Analysis of bulk RNA-seq data

Analysis of Next-Generation Sequencing Data

Friederike Dündar

Applied Bioinformatics Core

Slides at https://bit.ly/2CUdS9z[1]

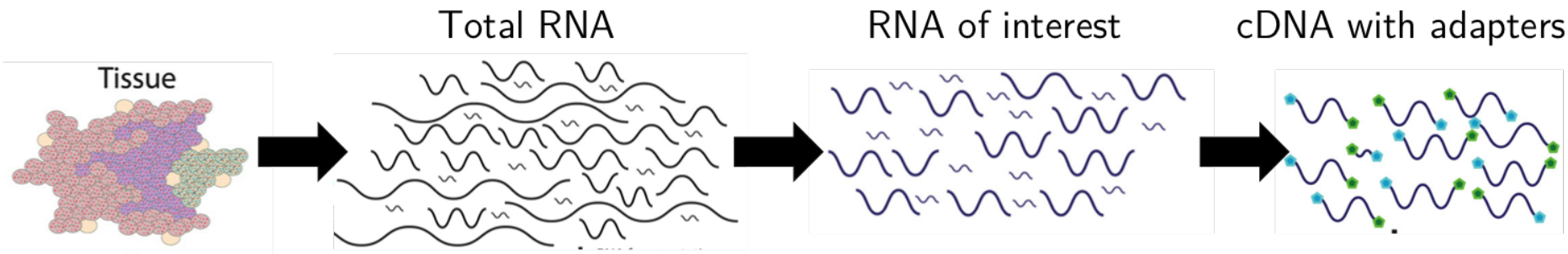
February 19, 2019

[1] http://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2018/

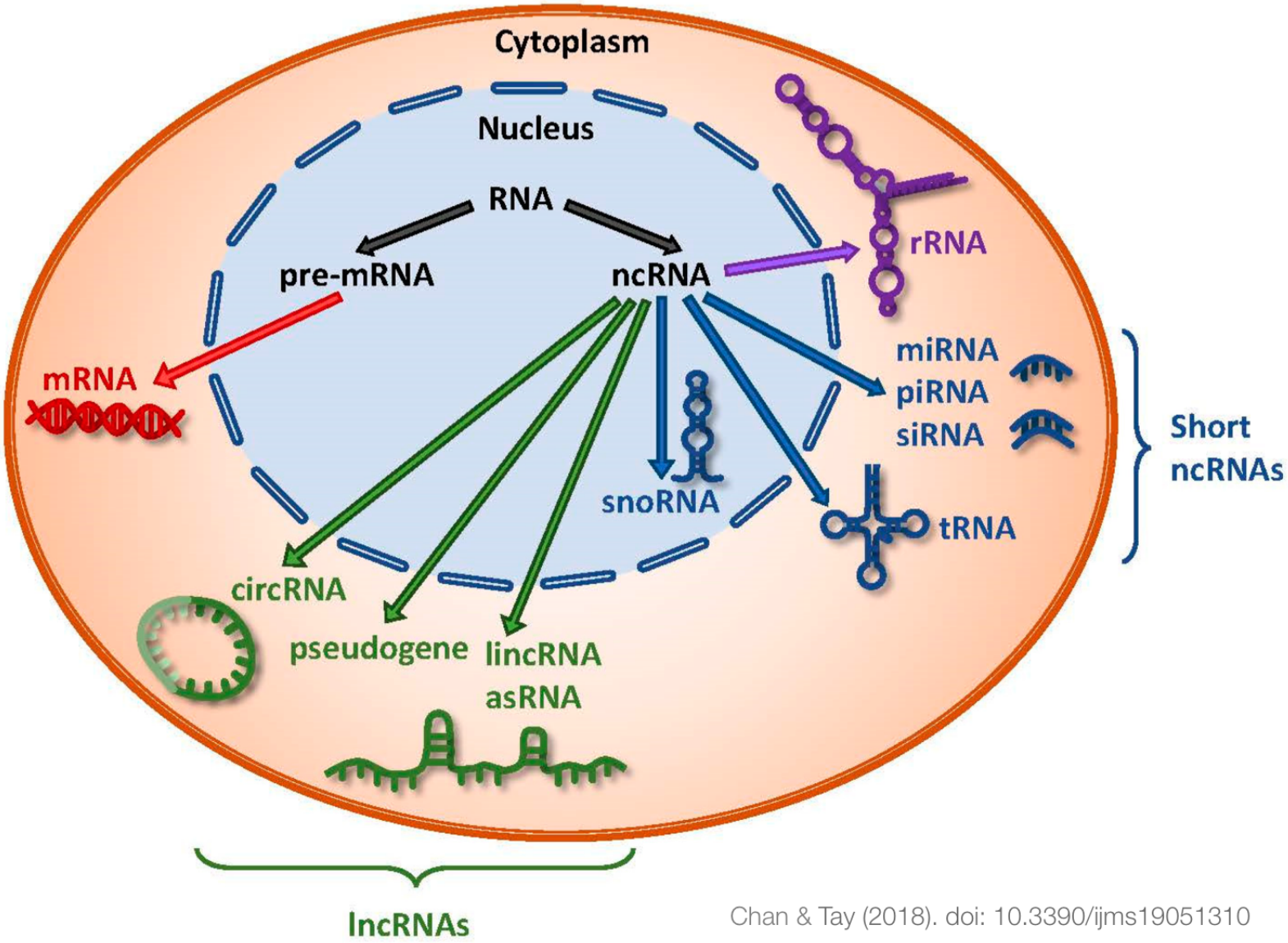
Different types of RNA – different library preps

General steps of RNA-seq preparation

1. RNA **extraction** (cell lysis, RNA purification)
2. **enrichment** of the RNA of interest
3. **fragmentation** (ca. 200 bp)
4. **cDNA** synthesis
5. library prep to obtain cDNA with **adapters** for sequencing

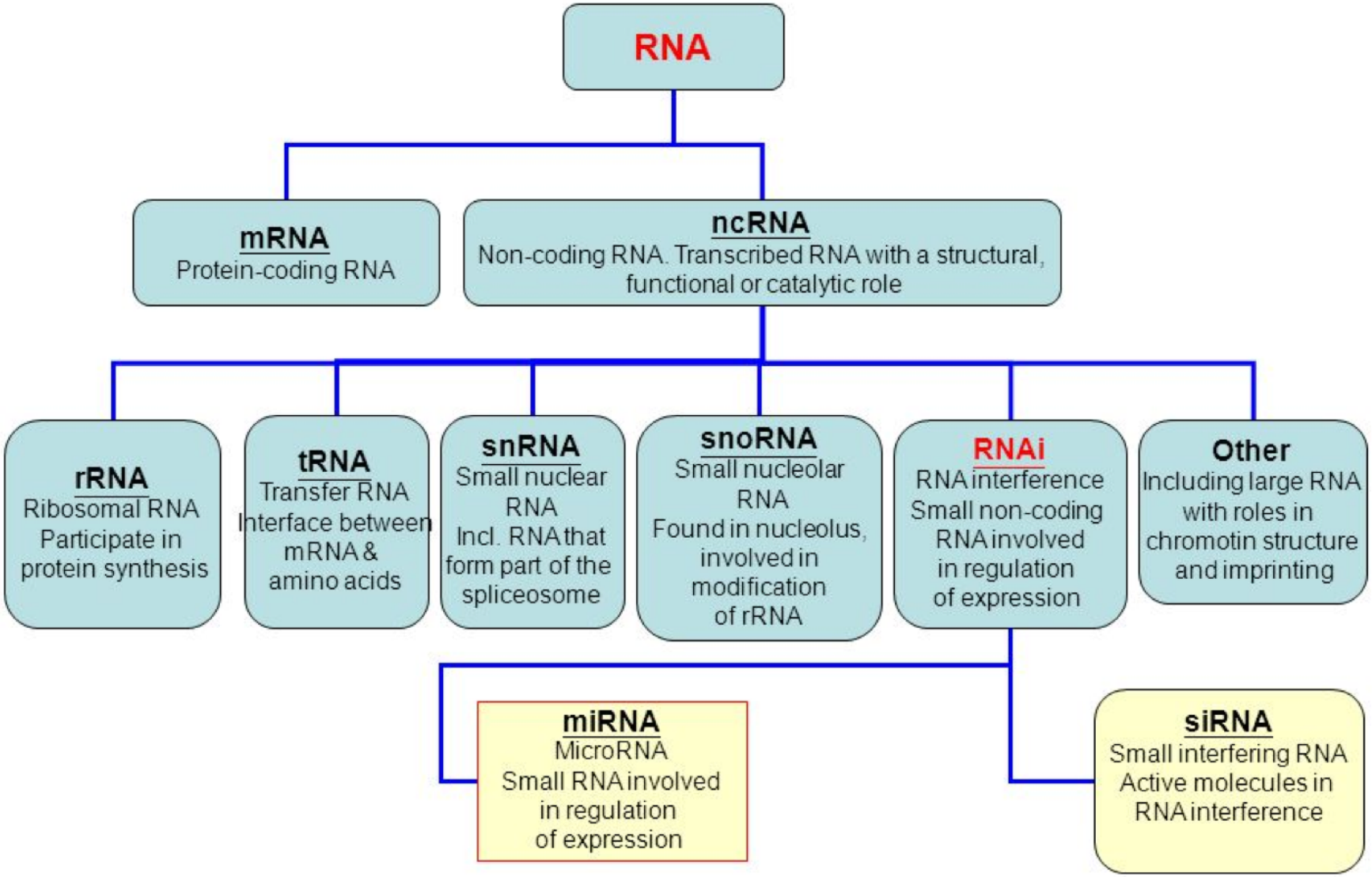


Different types of RNA (there are more!)



Chan & Tay (2018). doi: 10.3390/ijms19051310

Different types of RNA (there are more!)

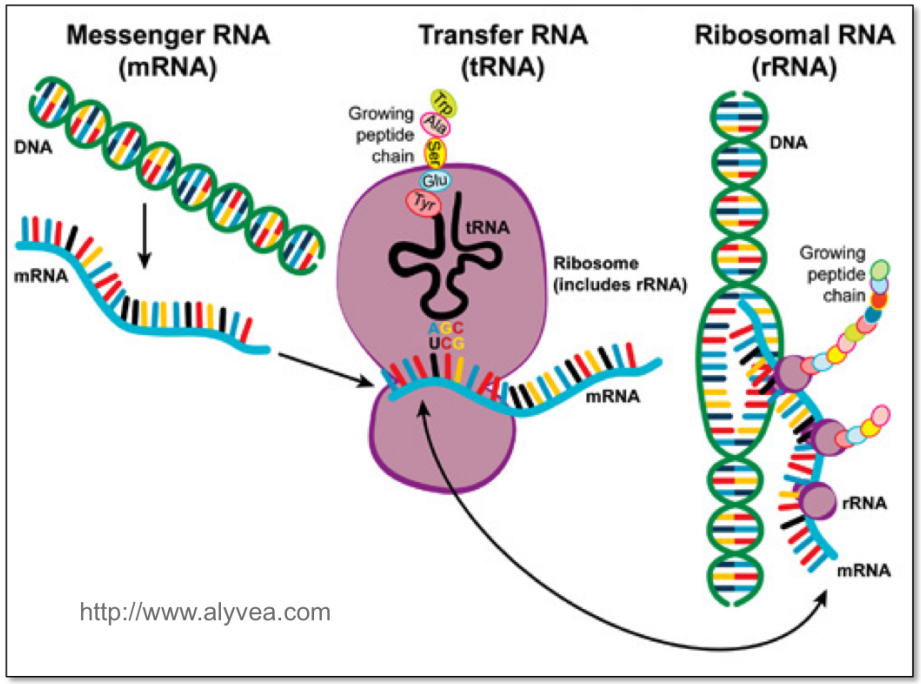


See, e.g., Wilkes et al. [2017] and Bartoszewski and Sikorski [2018] for an introduction

Sequencing prep protocol depends on the RNA properties

It is not a one-size-fits-all situation!

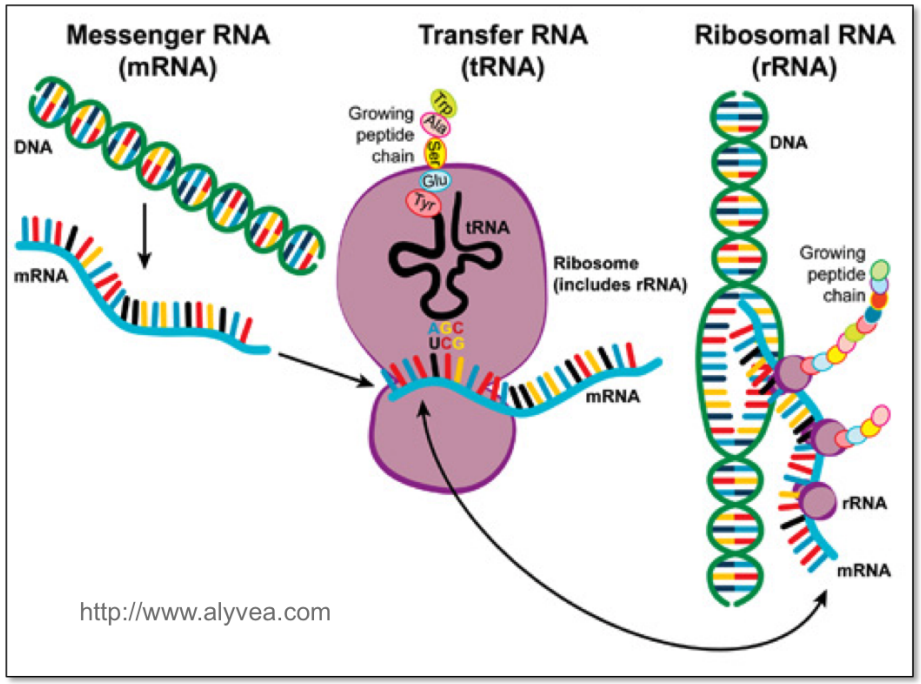
- **abundance** and stability
  - rRNA: 90-95% (!)
  - tRNA: 3-5%
  - mRNA: 2%
  - all other non-coding RNAs: *well* below 1%
- **cellular location**
  - most are in the cytoplasm
- **size**
  - miRNAs: 18-23bp
  - mRNA: several 100 to 1000 bp
- **specific sequences/modifications**
  - poly(A) tails of mRNA
  - 2D structure
  - antisense transcripts



Sequencing prep protocol depends on the RNA properties

It is not a one-size-fits-all situation!

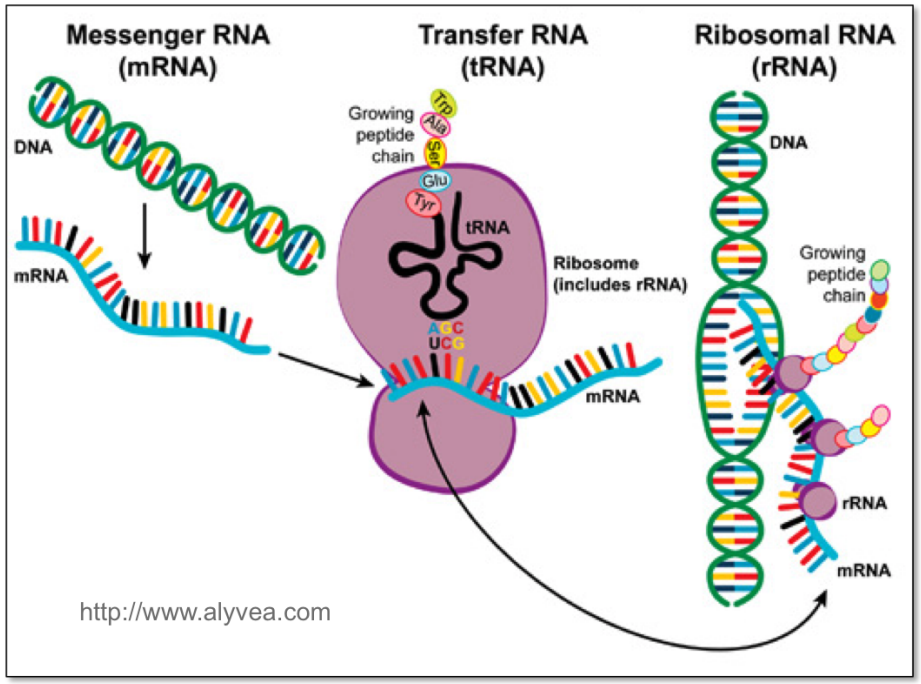
- **abundance** and stability
  - rRNA: 90-95% (!)
  - tRNA: 3-5%
  - mRNA: 2%
  - all other non-coding RNAs: *well* below 1%
- **cellular location**
  - most are in the cytoplasm
- **size**
  - miRNAs: 18-23bp
  - mRNA: several 100 to 1000 bp
- **specific sequences/modifications**
  - poly(A) tails of mRNA
  - 2D structure
  - antisense transcripts



Sequencing prep protocol depends on the RNA properties

It is not a one-size-fits-all situation!

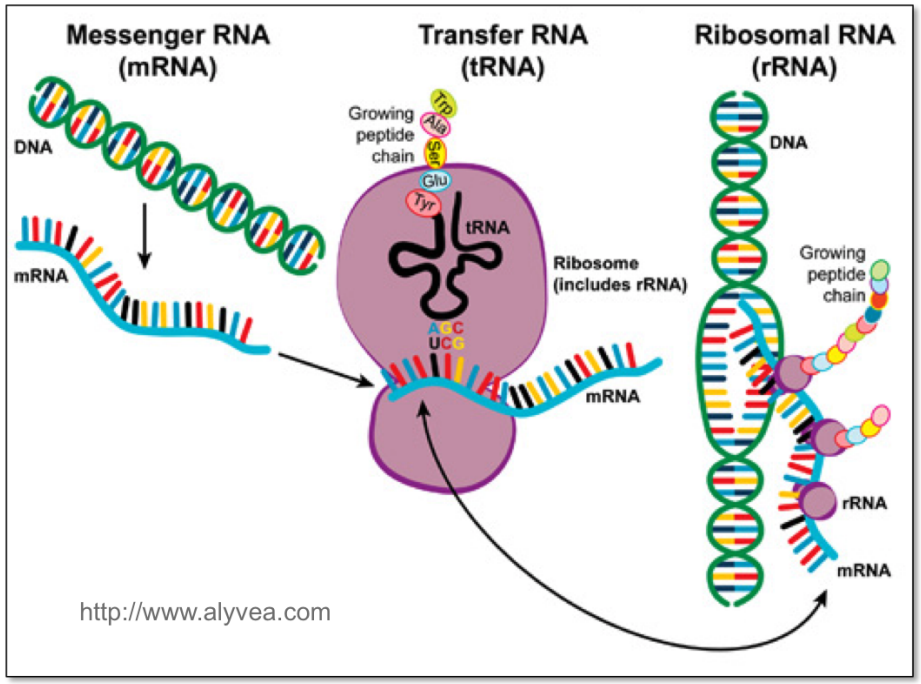
- **abundance** and stability
  - rRNA: 90-95% (!)
  - tRNA: 3-5%
  - mRNA: 2%
  - all other non-coding RNAs: *well* below 1%
- **cellular location**
  - most are in the cytoplasm
- **size**
  - miRNAs: 18-23bp
  - mRNA: several 100 to 1000 bp
- **specific sequences/modifications**
  - poly(A) tails of mRNA
  - 2D structure
  - antisense transcripts



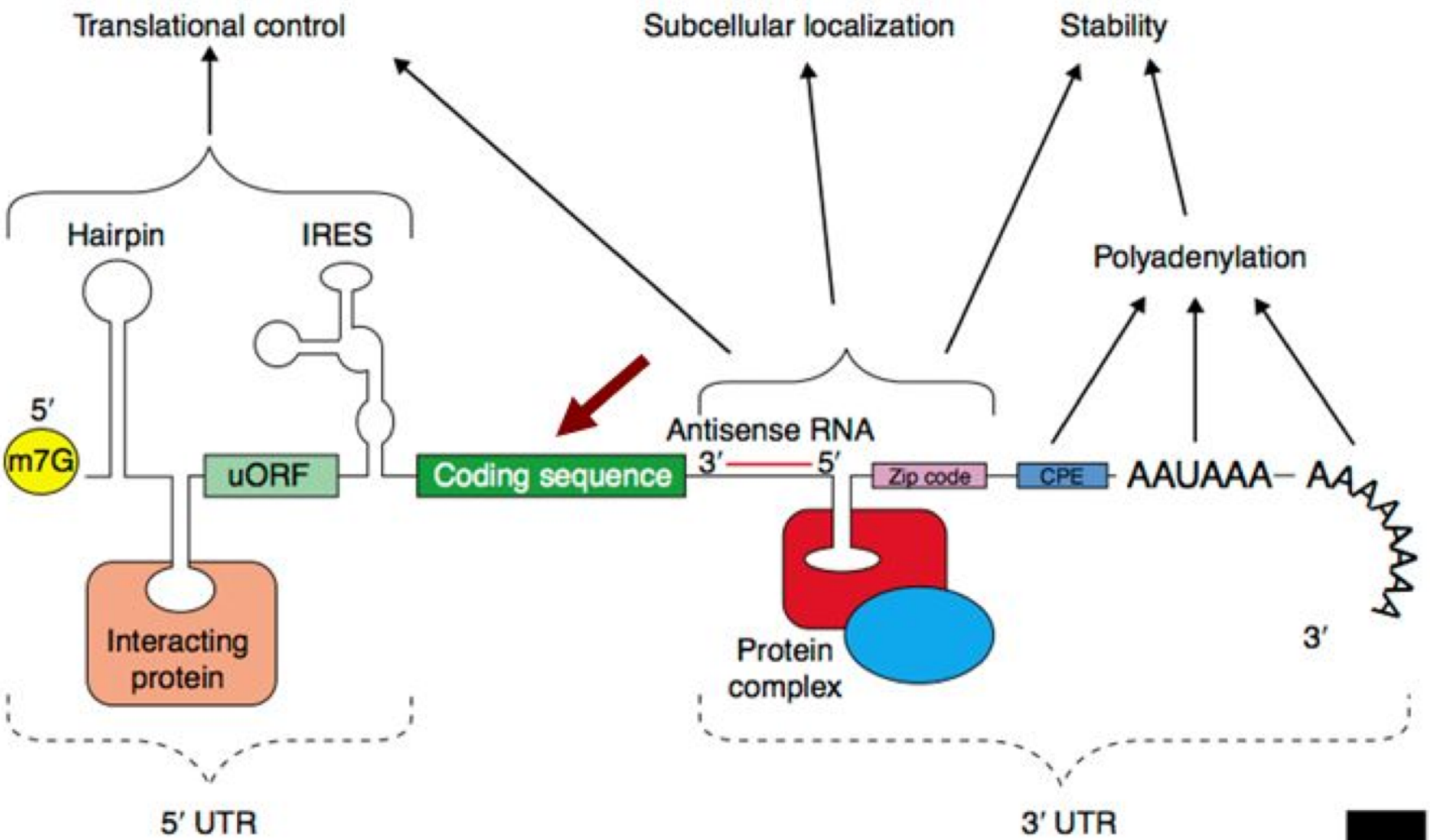
Sequencing prep protocol depends on the RNA properties

It is not a one-size-fits-all situation!

- **abundance** and stability
  - rRNA: 90-95% (!)
  - tRNA: 3-5%
  - mRNA: 2%
  - all other non-coding RNAs: *well* below 1%
- **cellular location**
  - most are in the cytoplasm
- **size**
  - miRNAs: 18-23bp
  - mRNA: several 100 to 1000 bp
- **specific sequences/modifications**
  - poly(A) tails of mRNA
  - 2D structure
  - antisense transcripts



mRNA alone has numerous facets

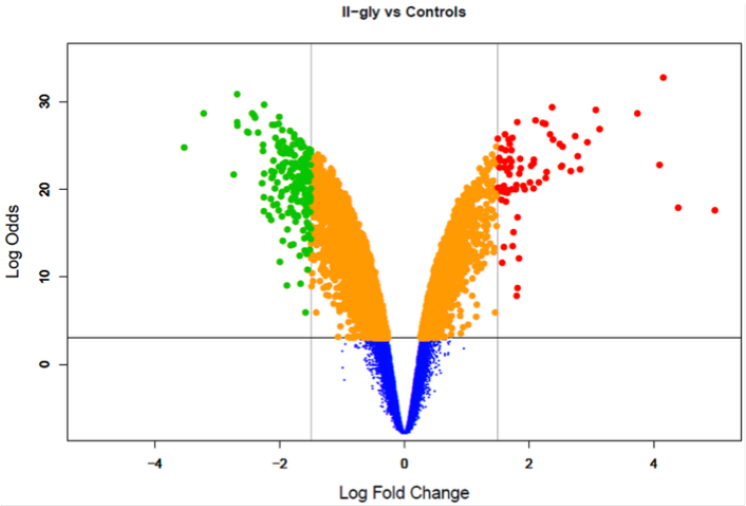


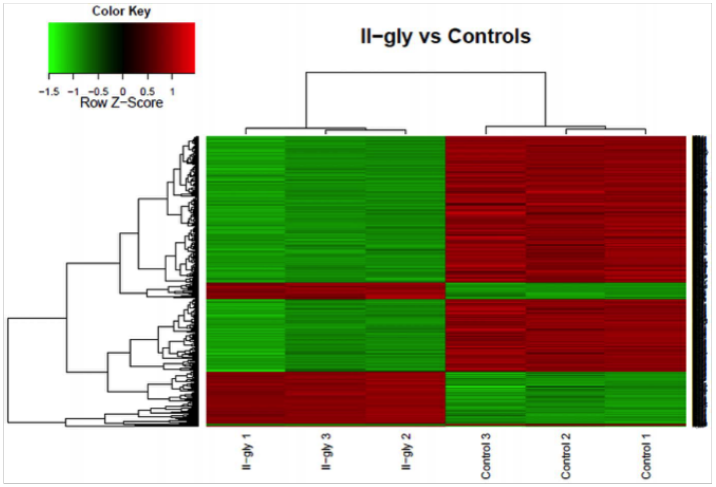
Mignone et al. (2002). doi: 10.1186/gb-2002-3-3-reviews0004

Focus today

Bulk RNA-seq of mRNA

- expression quantification of (mostly) **mRNA transcripts**
- extracted from **populations of cells**
- and tested for **gene-specific differences** between distinct **conditions**

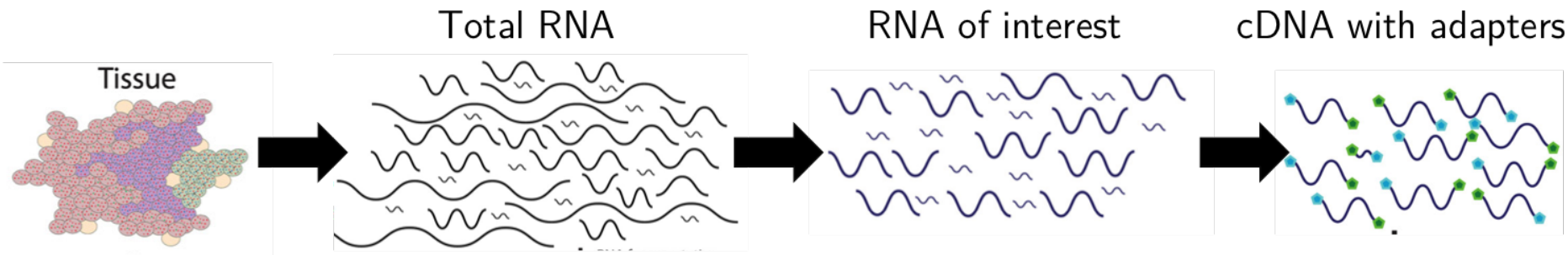




Valencia-Cruz et al. (2013). doi: 10.1371/journal.pone.0054664

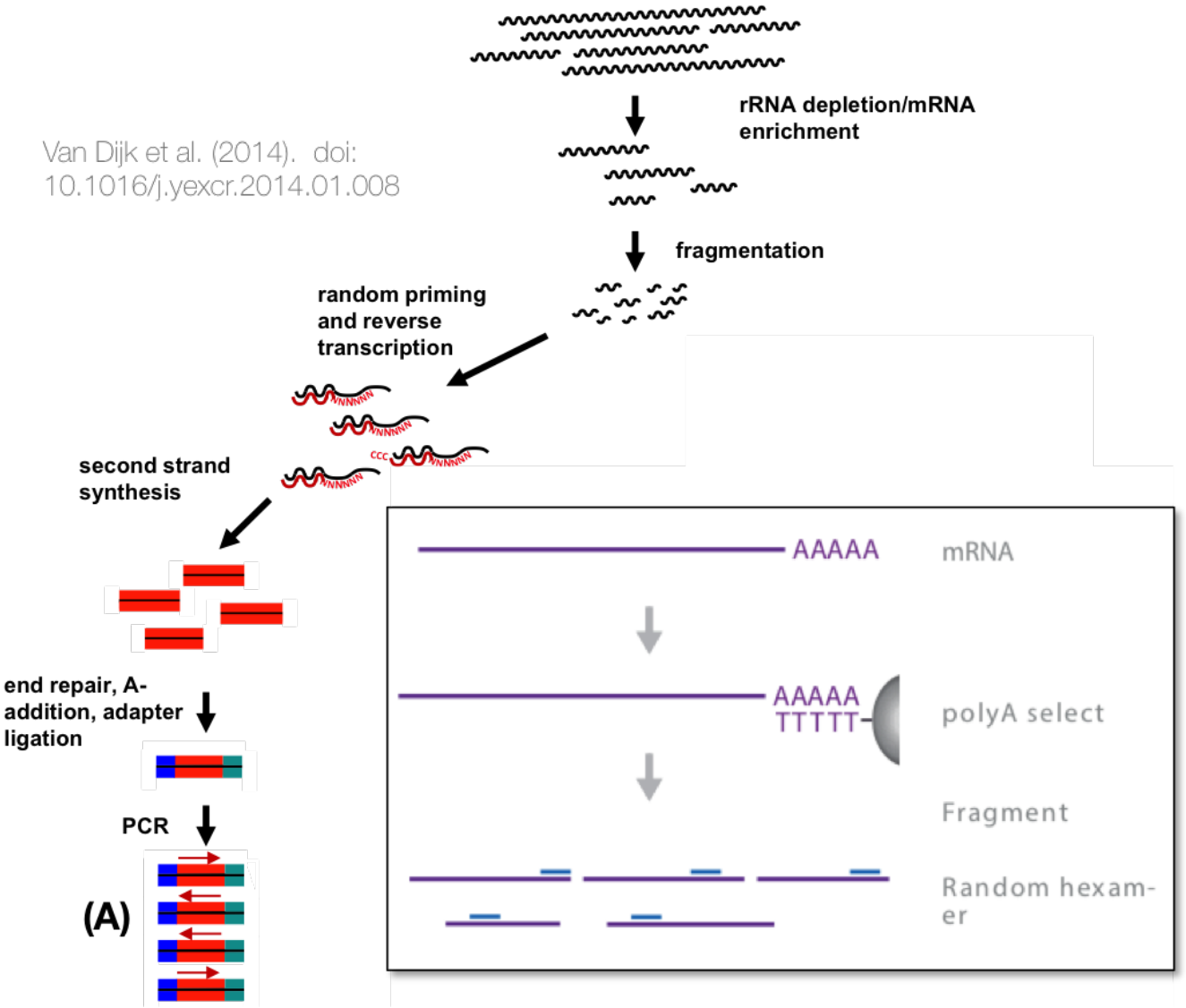
General steps of RNA-seq preparation

1. RNA **extraction**[2] (cell lysis, RNA purification)
2. **enrichment** of the RNA of interest
3. **fragmentation** (ca. 200 bp)
4. **cDNA** synthesis
5. library prep to obtain cDNA with **adapters** for sequencing



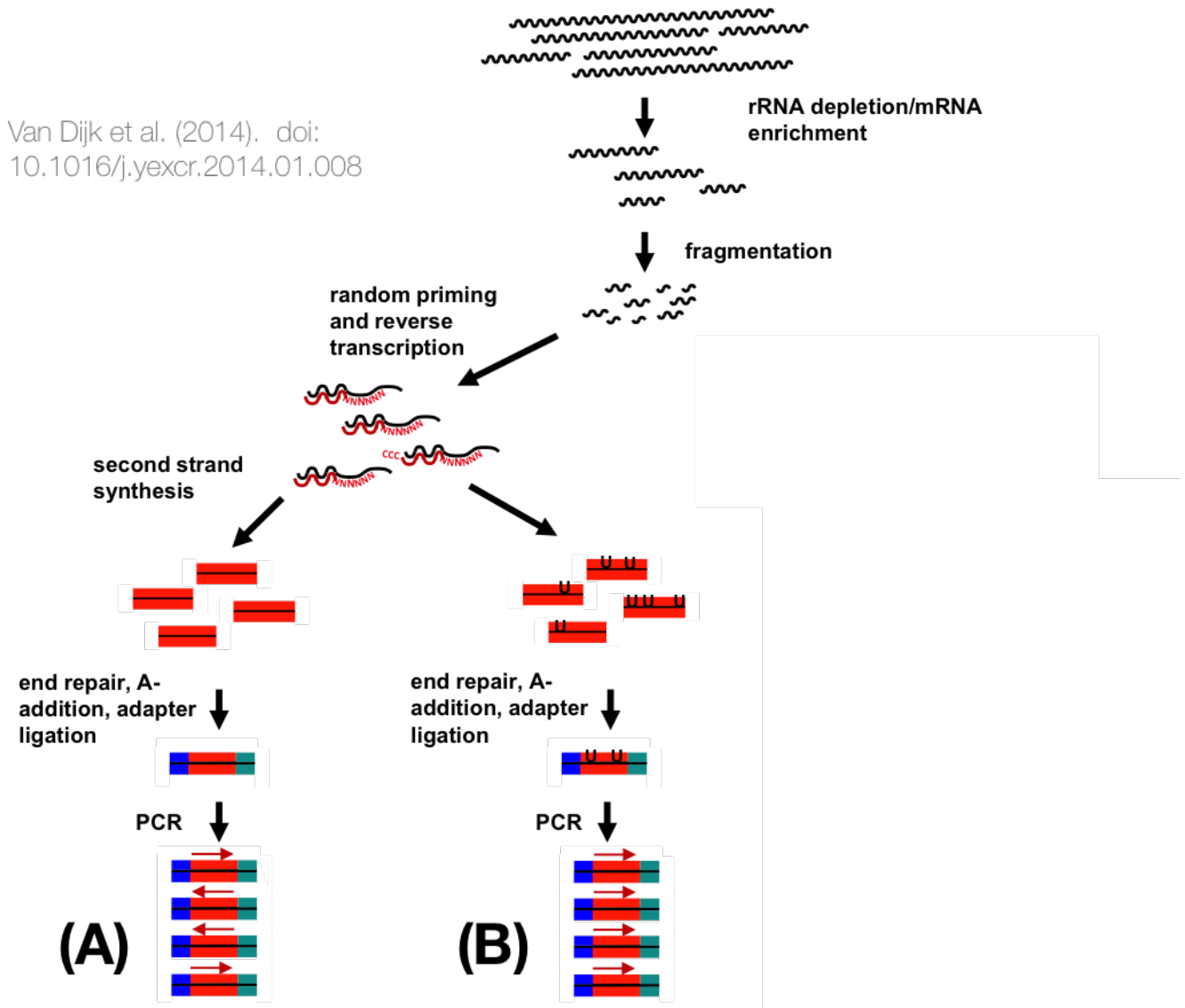
[2] Most standard extraction methods will lose RNA <100 bp!

The most common library preparation methods



- **(A)** classical unstranded mRNA library prep

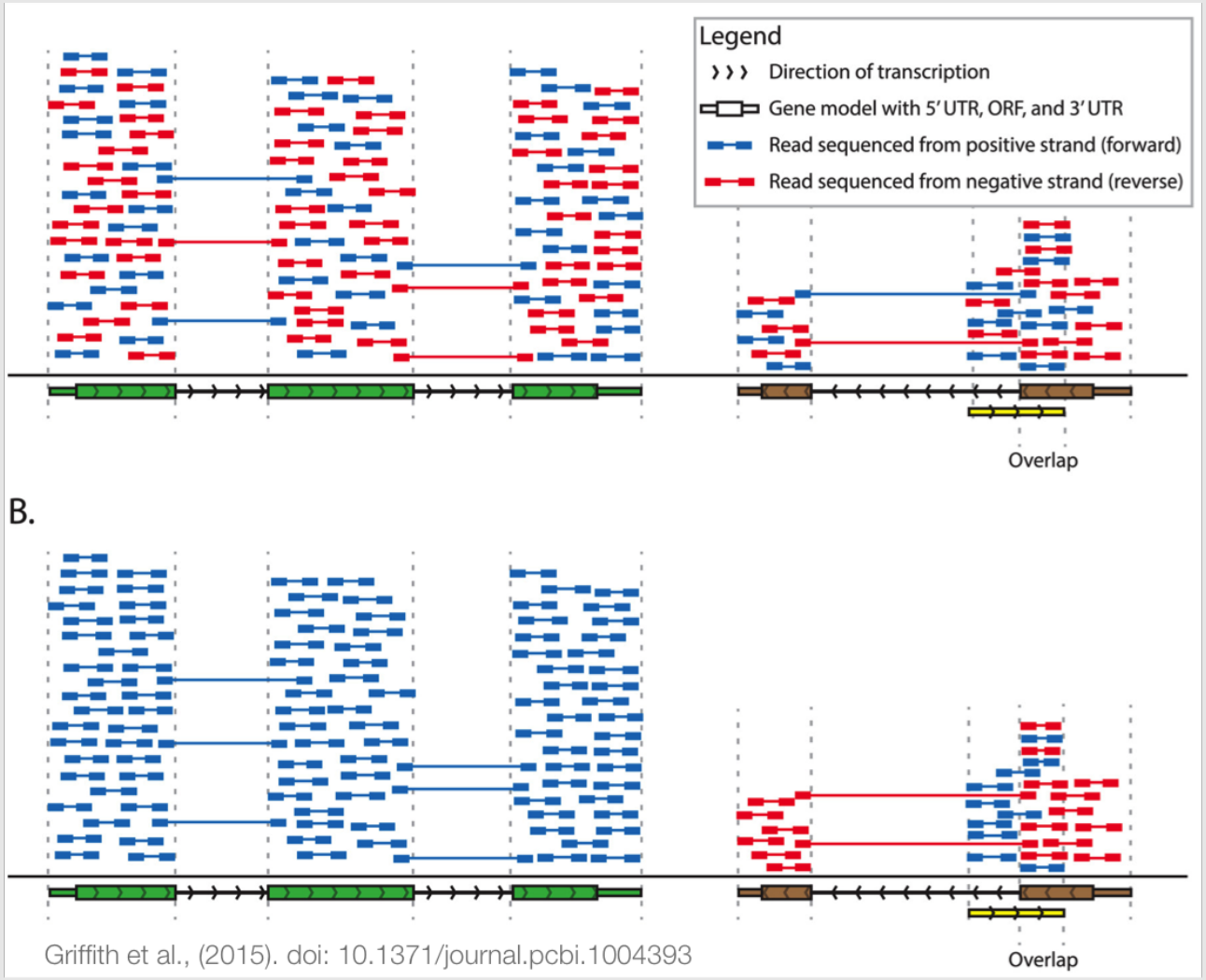
The most common library preparation methods



- **(A)** classical unstranded mRNA library prep
- **(B)** stranded mRNA (dUTP-based) (see Levin et al. [2010] and Zhao et al. [2015] for details)

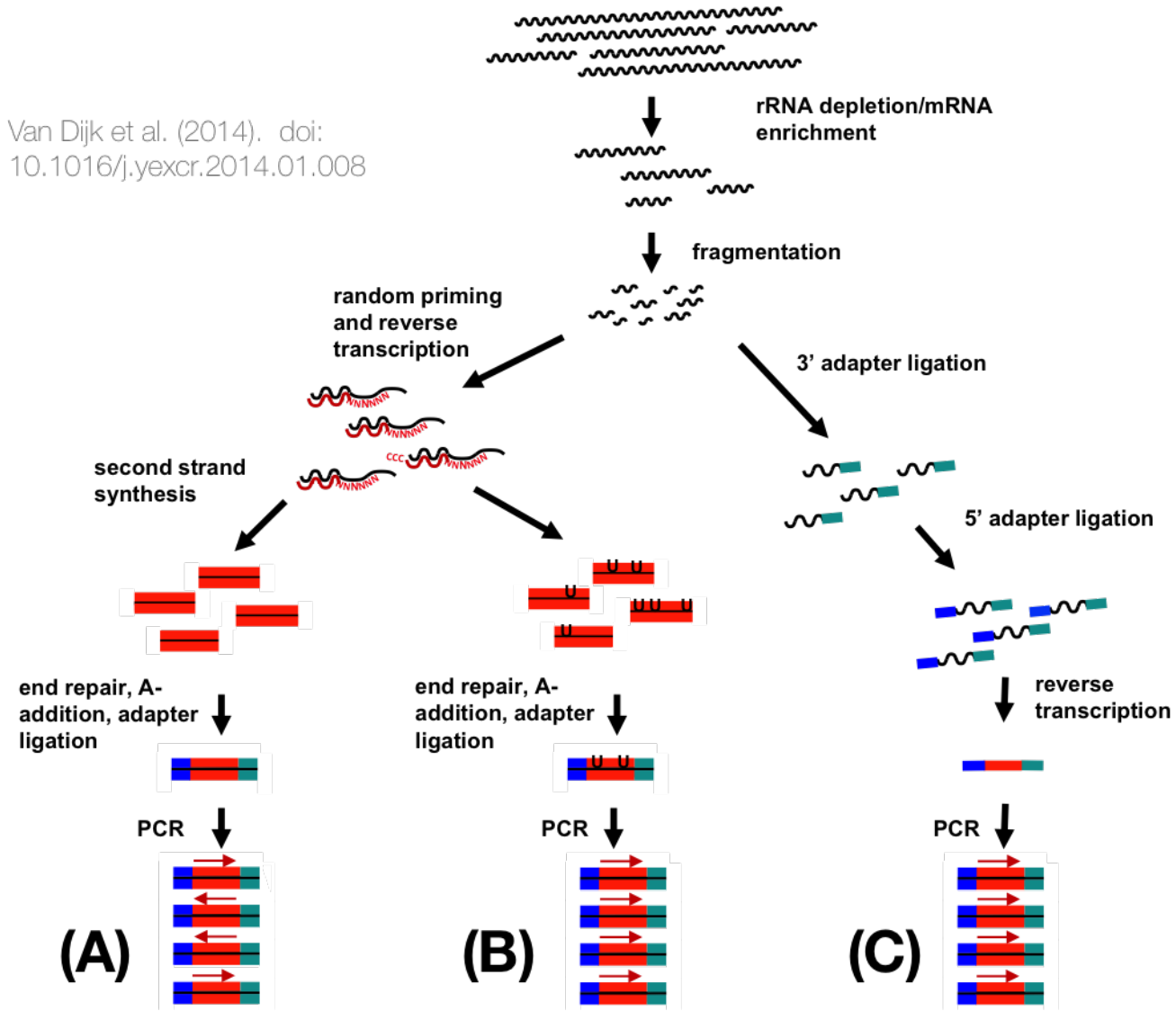
The most common library preparation methods

Unstranded vs. stranded



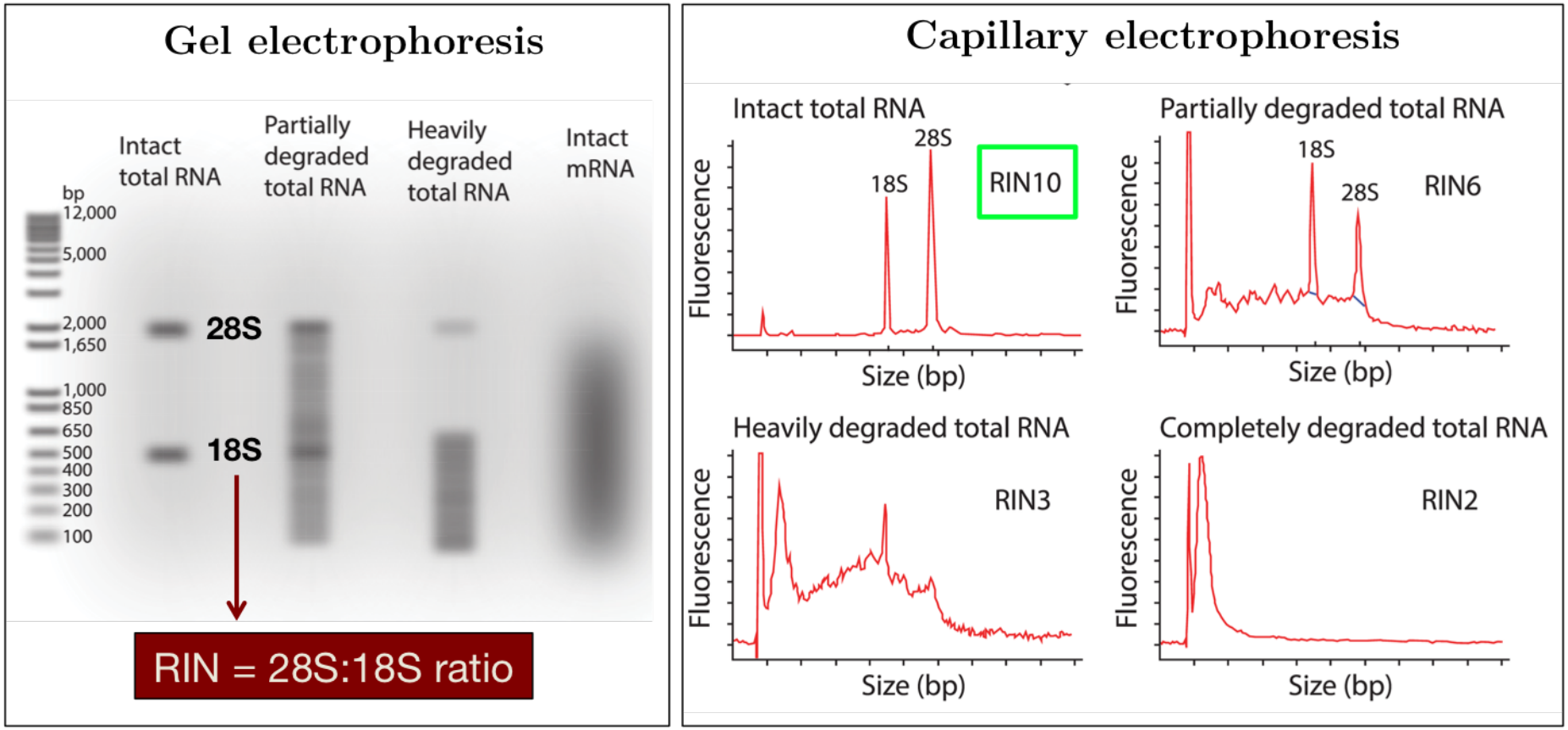
Griffith et al., (2015). doi: 10.1371/journal.pcbi.1004393

The most common library preparation methods



- **(A)** classical unstranded mRNA library prep
- **(B)** stranded mRNA (dUTP-based) (see Levin et al. [2010] and Zhao et al. [2015] for details)
- **(C)** small RNAs (miRNA, piRNA, tRNA, ... <100 bp) using 2 adapters – not optimal for differential expression analyses!

QC of RNA extraction

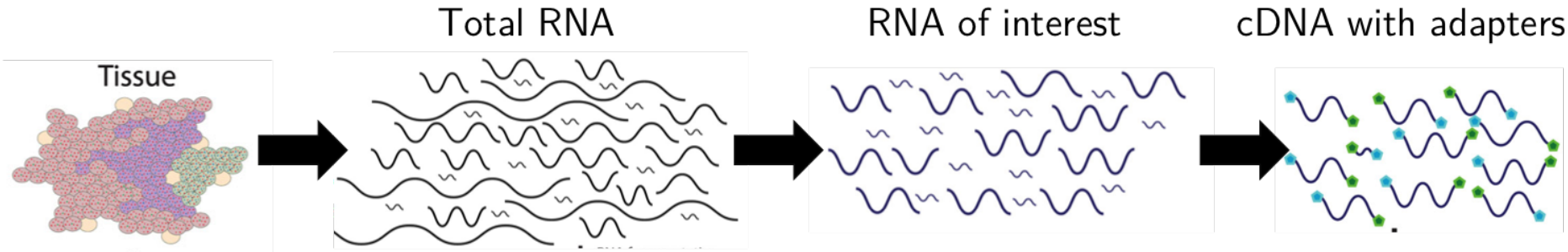


Griffith et al. (2015). doi: 10.1371/journal.pcbi.1004393

Avoid degraded RNA! Optimum: RNA Integraty Score (RIN) of 10.

General steps of RNA-seq preparation

1. RNA **extraction** (cell lysis, RNA purification)
2. **enrichment** of the RNA of interest
   - mRNA: poly(A) enrichment vs. ribosomal-depletion
   - small RNAs: size-based enrichment
3. **fragmentation** (ca. 200 bp)
4. **cDNA** synthesis
5. library prep to obtain cDNA with **adapters** for sequencing

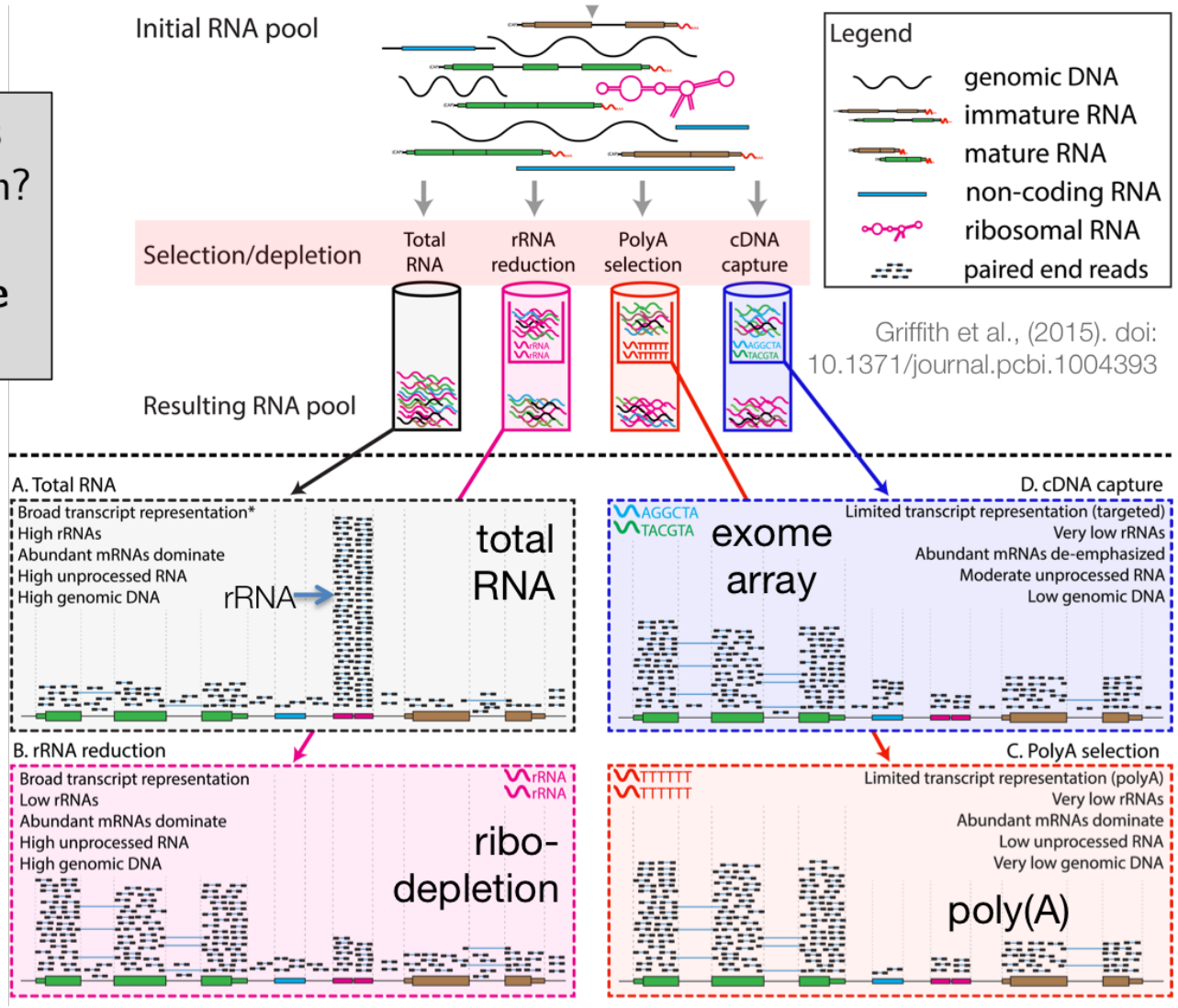


Every step has consequences – example: mRNA enrichment

which **transcripts** are you interested in?

what type of **noise** can you tolerate?

- rRNA
- protein coding (strongly expressed)
- protein coding (lowly expressed)



Every step has consequences

- Do not mix different strategies for samples that are to be compared to each other!
  - extraction, enrichment, library prep

There are many papers comparing different aspects of different RNA-seq approaches, e.g.

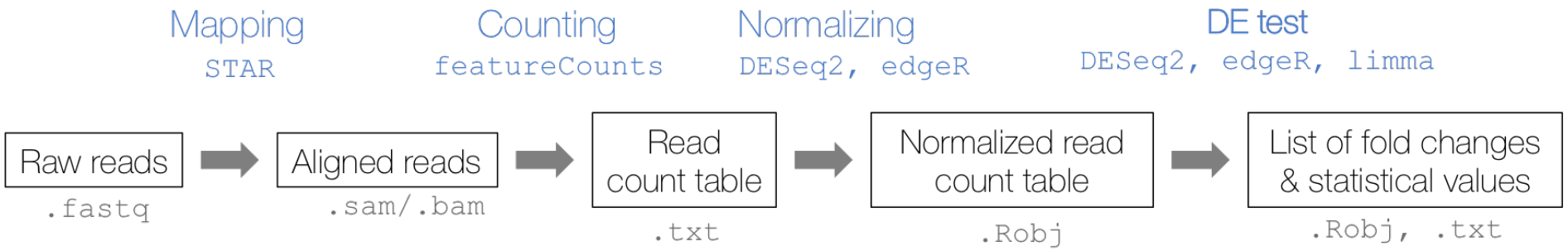
- *Library preparation methods for next-generation sequencing: Tone down the bias* [van Dijk et al., 2014]
- *Systematic comparison of small RNA library preparation protocols for next-generation sequencing* [Dard-Dascot et al., 2018]
- *A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples.* [Schuierer et al., 2017]
- many more – PubMed is your friend!

Make an informed decision!

Gene expression quantification

General bioinformatics workflow for bulk RNA-seq data

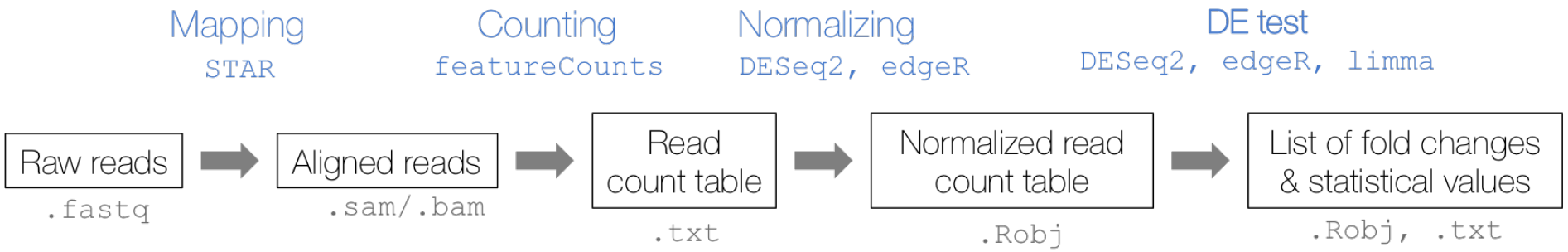
Gene expression quantification (counting reads per gene following alignment) is typically followed by differential gene expression (DE or DGE) analysis.



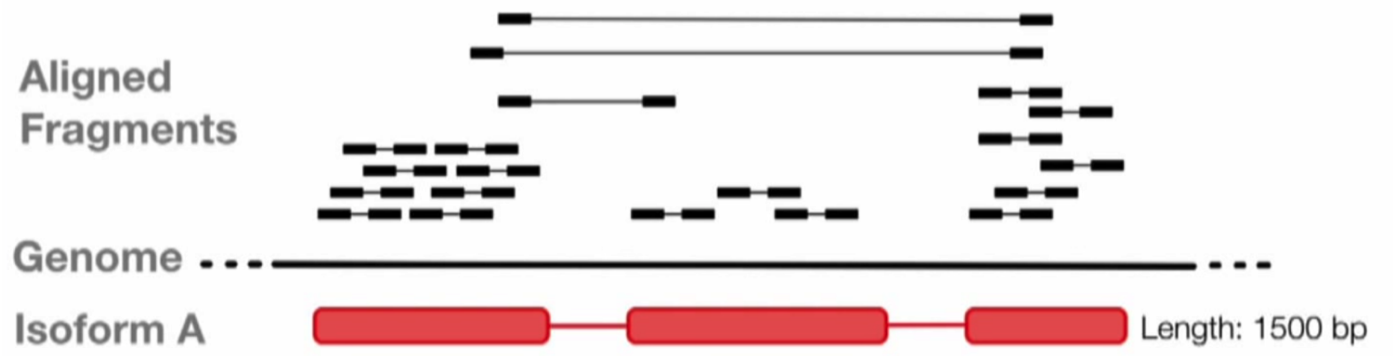
Null hypothesis

There is no difference in the expression levels of inidividual genes in condition A and condition B.

Quantification of gene expression



1. **Align**
   - with splice-aware alignment tools! e.g. STAR
2. **Count** reads that overlap with annotated genes



1. Aligning reads using STAR

```
$ mkdir alignment
$ cd alignment/
$ ln -s ~frd2007/ANGSD_2019/RNA-seq/raw_reads_Gierlinski_yeast/
```

I had previously downloaded numerous samples of the Gierlinski data set. These are stored in the folder `RNA-seq/raw_reads*` to which I have now created a symbolic link:

```
$ ls -lahF raw_reads_Gierlinski_yeast/
 total 44K
 drwxr-xr-x 12 frd2007 abc  126 Jan 31 14:51 ./
 drwxr-xr-x  3 frd2007 abc 4.0K Jan 31 14:55 ../
 drwxr-xr-x  2 frd2007 abc 4.0K Jan 31 14:51 SNF2_1/
 drwxr-xr-x  2 frd2007 abc 4.0K Jan 31 14:51 SNF2_2/
 drwxr-xr-x  2 frd2007 abc 4.0K Jan 31 14:51 SNF2_3/
 drwxr-xr-x  2 frd2007 abc 4.0K Jan 31 14:51 SNF2_4/
 drwxr-xr-x  2 frd2007 abc 4.0K Jan 31 14:51 SNF2_5/
 drwxr-xr-x  2 frd2007 abc 4.0K Jan 31 14:51 WT_1/
 drwxr-xr-x  2 frd2007 abc 4.0K Jan 31 14:49 WT_2/
 drwxr-xr-x  2 frd2007 abc 4.0K Jan 31 14:50 WT_3/
 drwxr-xr-x  2 frd2007 abc 4.0K Jan 31 14:50 WT_4/
 drwxr-xr-x  2 frd2007 abc 4.0K Jan 31 14:50 WT_5/
```

1. Aligning reads using STAR

Every subfolder contains the technical replicates of the respective sample:

```
$ ls -lahF raw_reads_Gierlinski_yeast/SNF2_1/
 total 627M
 drwxr-xr-x  2 frd2007 abc 4.0K Jan 31 14:51 ./
 drwxr-xr-x 12 frd2007 abc  126 Jan 31 14:51 ../
 -rw-r--r--  1 frd2007 abc  98M Jan 31 14:50 ERR458500.fastq.gz
 -rw-r--r--  1 frd2007 abc  97M Jan 31 14:51 ERR458501.fastq.gz
 -rw-r--r--  1 frd2007 abc  96M Jan 31 14:51 ERR458502.fastq.gz
 -rw-r--r--  1 frd2007 abc  88M Jan 31 14:51 ERR458503.fastq.gz
 -rw-r--r--  1 frd2007 abc  76M Jan 31 14:51 ERR458504.fastq.gz
 -rw-r--r--  1 frd2007 abc  77M Jan 31 14:51 ERR458505.fastq.gz
 -rw-r--r--  1 frd2007 abc  98M Jan 31 14:51 ERR458506.fastq.gz
```

For the alignment, I will use STAR.

```
$ spack find | egrep -i STAR
 star@2.5.3a
 star@2.6.1a

$ spack load star@2.6.1a
```

1. Aligning reads using STAR

To determine suitable numbers for IntronMin and IntronMax parameters, we downloaded a bed file for the yeast introns from UCSC table browser (details https://www.biostars.org/p/13290/)

```
ln -s ~frd2007/ANGSD_2019/RNA-seq/refGenome_S_cerevisiae/introns_yeast.bed
 # get min. intron size
 awk '{print $3-$2}' introns_yeast.bed | sort -k1n | uniq | head -n 3
 1
 31
 35

 # get max. intron size
 awk '{print $3-$2}' introns_yeast.bed | sort -k1n | uniq | tail -n 3
 1623
 2448
 2483
```

Now that we have a feeling for what the sizes of annotated introns look like, we can run STAR.

1. Aligning reads using STAR: Genome Index with exon boundary info

REMEMBER!

1. build an index
2. align

```
 ## GENOME & TRANSCRIPTOME INDEX BUILDING
mkdir /home/frd2007/ANGSD_2019/RNA-seq/refGenome_S_cerevisiae/STARindex
ln -s /home/frd2007/ANGSD_2019/RNA-seq/refGenome_S_cerevisiae/

 # Run STAR in "genomeGenerate" mode
 $ STAR --runMode genomeGenerate
  --genomeDir refGenome_S_cerevisiae/STARindex # where index will be stored
  --genomeFastaFiles refGenome_S_cerevisiae/sacCer3.fa # ref. genome seq.
  --sjdbGTFfile refGenome_S_cerevisiae/sacCer3.gtf # annotation file
  --sjdbOverhang 49 # should be read length minus 1
 --runThreadN 1  # can be used to define more processors
```

1. Aligning reads using STAR

For the alignment, I will use a for-loop over all the samples (WT repl. 1-5 and SNF2 repl. 1-5)
In order to make STAR use all the reads from all the technical replicates per sample, we need to list the respective files as **comma-separated lists**. This is not as trivial as it sounds, so I double-check that my command works:

```
$ 'ls' raw_reads_Gierlinski_yeast/WT_1/*.fastq.gz | paste -s -d , -
raw_reads_Gierlinski_yeast/WT_1/ERR458493.fastq.gz,
raw_reads_Gierlinski_yeast/WT_1/ERR458494.fastq.gz,
raw_reads_Gierlinski_yeast/WT_1/ERR458495.fastq.gz,
raw_reads_Gierlinski_yeast/WT_1/ERR458496.fastq.gz,
raw_reads_Gierlinski_yeast/WT_1/ERR458497.fastq.gz,
raw_reads_Gierlinski_yeast/WT_1/ERR458498.fastq.gz,
raw_reads_Gierlinski_yeast/WT_1/ERR458499.fastq.gz
```

Looks good! Off to the alignment then! I'll write a short (not very robust or generic!) script that I will use in a for-loop on all the samples.

1. Aligning reads using STAR

```
$ cat align_Gierlinski.sh
#! /bin/bash
# Read in arguments
STAR_DIR=$1
FASTQ_DIR=$2
SAMPLE=$3

# Define the list of fastq files per sample
FILES=`'ls' ${FASTQ_DIR}/${SAMPLE}/*.fastq.gz | paste -s -d , -`

# Run STAR
STAR --genomeDir ${STAR_DIR}/ --readFilesIn $FILES \
  --readFilesCommand gunzip -c --outFileNamePrefix ${SAMPLE}_  \
  --outFilterMultimapNmax 1 \
  --outSAMtype BAM SortedByCoordinate \
  --runThreadN 4 --twopassMode Basic  \
  --alignIntronMin 1 --alignIntronMax 3000
```

You can see the entire script here:
`~frd2007/ANGSD_2019/alignment/align_Gierlinski.sh`.

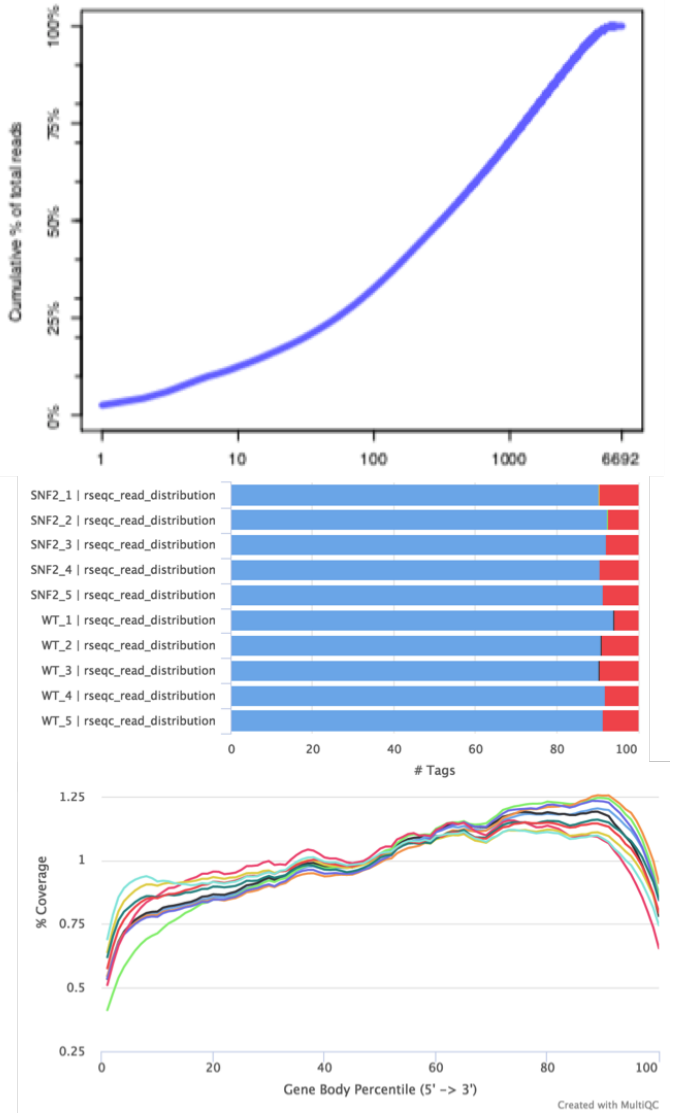
1. Aligning reads using STAR

```
# Make the script executable:
$ chmod 755 align_Gierlinski.sh

# Run it for all the samples of interest:
for SAMPLE in WT_1 WT_2 WT_3 WT_4 WT_5 SNF2_1 SNF2_2 SNF2_3 SNF2_4 SNF2_5
do
./align_Gierlinski.sh refGenome_S_cerevisiae/STARindex/ \
  raw_reads_Gierlinski_yeast/ $SAMPLE
done

# Should have added the indexing of the BAM files to the script,
# now I have to do it manually:
$ spack load samtools@1.9%gcc@6.3.0
$ for i in *bam
    do
        samtools index $i
    done
```

Alignment QC: RNA-seq-specific biases

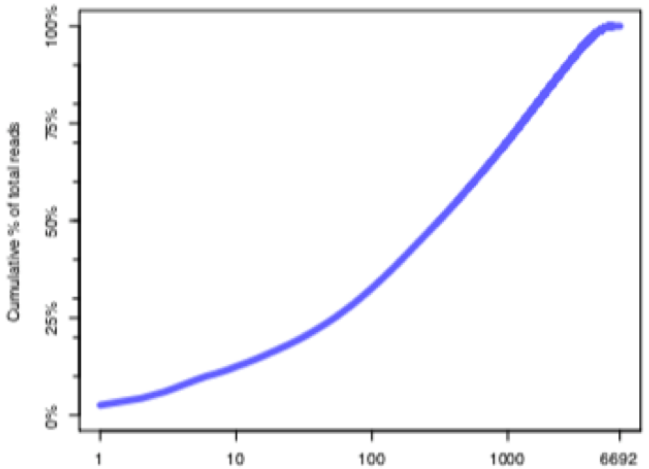
Typical biases of aligned reads of RNA-seq

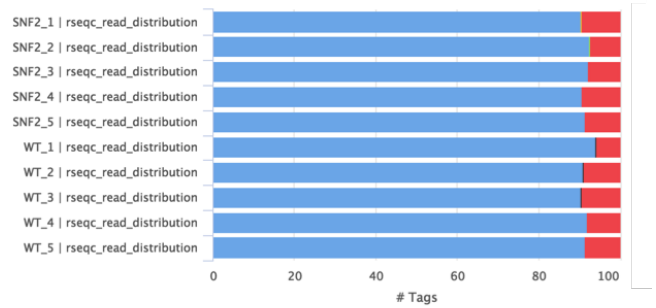
- lack of **gene diversity**: dominance of rRNAs, tRNAs (and/or other highly abundant transripts)
  - should be visible in `FastQC` results already
- **read distribution**
  - high **intron** coverage: incomplete poly(A) enrichment
  - many **intergenic** reads: gDNA contamination
- **gene body coverage**
  - **3' bias**: RNA degradation (and indicator of poly(A) enrichment)

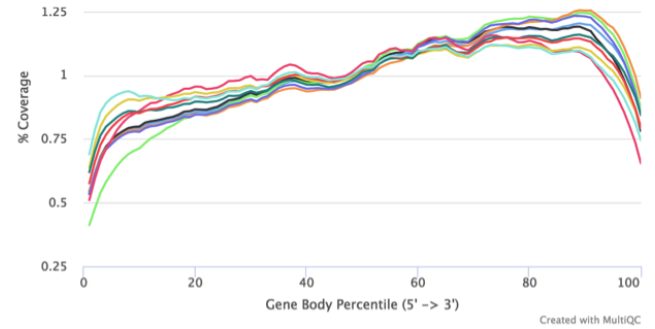


Typical biases of aligned reads of RNA-seq

- lack of **gene diversity**: dominance of rRNAs, tRNAs (and/or other highly abundant transripts)
  - should be visible in `FastQC` results already
- **read distribution**
  - high **intron** coverage: incomplete poly(A) enrichment
  - many **intergenic** reads: gDNA contamination
- **gene body coverage**
  - **3' bias**: RNA degradation (and indicator of poly(A) enrichment)

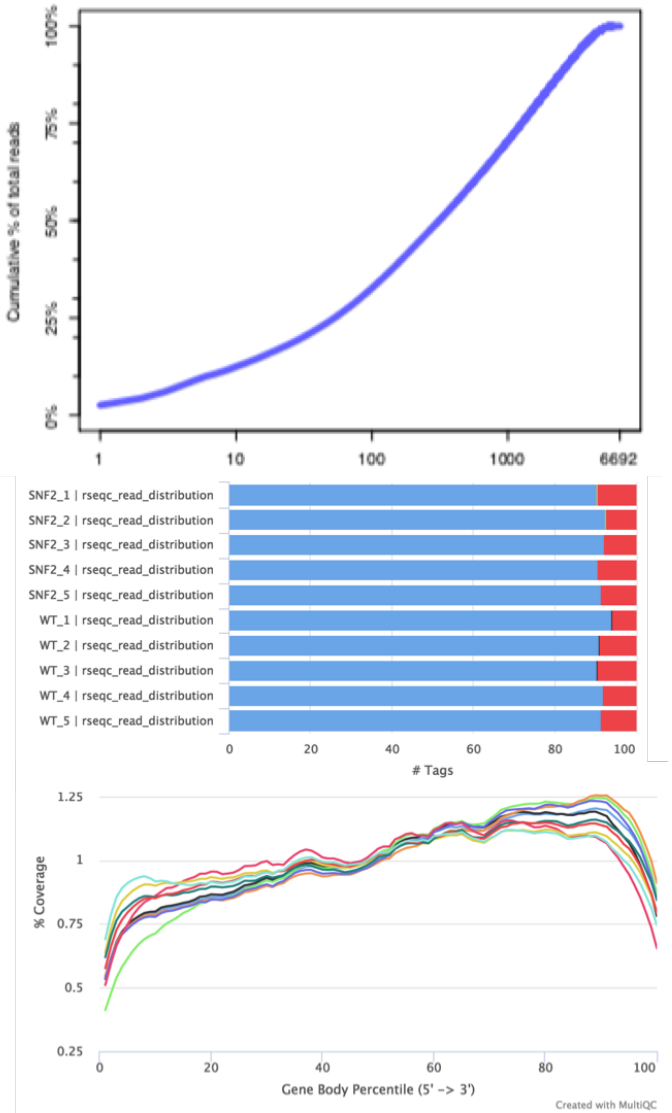






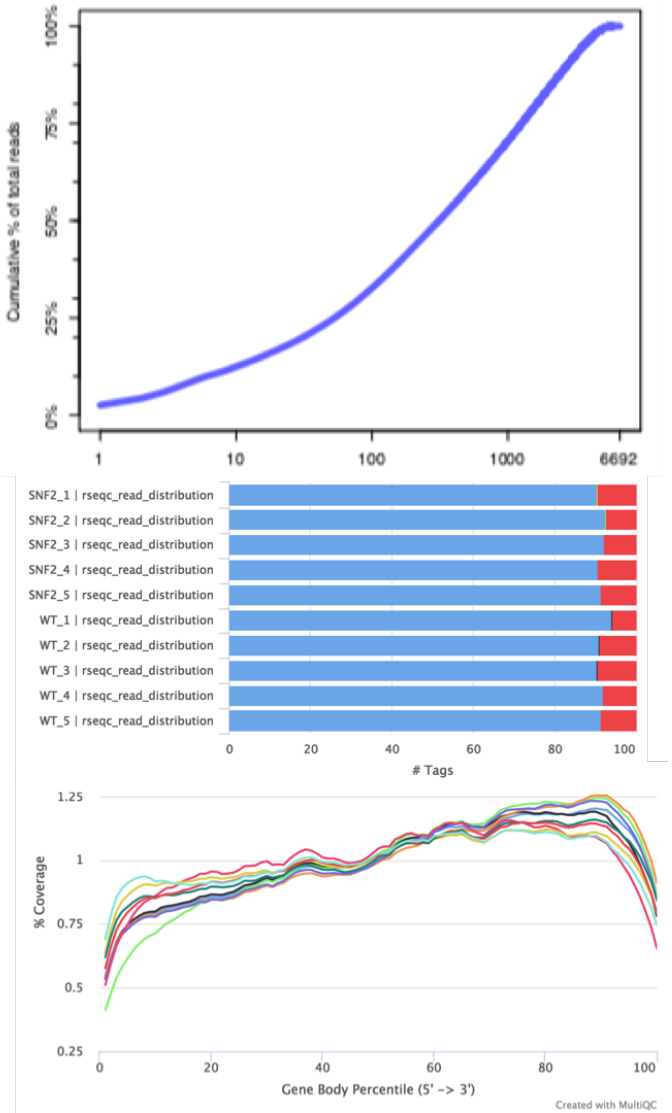
Typical biases of aligned reads of RNA-seq

- lack of **gene diversity**: dominance of rRNAs, tRNAs (and/or other highly abundant transripts)
  - should be visible in `FastQC` results already
- **read distribution**
  - high **intron** coverage: incomplete poly(A) enrichment
  - many **intergenic** reads: gDNA contamination
- **gene body coverage**
  - **3' bias**: RNA degradation (and indicator of poly(A) enrichment)



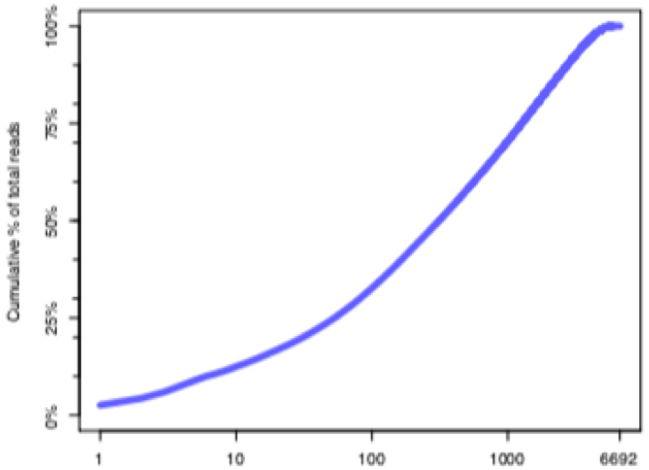
Typical biases of aligned reads of RNA-seq

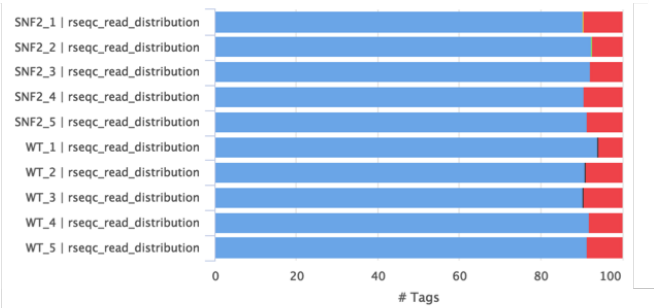
- lack of **gene diversity**: dominance of rRNAs, tRNAs (and/or other highly abundant transripts)
  - should be visible in `FastQC` results already
- **read distribution**
  - high **intron** coverage: incomplete poly(A) enrichment
  - many **intergenic** reads: gDNA contamination
- **gene body coverage**
  - **3' bias**: RNA degradation (and indicator of poly(A) enrichment)

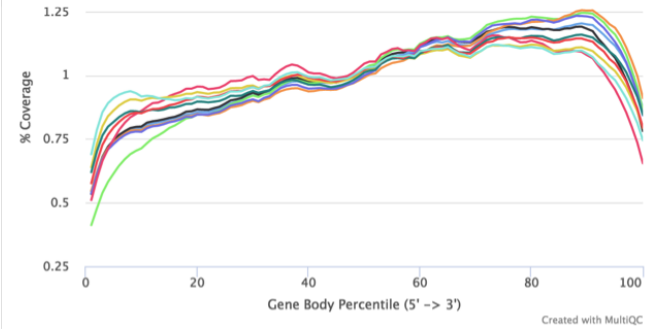


Typical biases of aligned reads of RNA-seq

- lack of **gene diversity**: dominance of rRNAs, tRNAs (and/or other highly abundant transripts)
  - should be visible in `FastQC` results already
- **read distribution**
  - high **intron** coverage: incomplete poly(A) enrichment
  - many **intergenic** reads: gDNA contamination
- **gene body coverage**
  - **3' bias**: RNA degradation (and indicator of poly(A) enrichment)

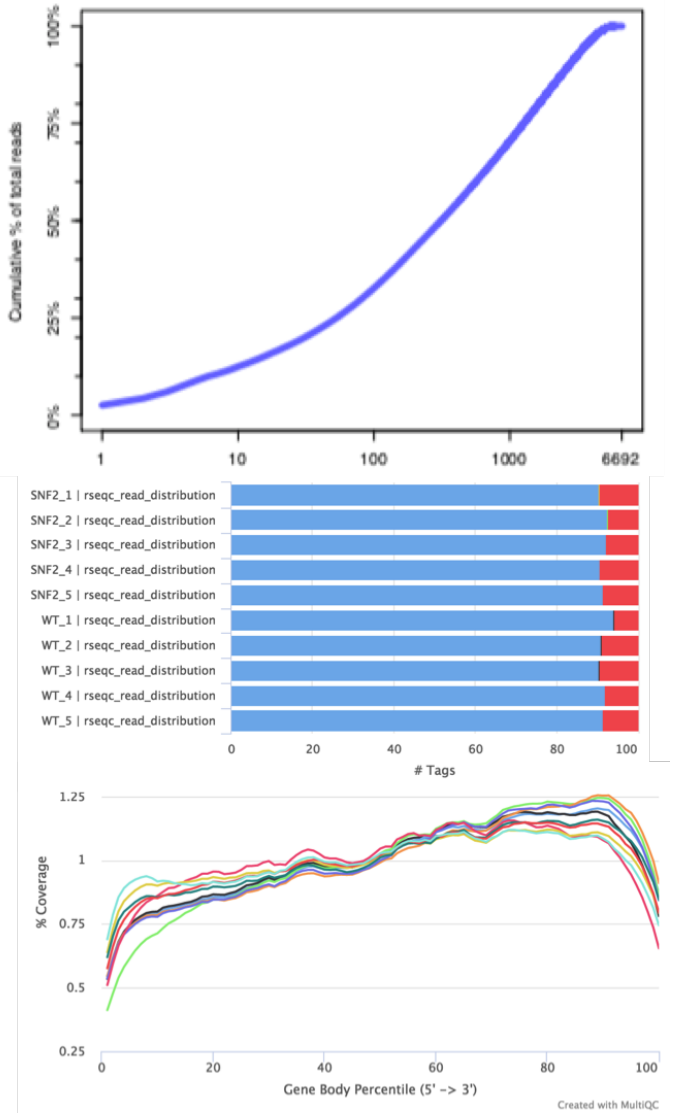






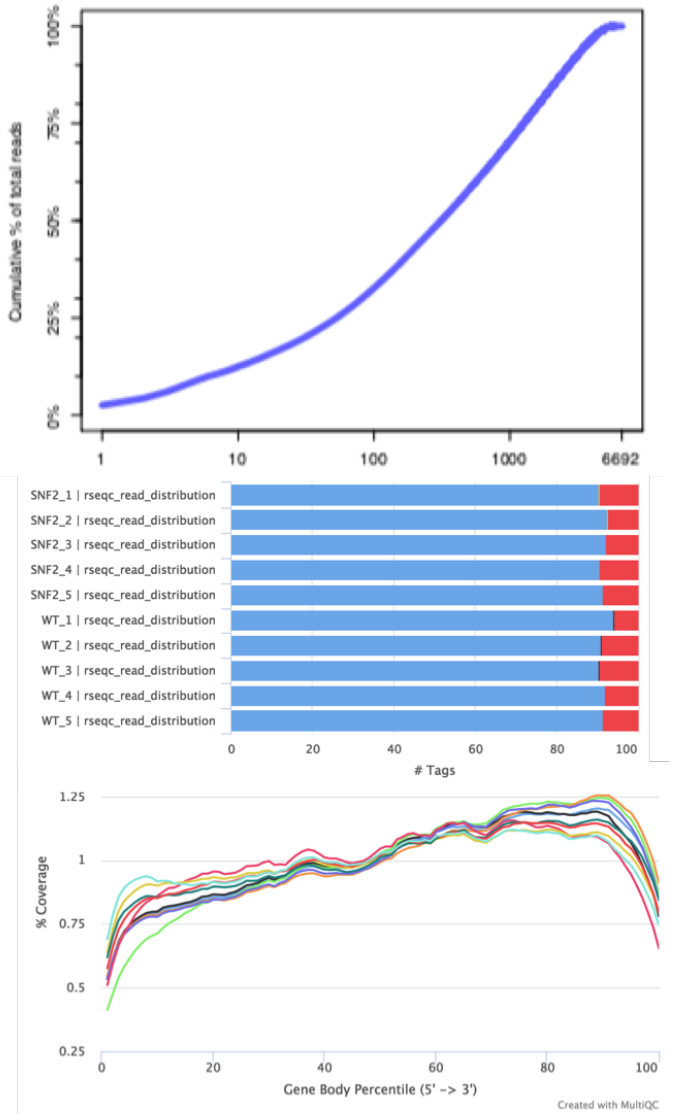
Typical biases of aligned reads of RNA-seq

- lack of **gene diversity**: dominance of rRNAs, tRNAs (and/or other highly abundant transripts)
  - should be visible in `FastQC` results already
- **read distribution**
  - high **intron** coverage: incomplete poly(A) enrichment
  - many **intergenic** reads: gDNA contamination
- **gene body coverage**
  - **3' bias**: RNA degradation (and indicator of poly(A) enrichment)



Typical biases of aligned reads of RNA-seq

- lack of **gene diversity**: dominance of rRNAs, tRNAs (and/or other highly abundant transripts)
  - should be visible in `FastQC` results already
- **read distribution**
  - high **intron** coverage: incomplete poly(A) enrichment
  - many **intergenic** reads: gDNA contamination
- **gene body coverage**
  - **3' bias**: RNA degradation (and indicator of poly(A) enrichment)



RSeQC package

```
$ spack find | egrep -i rseqc
py-rseqc@2.6.4
$ spack load -r py-rseqc@2.6.4 # note the -r to load all dependencies
# for this python-based tool
```

- publication: Wang et al. [2012]
- http://rseqc.sourceforge.net contains the documentation
- see Table 11 of the RNA-seq workshop for a list of its scripts
  - the ones we use most often are are `read_distribution` and `geneBody_coverage.py`
- commands are not well standardized
  - e.g. sometimes the results are just printed to the screen, sometimes it generates a result file silently, sometimes you need to define a file name via `-o`
- result files are not well standardized, either
  - from text output to R scripts to PDF documents

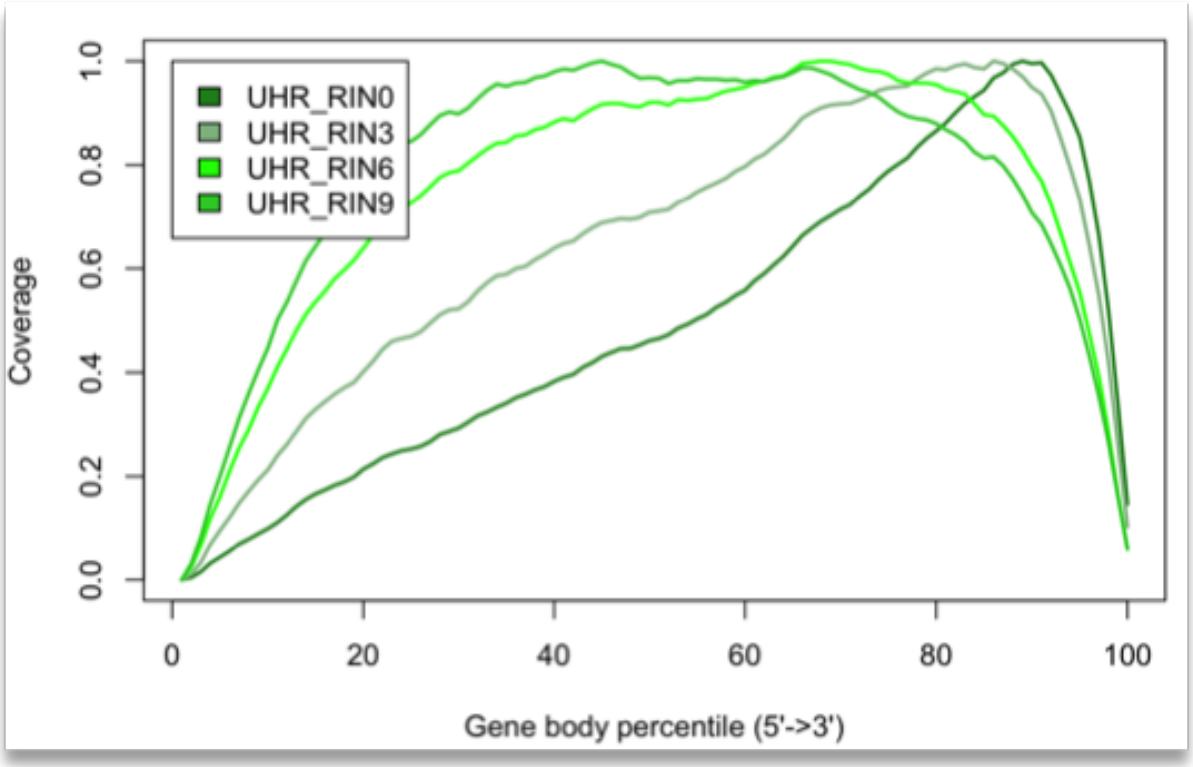
RSeQC: Read distribution

How many reads fall into exons? Based on annotation file (BED!)

```
$ for SAMPLE in WT_1 WT_2 WT_3 WT_4 WT_5 SNF2_1 SNF2_2 SNF2_3 SNF2_4 SNF2_5
 do
 read_distribution.py -i bams/${SAMPLE}*.bam
 -r ../RNA-seq/refGenome_S_cerevisiae/sacCer3.bed > \
 ${SAMPLE}/rseqc_read_distribution.out
 done

$ head -n10 WT_1/rseqc_read_distribution.out
Total Reads                     1049466
Total Tags                      1059871
Total Assigned Tags             992608
=====================================================================
Group                   Total_bases             Tag_count               Tags/Kb
CDS_Exons               8832031                 990363                  112.13
5'UTR_Exons             0                       0                       0.00
3'UTR_Exons             0                       0                       0.00
Introns                 69259                   630                     9.10
TSS_up_1kb              2421198                 1260                    0.52
```

RSeQC: Gene body coverage

```
$ for SAMPLE in WT_1 WT_2 WT_3 WT_4 WT_5 SNF2_1 SNF2_2 SNF2_3 SNF2_4 SNF2_5
   do
    geneBody_coverage.py -i bams/${SAMPLE}*.bam \
     -r ../RNA-seq/refGenome_S_cerevisiae/sacCer3.bed \
     -o ${SAMPLE}/rseqc_geneBody_coverage.out &
 done
```



QoRTs – an alternative to RSeQC

```
$ spack find | egrep -i qorts
$ spack load qorts@1.2.42
# we need the location of the java executable
$ QORTS_LOC=`spack location -i qorts`

# run QoRTs in summary mode
$ for SAMPLE in WT_1 WT_2 WT_3 WT_4 WT_5 SNF2_1 SNF2_2 SNF2_3 SNF2_4 SNF2_5
 do
   java -Xmx4G -jar ${QORTS_LOC}/bin/QoRTs.jar QC --singleEnded
    --generatePdfReport \
    bams/${SAMPLE}*.bam  \
   ../RNA-seq/refGenome_S_cerevisiae/sacCer3.gtf $SAMPLE
done
```

- more convenient and standardized usage than RSeQC
- offers **gene diversity** plot and more fine-grained plots where genes are stratified by expression strength [Hartley and Mullikin, 2015]
- will bundle numerous analyses in one PDF and allows for direct cross-comparisons, but MultiQC doesn't handle it very robustly

Summary of RNA-seq alignment QC

- **raw reads QC (fastq)**
  - adapter/primer/other contaminating and over-represented sequences
  - sequencing quality
  - GC distributions
  - duplication levels

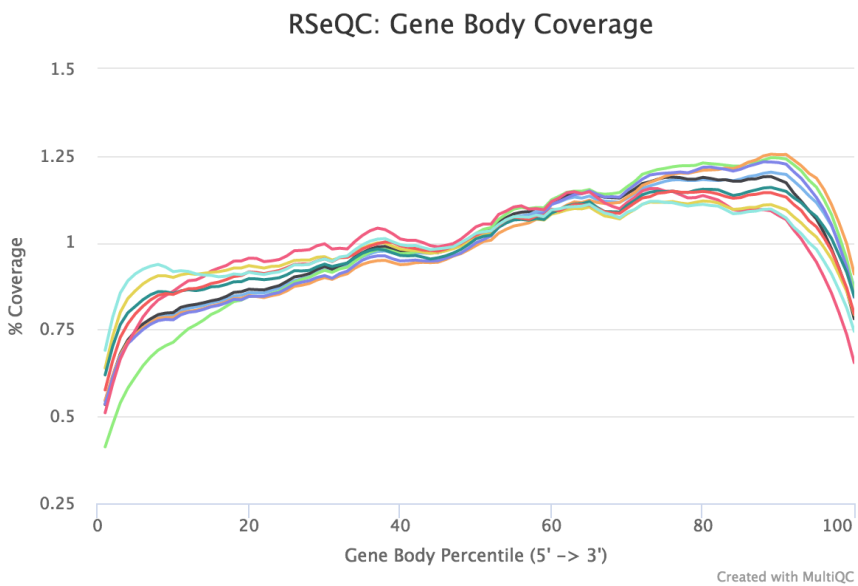
FastQC
(QoRTs)

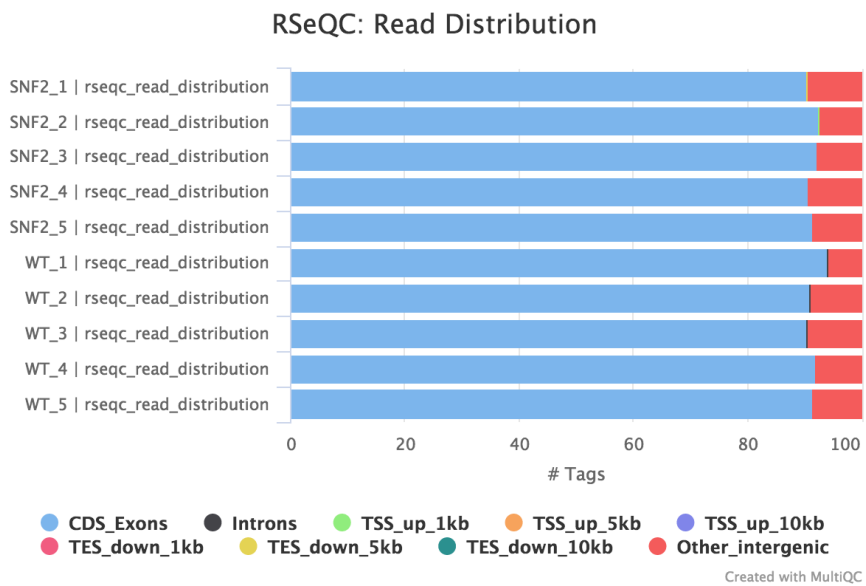
- **aligned reads QC (bam)**
  - % (uniquely) aligned reads
  - % exonic vs. intronic/intergenic
  - gene diversity
  - gene body coverage

aligner's log files
samtools flagstat
RSeQC
QoRTs
MultiQC
…

Summary of RNA-seq alignment QC

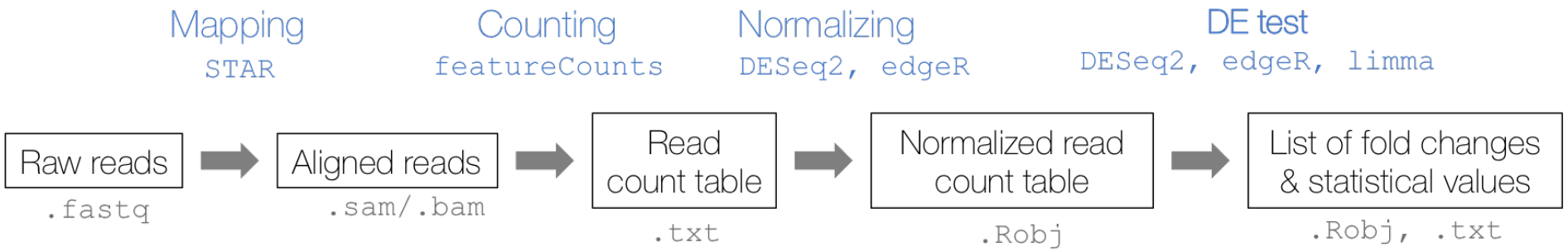
1. Did you capture a diverse set of mRNAs? (or RNAs of the type that you expect)?
2. Are the gene bodies covered similarly across different samples?
3. Is there evidence for contaminations, either from highly abundant, irrelevant transcripts or from genomic DNA?



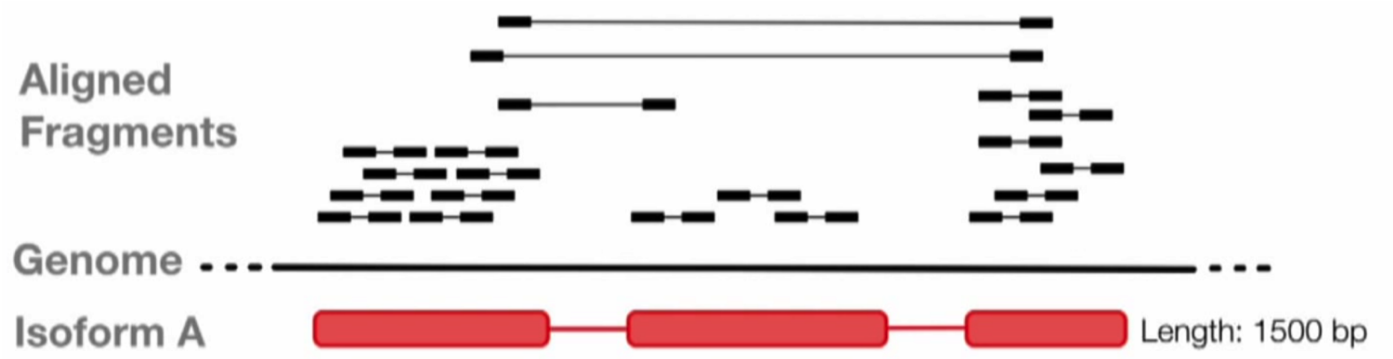


Quantification of gene expression - Part II

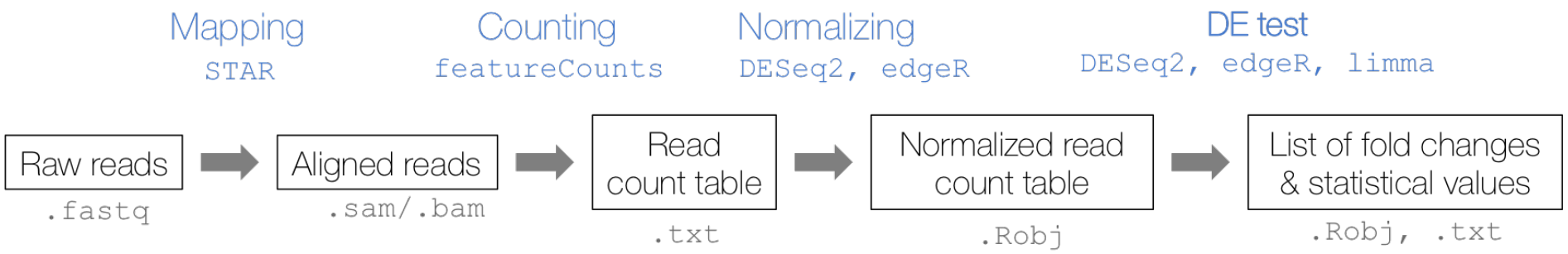
Quantification of gene expression



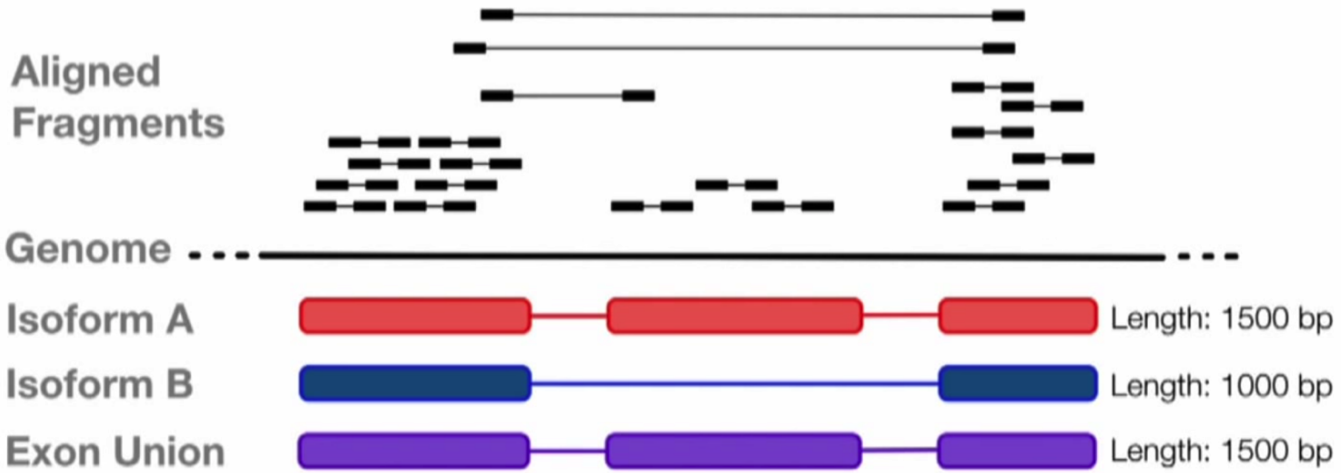
1. **Align**
   - with splice-aware alignment tools! e.g. STAR
2. **Count** reads that overlap with annotated genes



Quantification of gene expression



1. **Align**
   - with splice-aware alignment tools! e.g. STAR
2. **Count** reads that overlap with annotated genes
   - complicated by alternative isoforms: **genes != transcripts**



Different philosophies of expression quantification

- (splice-aware) **alignment** followed by **counting** of reads overlapping with a **gene**
  - "traditional" way of obtaining expression values per gene
  - `STAR` + `featureCounts` + normalizations
- (splice-aware) **alignment** followed by identification of the minimal number of **transcripts** that are supported by the reads aligning to a given locus
  - `TopHat` + `Cufflinks` (DO NOT USE THIS!)
- direct **transcript abundance estimation without alignment** by determining which known transcripts are compatible with a given pool of sequenced reads
  - `kallisto`, `salmon`, `sailfish`, `RSEM`

Different philosophies of expression quantification

- (splice-aware) **alignment** followed by **counting** of reads overlapping with a **gene**
  - "traditional" way of obtaining expression values per gene
  - `STAR` + `featureCounts` + normalizations
- (splice-aware) **alignment** followed by identification of the minimal number of **transcripts** that are supported by the reads aligning to a given locus
  - `TopHat` + `Cufflinks` (DO NOT USE THIS!)
- direct **transcript abundance estimation without alignment** by determining which known transcripts are compatible with a given pool of sequenced reads
  - `kallisto`, `salmon`, `sailfish`, `RSEM`

Different philosophies of expression quantification

- (splice-aware) **alignment** followed by **counting** of reads overlapping with a **gene**
  - "traditional" way of obtaining expression values per gene
  - `STAR` + `featureCounts` + normalizations
- (splice-aware) **alignment** followed by identification of the minimal number of **transcripts** that are supported by the reads aligning to a given locus
  - `TopHat` + `Cufflinks` (DO NOT USE THIS!)
- direct **transcript abundance estimation without alignment** by determining which known transcripts are compatible with a given pool of sequenced reads
  - `kallisto`, `salmon`, `sailfish`, `RSEM`

Different philosophies of expression quantification

- (splice-aware) **alignment** followed by **counting** of reads overlapping with a **gene**
  - "traditional" way of obtaining expression values per gene
  - `STAR` + `featureCounts` + normalizations
- (splice-aware) **alignment** followed by identification of the minimal number of **transcripts** that are supported by the reads aligning to a given locus
  - `TopHat` + `Cufflinks` (DO NOT USE THIS!)
- direct **transcript abundance estimation without alignment** by determining which known transcripts are compatible with a given pool of sequenced reads
  - `kallisto`, `salmon`, `sailfish`, `RSEM`

Different philosophies of expression quantification

- (splice-aware) **alignment** followed by **counting** of reads overlapping with a **gene**
  - "traditional" way of obtaining expression values per gene
  - `STAR` + `featureCounts` + normalizations
- (splice-aware) **alignment** followed by identification of the minimal number of **transcripts** that are supported by the reads aligning to a given locus
  - `TopHat` + `Cufflinks` (DO NOT USE THIS!)
- direct **transcript abundance estimation without alignment** by determining which known transcripts are compatible with a given pool of sequenced reads
  - `kallisto`, `salmon`, `sailfish`, `RSEM`

Different philosophies of expression quantification

- (splice-aware) **alignment** followed by **counting** of reads overlapping with a **gene**
  - "traditional" way of obtaining expression values per gene
  - `STAR` + `featureCounts` + normalizations
- (splice-aware) **alignment** followed by identification of the minimal number of **transcripts** that are supported by the reads aligning to a given locus
  - `TopHat` + `Cufflinks` (DO NOT USE THIS!)
- direct **transcript abundance estimation without alignment** by determining which known transcripts are compatible with a given pool of sequenced reads
  - `kallisto`, `salmon`, `sailfish`, `RSEM`

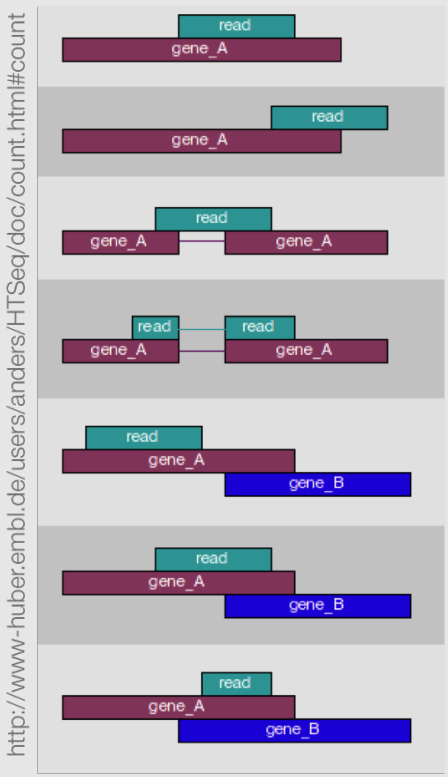
Different philosophies of expression quantification

- (splice-aware) **alignment** followed by **counting** of reads overlapping with a **gene**
  - "traditional" way of obtaining expression values per gene
  - `STAR` + `featureCounts` + normalizations
- (splice-aware) **alignment** followed by identification of the minimal number of **transcripts** that are supported by the reads aligning to a given locus
  - `TopHat` + `Cufflinks` (DO NOT USE THIS!)
- direct **transcript abundance estimation without alignment** by determining which known transcripts are compatible with a given pool of sequenced reads
  - `kallisto`, `salmon`, `sailfish`, `RSEM`

Different philosophies of expression quantification

1. Counting read-gene overlaps with featureCounts

- **features** = single rows within the GTF file, e.g. exons
- **meta-features** = how single rows may be grouped together, e.g. by `transcript-id` or `gene-id` (define via `-g` option)
- see http://bioinf.wehi.edu.au/featureCounts/ and Chapter 7 of SubreadUsersGuide.pdf for details!



featureCounts will use read-gene overlaps as small as 1 bp

multi-overlap reads will be discarded

Different philosophies of expression quantification

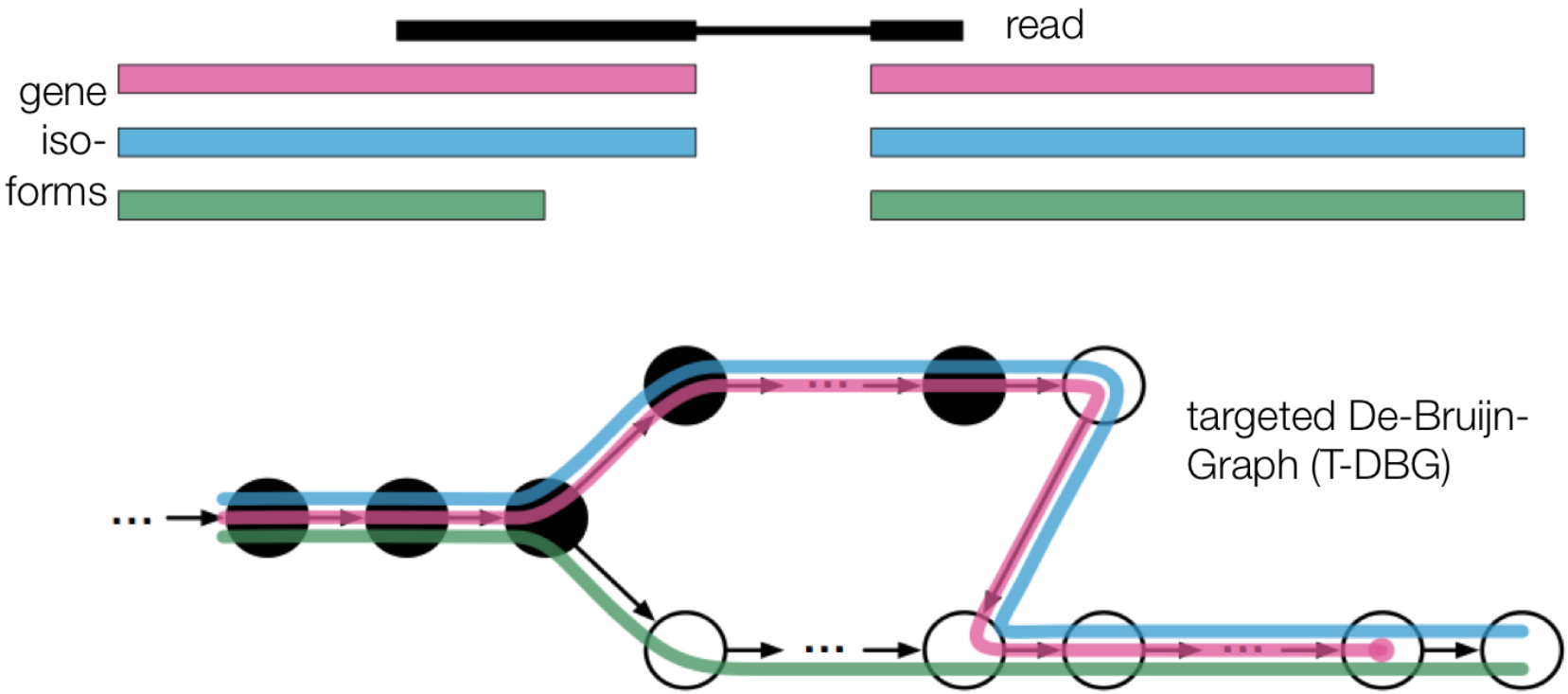
1. Counting read-gene overlaps with featureCounts

Let's do it!

- **Count the reads that overlap with genes (union of all exons per gene).**
- Note: `featureCounts` is part of the `subread` package.

Different philosophies of expression quantification

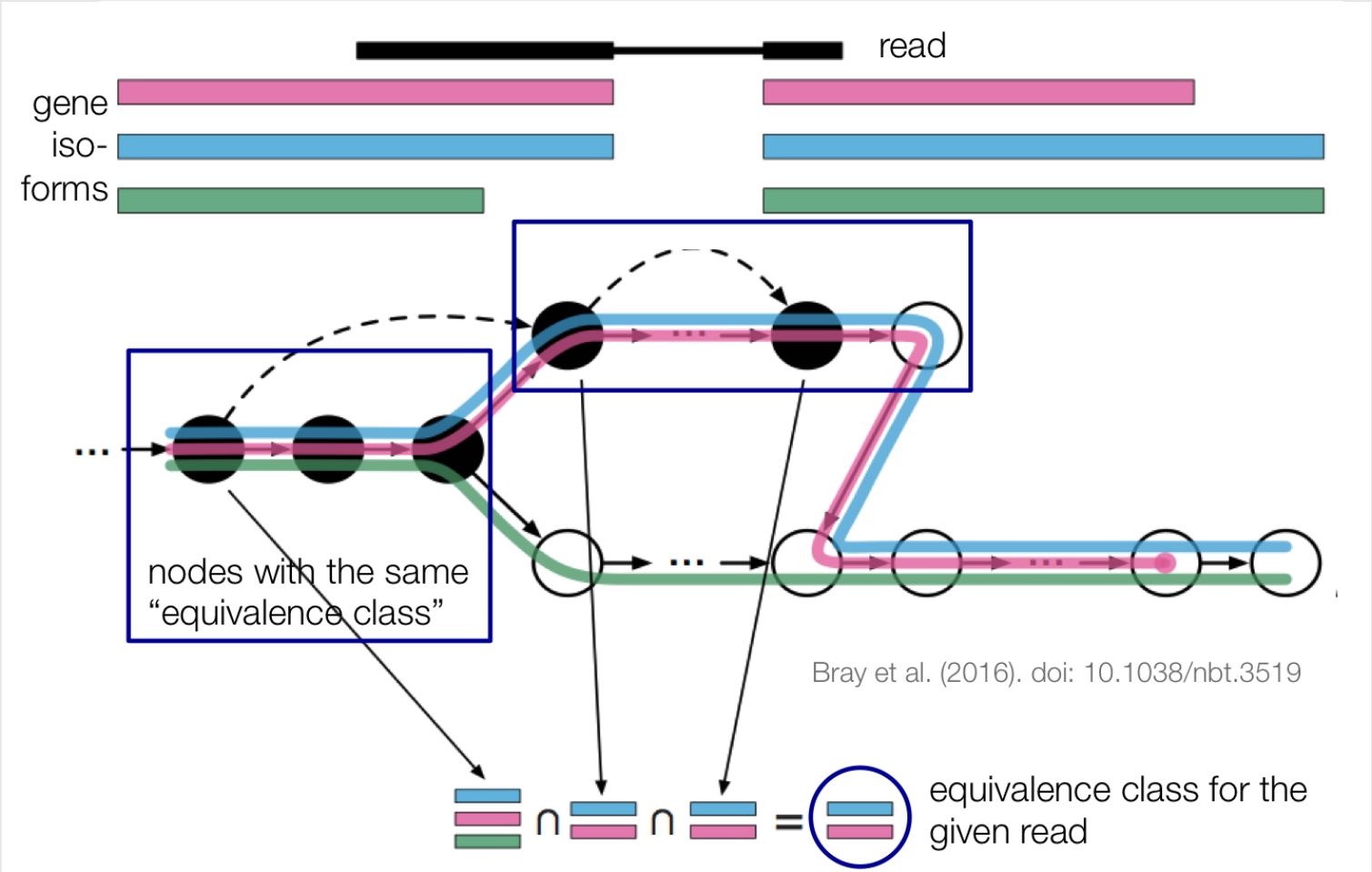
2. Transcript abundance estimation via pseudoalignment



Bray et al. (2016). doi: 10.1038/nbt.3519
http://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html

Different philosophies of expression quantification

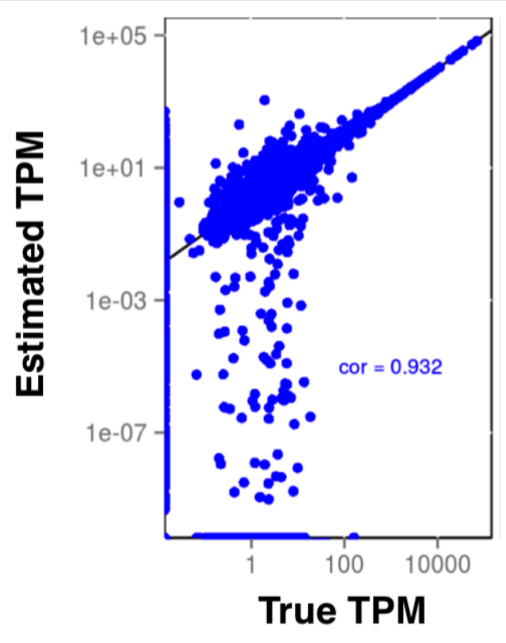
2. Transcript abundance estimation via pseudoalignment

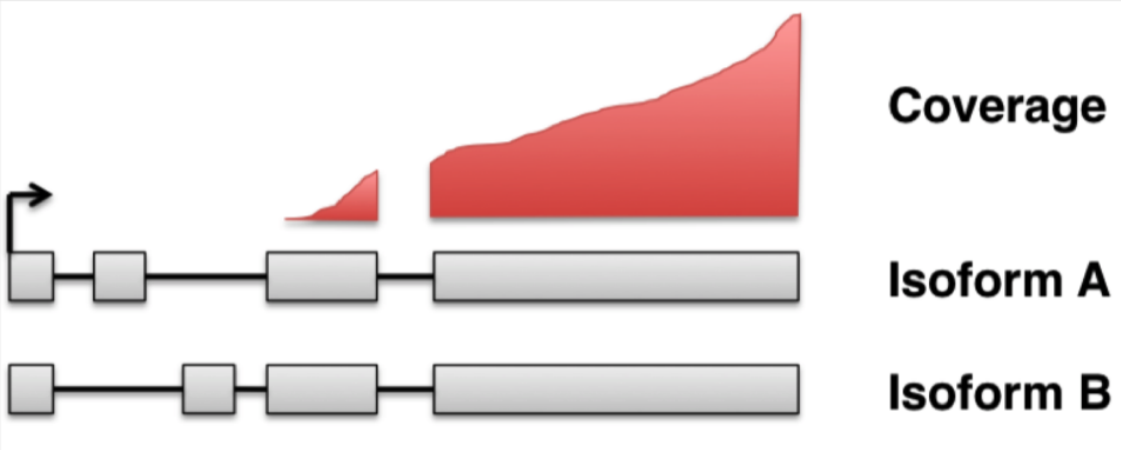


Different philosophies of expression quantification

2. Transcript abundance estimation via pseudoalignment – CAUTION!

- abundance estimates for lowly expressed transcripts are highly variable
- problem when coverage of the region defining an isoform is low





For very similar transcripts, collapsing all abundances per gene into a **gene-centric measure** is more robust and accurate. [Soneson et al., 2015]

Comparing "read count overlaps" to "pseudoalignments"

	Traditional	Pseudoalignment
Ex. workflow:	STAR + featureCounts	kallisto
Read mapping based on:	Where does a read match best?	Which equivalence class (EC) does a read match best?
Reference:	Genome seq. + exon boundaries	cDNA sequences
Mapping result:	Genome coordinates (BAM)	Read-EC table[3]
Expression quantification:	Counting how many reads *overlap* a gene[4].	Summing up the reads assigned to each EC.
Output:	Read counts (integers)	Estimated transcript abundances
Speed:	+++	++++

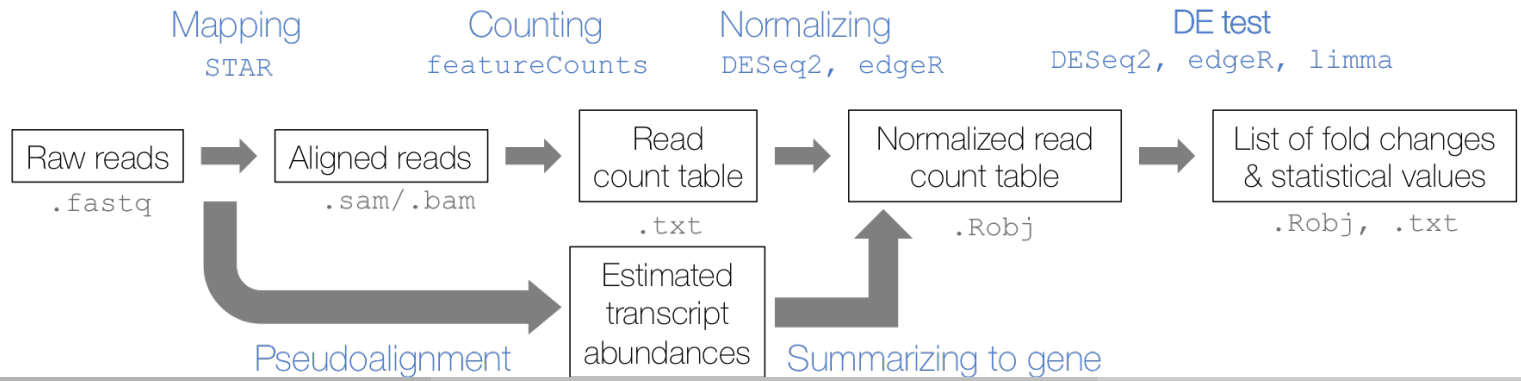
[3] As simple a table as it gets.

[4] The read sequence is irrelevant at this point.

General bioinformatics workflow – updated

Understand your null hypothesis!(See Soneson et al. [2015], Love et al. [2018])

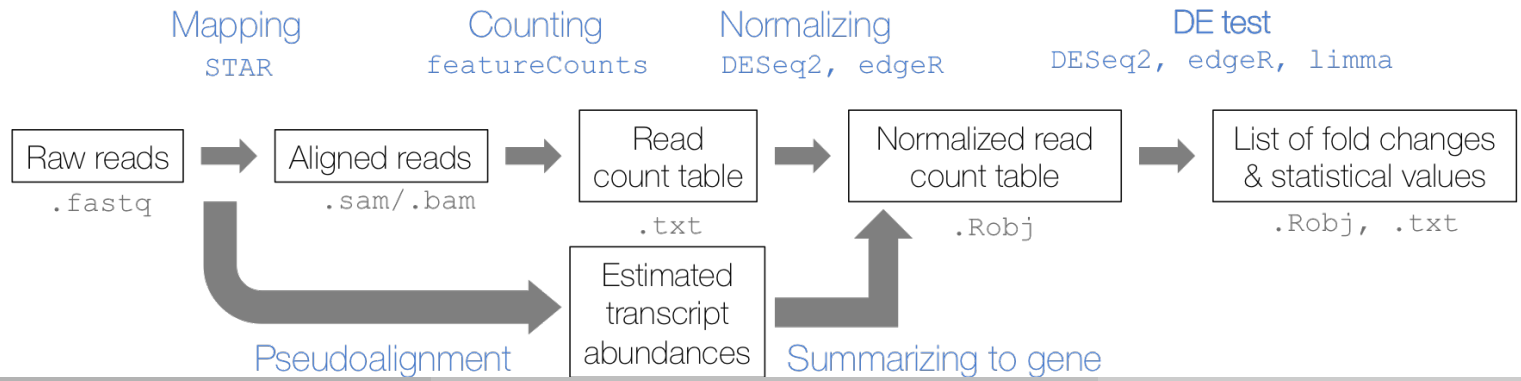
- **DGE**: Differential **Gene** Expression
  - Has the total ouput of a gene changed?
  - input for the statistical testing: (estimated) counts per gene used by `DESeq2/edgeR/limma`
- **DTU**: Differential **Transcript Usage**
  - Has the isoform composition for a given gene changed? I.e. are there different dominant isoforms depending on the condition?
  - common when comparing different cell types (incl. healthy vs. cancer)
  - input for the statistical testing: (estimated) counts per transcript used by `DEXSeq` (!)



General bioinformatics workflow – updated

Understand your null hypothesis!(See Soneson et al. [2015], Love et al. [2018])

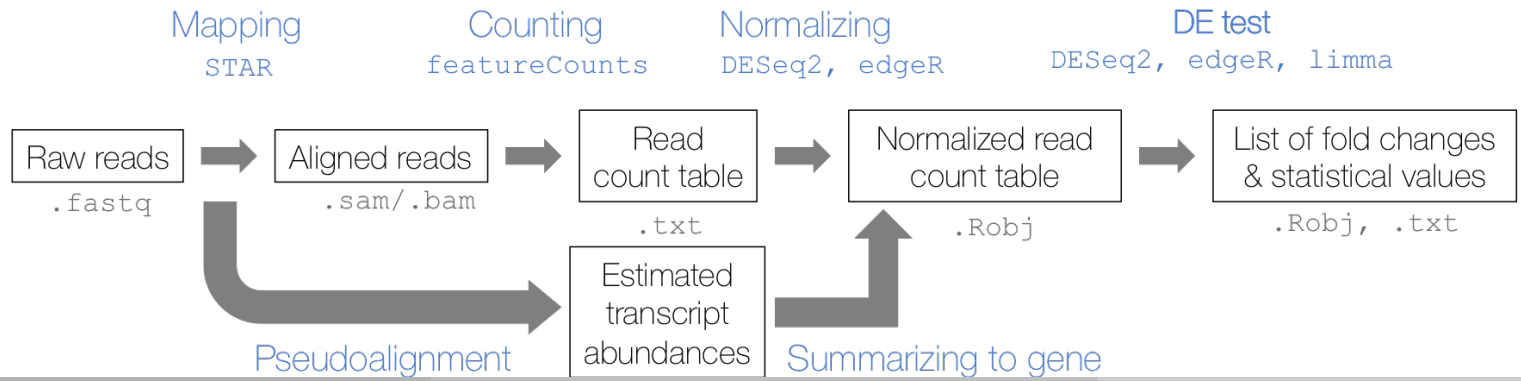
- **DGE**: Differential **Gene** Expression
  - Has the total ouput of a gene changed?
  - input for the statistical testing: (estimated) counts per gene used by `DESeq2/edgeR/limma`
- **DTU**: Differential **Transcript Usage**
  - Has the isoform composition for a given gene changed? I.e. are there different dominant isoforms depending on the condition?
  - common when comparing different cell types (incl. healthy vs. cancer)
  - input for the statistical testing: (estimated) counts per transcript used by `DEXSeq` (!)



General bioinformatics workflow – updated

Understand your null hypothesis!(See Soneson et al. [2015], Love et al. [2018])

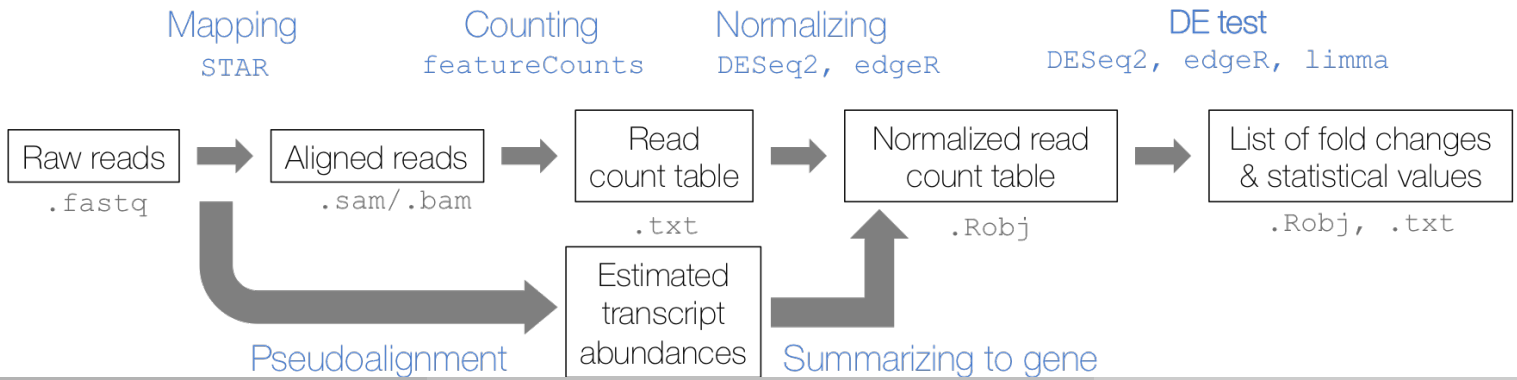
- **DGE**: Differential **Gene** Expression
  - Has the total ouput of a gene changed?
  - input for the statistical testing: (estimated) counts per gene used by `DESeq2/edgeR/limma`
- **DTU**: Differential **Transcript Usage**
  - Has the isoform composition for a given gene changed? I.e. are there different dominant isoforms depending on the condition?
  - common when comparing different cell types (incl. healthy vs. cancer)
  - input for the statistical testing: (estimated) counts per transcript used by `DEXSeq` (!)



General bioinformatics workflow – updated

Understand your null hypothesis!(See Soneson et al. [2015], Love et al. [2018])

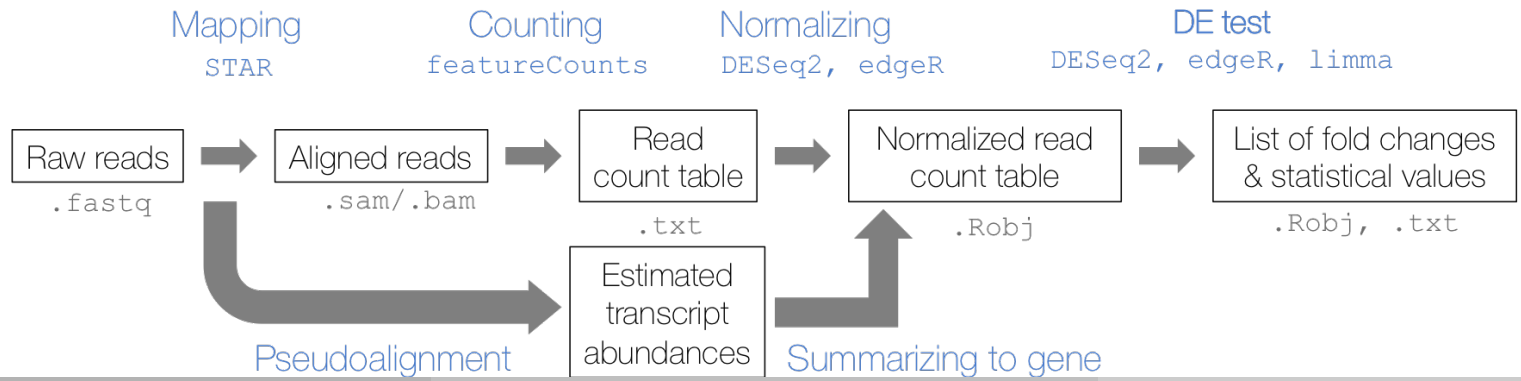
- **DGE**: Differential **Gene** Expression
  - Has the total ouput of a gene changed?
  - input for the statistical testing: (estimated) counts per gene used by `DESeq2/edgeR/limma`
- **DTU**: Differential **Transcript Usage**
  - Has the isoform composition for a given gene changed? I.e. are there different dominant isoforms depending on the condition?
  - common when comparing different cell types (incl. healthy vs. cancer)
  - input for the statistical testing: (estimated) counts per transcript used by `DEXSeq` (!)



General bioinformatics workflow – updated

Understand your null hypothesis!(See Soneson et al. [2015], Love et al. [2018])

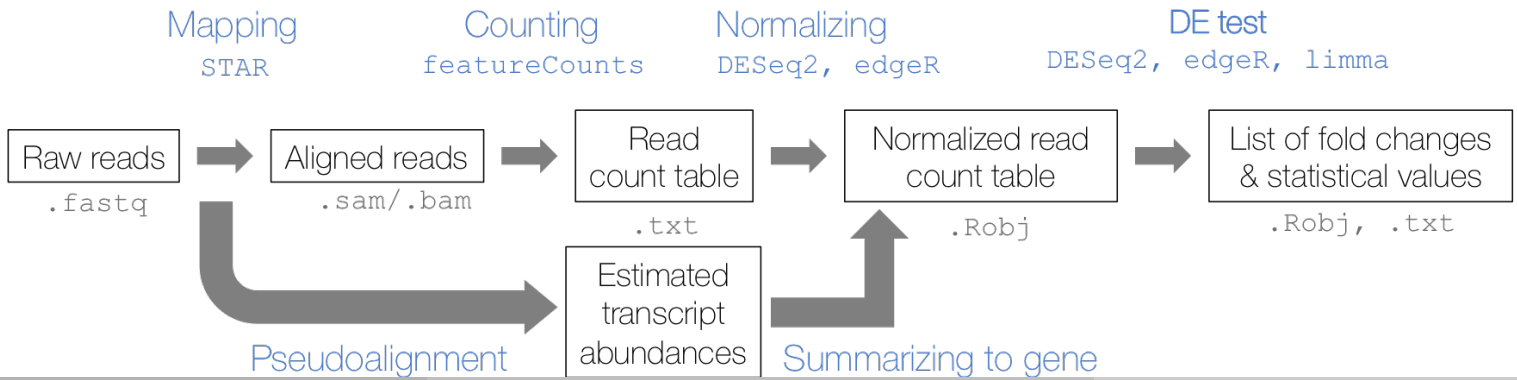
- **DGE**: Differential **Gene** Expression
  - Has the total ouput of a gene changed?
  - input for the statistical testing: (estimated) counts per gene used by `DESeq2/edgeR/limma`
- **DTU**: Differential **Transcript Usage**
  - Has the isoform composition for a given gene changed? I.e. are there different dominant isoforms depending on the condition?
  - common when comparing different cell types (incl. healthy vs. cancer)
  - input for the statistical testing: (estimated) counts per transcript used by `DEXSeq` (!)



General bioinformatics workflow – updated

Understand your null hypothesis!(See Soneson et al. [2015], Love et al. [2018])

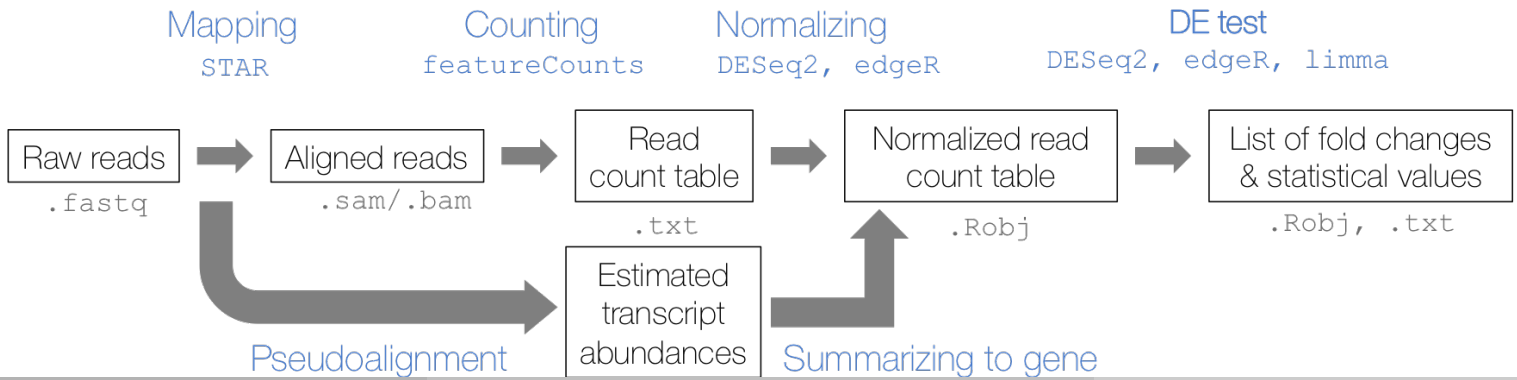
- **DGE**: Differential **Gene** Expression
  - Has the total ouput of a gene changed?
  - input for the statistical testing: (estimated) counts per gene used by `DESeq2/edgeR/limma`
- **DTU**: Differential **Transcript Usage**
  - Has the isoform composition for a given gene changed? I.e. are there different dominant isoforms depending on the condition?
  - common when comparing different cell types (incl. healthy vs. cancer)
  - input for the statistical testing: (estimated) counts per transcript used by `DEXSeq` (!)



General bioinformatics workflow – updated

Understand your null hypothesis!(See Soneson et al. [2015], Love et al. [2018])

- **DGE**: Differential **Gene** Expression
  - Has the total ouput of a gene changed?
  - input for the statistical testing: (estimated) counts per gene used by `DESeq2/edgeR/limma`
- **DTU**: Differential **Transcript Usage**
  - Has the isoform composition for a given gene changed? I.e. are there different dominant isoforms depending on the condition?
  - common when comparing different cell types (incl. healthy vs. cancer)
  - input for the statistical testing: (estimated) counts per transcript used by `DEXSeq` (!)



Normalization

Read counts are influenced by numerous factors, not just expression strength

Raw counts[5]: number of reads (or fragments) overlapping with the union of exons of a gene.

The raw counts are not just a reflection of the actual number of captured transcripts!

strongly influenced by:

- gene length
- transcript sequence (% GC)
- sequencing depth
- expression of all other genes in the same sample

may cause variations for **different genes** expressed at the same level

may cause variations for the **same gene** in different samples

[5] includes "estimated" gene counts from pseudoaligners

Different units for expression values

- **Raw counts**: number of reads/ fragments overlapping with the union of exons of a gene

$$X_i$$

- **[RF]PKM**: Reads/Fragments per Kilobase of gene per Million reads mapped – AVOID!

$$RPKM_i = \frac{X_i}{(\frac{l_i}{10^3})(\frac{N}{10^6})}$$

gene length seq. depth

- **TPM**: Transcripts Per Million

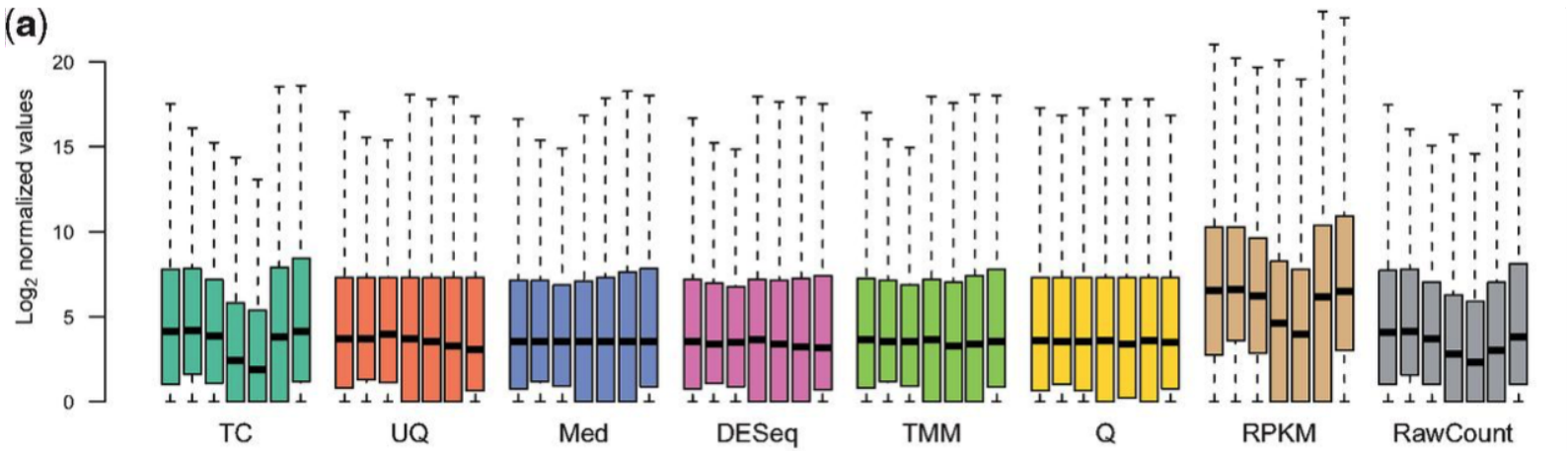
$$TPM_i = \frac{X_i}{l_i} * \frac{1}{\sum_j \frac{X_j}{l_k}} * 10^6$$

gene read counts per bp

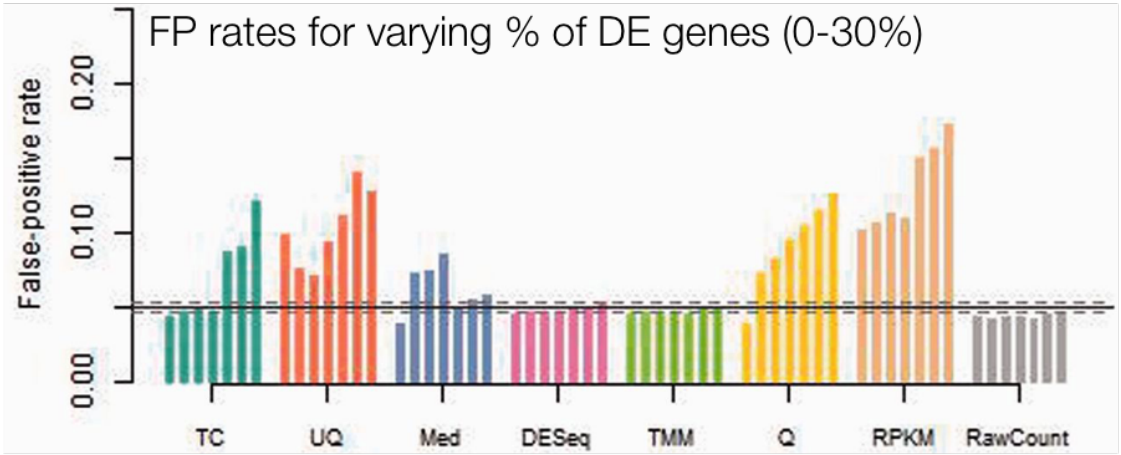
all gene counts over *all* gene bp

- **rlog**: log2-transformed count data normalized for small counts and library size (DESeq2)

Effects of normalization methods on FC calculation and DGE analysis



Dillies et al.(2012). Briefings in Bioinformatics. doi:10.1093/bib/bbs046



Avoid [RF]PKM and **total read count** normalization for DGE!

if you need normalized expression values, e.g. for exploratory plots, use **TPM or DESeq2's rlog**

Working with read counts

- Download the featureCounts results to your laptop.
- Read the featureCounts results into R.
- Let's normalize and explore!

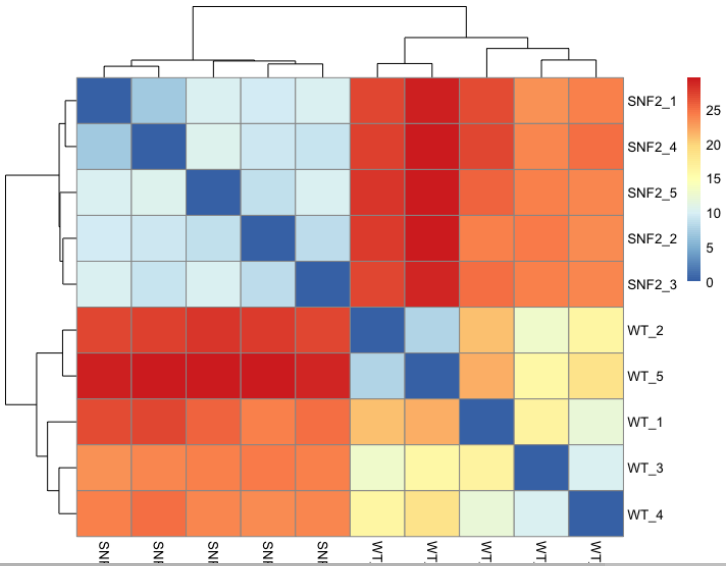
Exploratory analyses

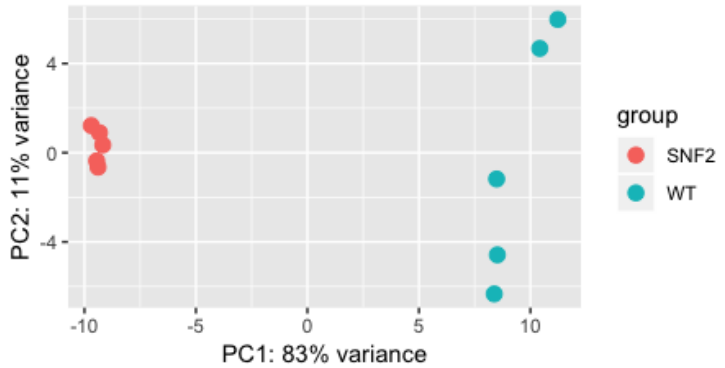
Exploratory analyses

CAVE

Exploratory analyses **do not test a null hypothesis**! They are meant to familiarize yourself with the data!

- correlations of gene expression
- (hierarchical) clustering
- dimensionality reduction methods, e.g. PCA





Which expression units should be used?

Exploratory analyses work better on **normalized and transformed** read counts because they are:

- strongly influenced by
  - gene length
  - sequencing depth — DESeq's size factor normalization
  - expression of all other genes in the same sample

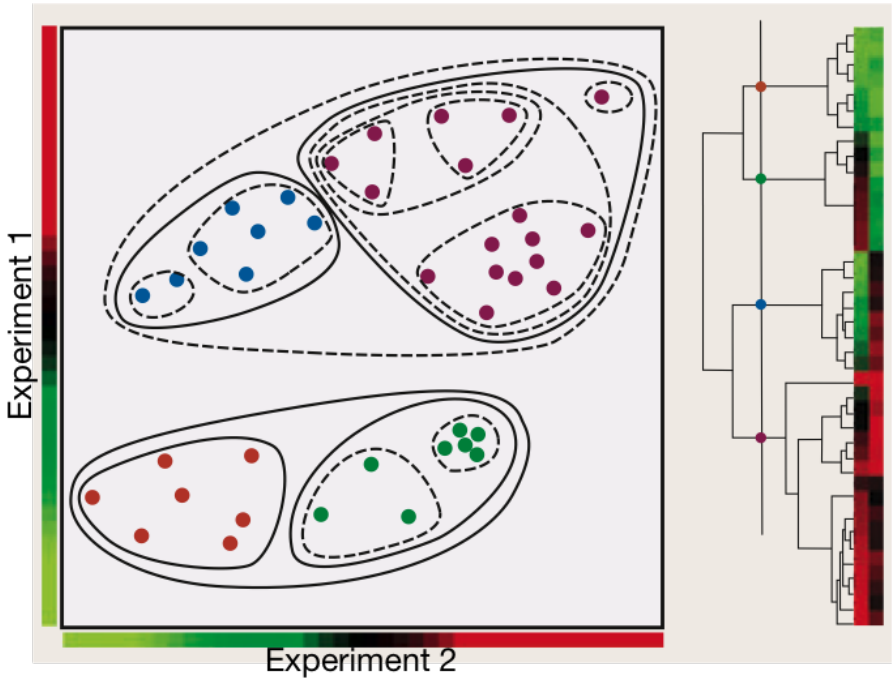
- large dynamic range
- discrete values

hetero-skedasticity

log transformation and variance stabilization (DESeq's rlog())

Hierarchical clustering

Goal: partition the objects into homogeneous groups, such that the within-group similarities are large.

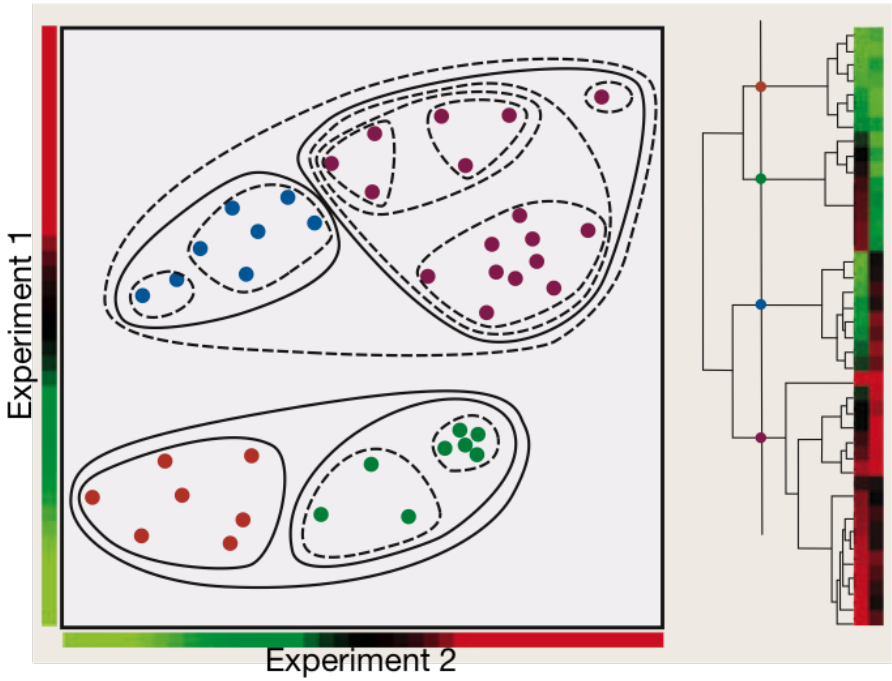


- Result: **dendrogram**
  - clustering is obtained by **cutting the dendrogram** at the desired level
- Similarity measure
  - Euclidean
  - Pearson
- Distance measure
  - Complete: largest distance
  - Average: average distance

single-sample (or single-gene) clusters are successively joined, starting with the least dissimilar two samples

Hierarchical clustering

Goal: partition the objects into homogeneous groups, such that the within-group similarities are large.

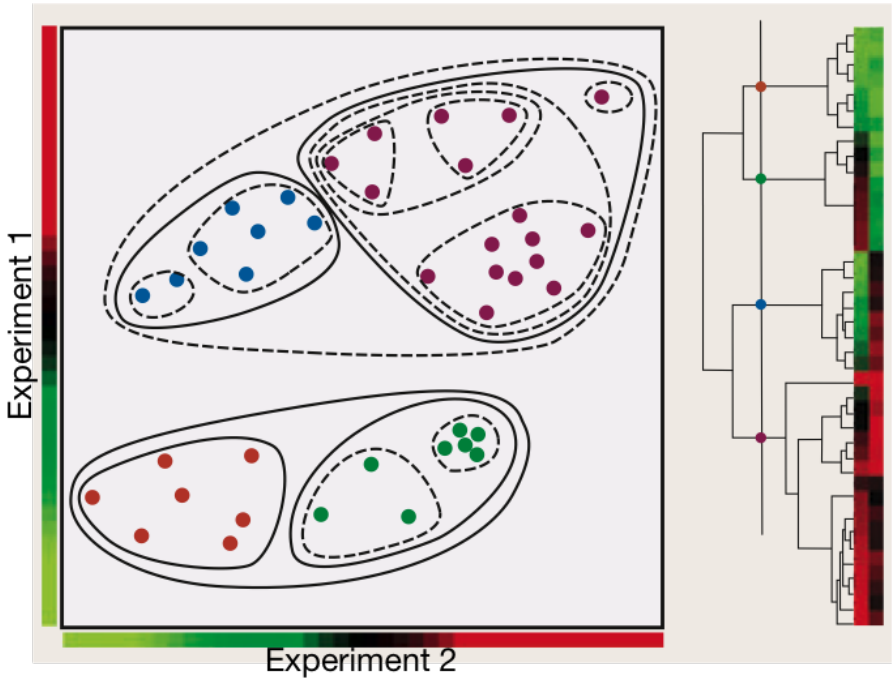


- Result: **dendrogram**
  - clustering is obtained by **cutting the dendrogram** at the desired level
- Similarity measure
  - Euclidean
  - Pearson
- Distance measure
  - Complete: largest distance
  - Average: average distance

single-sample (or single-gene) clusters are successively joined, starting with the least dissimilar two samples

Hierarchical clustering

Goal: partition the objects into homogeneous groups, such that the within-group similarities are large.

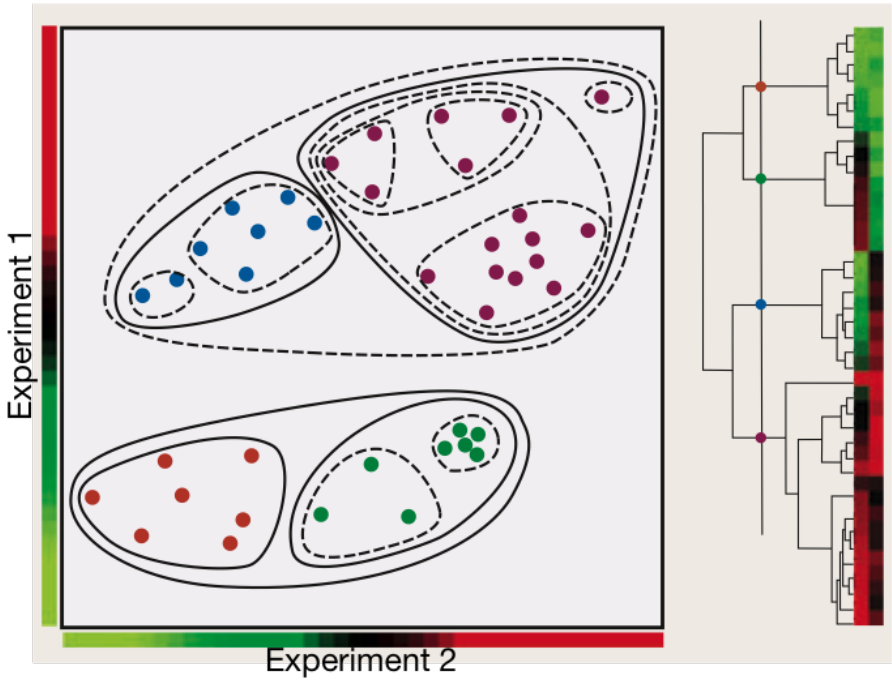


- Result: **dendrogram**
  - clustering is obtained by **cutting the dendrogram** at the desired level
- Similarity measure
  - Euclidean
  - Pearson
- Distance measure
  - Complete: largest distance
  - Average: average distance

single-sample (or single-gene) clusters are successively joined, starting with the least dissimilar two samples

Hierarchical clustering

Goal: partition the objects into homogeneous groups, such that the within-group similarities are large.

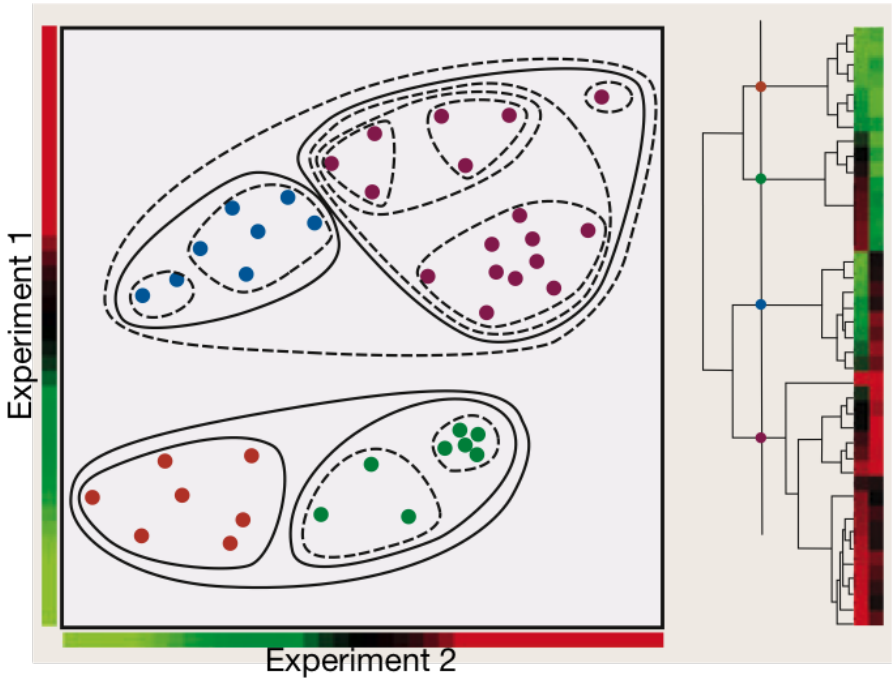


- Result: **dendrogram**
  - clustering is obtained by **cutting the dendrogram** at the desired level
- Similarity measure
  - Euclidean
  - Pearson
- Distance measure
  - Complete: largest distance
  - Average: average distance

single-sample (or single-gene) clusters are successively joined, starting with the least dissimilar two samples

Hierarchical clustering

Goal: partition the objects into homogeneous groups, such that the within-group similarities are large.

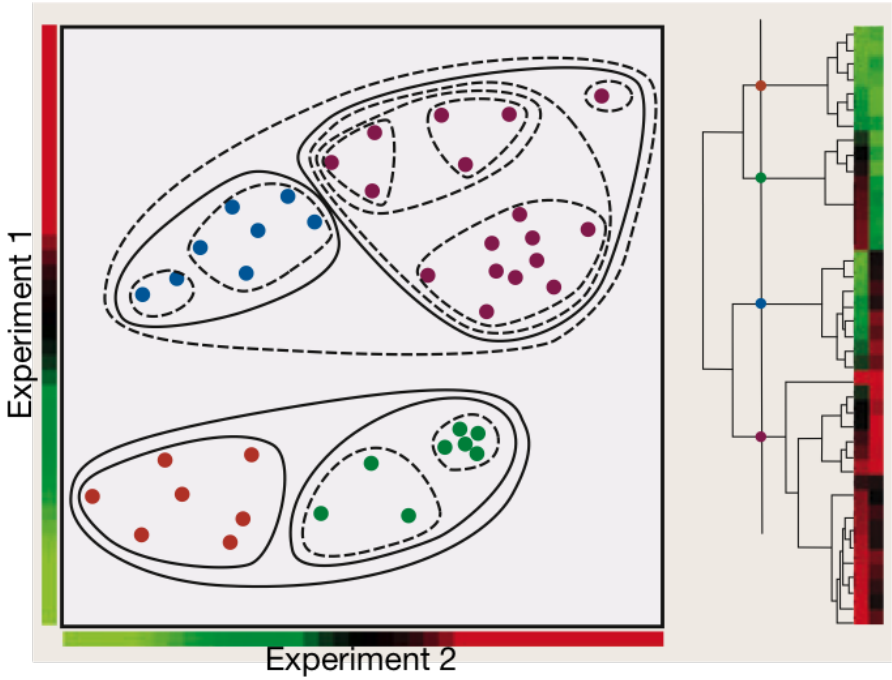


- Result: **dendrogram**
  - clustering is obtained by **cutting the dendrogram** at the desired level
- Similarity measure
  - Euclidean
  - Pearson
- Distance measure
  - Complete: largest distance
  - Average: average distance

single-sample (or single-gene) clusters are successively joined, starting with the least dissimilar two samples

Hierarchical clustering

Goal: partition the objects into homogeneous groups, such that the within-group similarities are large.



- Result: **dendrogram**
  - clustering is obtained by **cutting the dendrogram** at the desired level
- Similarity measure
  - Euclidean
  - Pearson
- Distance measure
  - Complete: largest distance
  - Average: average distance

single-sample (or single-gene) clusters are successively joined, starting with the least dissimilar two samples

PCA: Principal component analyses

starting point: matrix with expression values per gene and sample, e.g. 7,100 genes x 10 samples

	SNF2_1	SNF2_2	SNF2_3	SNF2_4	SNF2_5	WT_1	WT_2	WT_3	WT_4	WT_5
YDL248W	109	84	100	112	62	47	65	60	95	43
YDL247W.A	0	1	1	0	3	0	0	1	0	0
YDL247W	6	6	1	3	4	2	3	4	7	9
YDL246C	6	6	1	4	4	1	3	2	4	0
YDL245C	1	6	9	5	3	6	2	5	5	6
YDL244W	79	59	49	60	37	9	8	12	30	14

reduced to 2 **principal components** (or more) x 10 samples

	PC1	PC2
SNF2_1	-9.322866	0.8929154
SNF2_2	-9.390920	-0.6478100
SNF2_3	-9.176814	0.3460428
SNF2_4	-9.693035	1.2174519
SNF2_5	-9.450847	-0.3668670
WT_1	8.378671	-6.3321623
WT_2	10.421518	4.6749399
WT_3	8.486379	-1.1793146
WT_4	8.517490	-4.5814481
WT_5	11.230425	5.9762519

- linear combi of optimally weighted observed variables
- the vectors along which the variation between samples is maximal
- their number is $\leq$ number of original variables.

PCA vs. hierarchical clustering

- often similar results because both techniques should capture the most dominant patterns - first principal components should contain the information that are separating different subgroups of the samples from each other
- PCA will always be run on just a subset of the data! (both, genes and samples!)
- clustering will ALWAYS return clusters, PCA may not if the patterns of variation are too random

References

[Ballouz et al., 2018, Chan and Tay, 2018, D'haeseleer, 2005, Dillies et al., 2013, Griffith et al., 2015, Bray et al., 2016, Mignone et al., 2002, Pai and Luca, 2018, van Dijk et al., 2014, Dündar et al., 2018]

References

Sara Ballouz, Alexander Dobin, Thomas R Gingeras, and Jesse Gillis. The fractured landscape of RNA-seq alignment: the default in our STARs. *Nucleic Acids Research*, 46(10):5125–5138, 05 2018. doi: 10.1093/nar/gky325. URL https://dx.doi.org/10.1093/nar/gky325.

Rafal Bartoszewski and Aleksander F. Sikorski. Editorial focus: entering into the non-coding RNA era. *Cellular and Molecular Biology Letters*, 2018. doi: 10.1186/s11658-018-0111-3.

Nicolas L Bray, Harold Pimentel, Pall Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech*, 34(5): 525–527, 2016. doi: 10.1038/nbt.3519.

Jia Jia Chan and Yvonne Tay. Noncoding RNA: RNA regulatory networks in cancer. *International Journal of Molecular Sciences*, 2018. doi: 10.3390/ijms19051310.

Cloelia Dard-Dascot, Delphine Naquin, Yves D'Aubenton-Carafa, Karine Alix, Claude Thermes, and Erwin van Dijk. Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *BMC Genomics*, 2018. doi: 10.1186/s12864-018-4491-6.

Patrik D'haeseleer. How does gene expression clustering work? *Nature Biotechnology*, 23(12):1499–1501, 2005. doi: 10.1038/nbt1205-1499.

Marie Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Nicolas Servant Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom, Mickaël Guedj, and Florence Jaffrézic. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013. doi: 10.1093/bib/bbs046.

F. Dündar, L. Skrabanek, and P. Zumbo. Introduction to differential gene expression analysis using rna-seq, 2018. URL http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf.

Malachi Griffith, Jason R. Walker, Nicholas C. Spies, Benjamin J. Ainscough, and Obi L. Griffith. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLoS Computational Biology*, 11(8), 2015. doi: 10.1371/journal.pcbi.1004393.

Stephen W Hartley and James C Mullikin. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. *BMC Bioinformatics*, 16(1):224, January 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0670-5.

Joshua Z Levin, Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, Andreas Gnirke, and Aviv Regev. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods*, 7(9):709–715, 2010. doi: 10.1038/nmeth.1491.

Michael I Love, Charlotte Soneson, Rob Patro, Kristoffer Vitting-seerup, Alicia Oshlack, and Royal Children. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. *F1000Research*, 7(952), 2018. doi: 10.12688/f1000research.15398.1.

Flavio Mignone, Carmela Gissi, Sabino Liuni, Graziano Pesole, and Others. Untranslated regions of mRNAs. *Genome Biol*, 2002. doi: 10.1186/gb-2002-3-3-reviews0004.

Athma A. Pai and Francesca Luca. Environmental influences on RNA processing: Biochemical, molecular and genetic regulators of cellular response. *Wiley Interdisciplinary Reviews: RNA*, 2018. ISSN 17577012. doi: 10.1002/wrna.1503.

Sven Schuierer, Walter Carbone, Judith Knehr, Virginie Petitjean, Anita Fernandez, Marc Sultan, and Guglielmo Roma. A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples. *BMC Genomics*, 2017. doi: 10.1186/s12864-017-3827-y.

Charlotte Soneson, Michael I. Love, and Mark D. Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4(0):1521, 2015. doi: 10.12688/f1000research.7563.2.

Erwin L van Dijk, Yan Jaszczyszyn, and Claude Thermes. Library preparation methods for next-generation sequencing: tone down the bias. *Experimental cell research*, 322(1):12–20, Mar 2014. doi: 10.1016/j.yexcr.2014.01.008.

Liguo Wang, Shengqin Wang, and Wei Li. RSeQC: Quality control of RNA-seq experiments. *Bioinformatics*, 28(16):2184–2185, 2012. doi: 10.1093/bioinformatics/bts356.

M.C. Wilkes, C.E. Repellin, and K.M. Sakamoto. Beyond mRNA: The role of non-coding RNAs in normal and aberrant hematopoiesis. *Molecular Genetics and Metabolism*, 2017. doi: 10.1016/j.ymgme.2017.07.008.

Shanrong Zhao, Ying Zhang, William Gordon, Jie Quan, Hualin Xi, Sarah Du, David von Schack, and Baohong Zhang. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics*, 2015. doi: 10.1186/s12864-015-1876-7.