Analysis of bulk RNA-seq data

Analysis of Next-Generation Sequencing Data

Friederike Dündar

Applied Bioinformatics Core

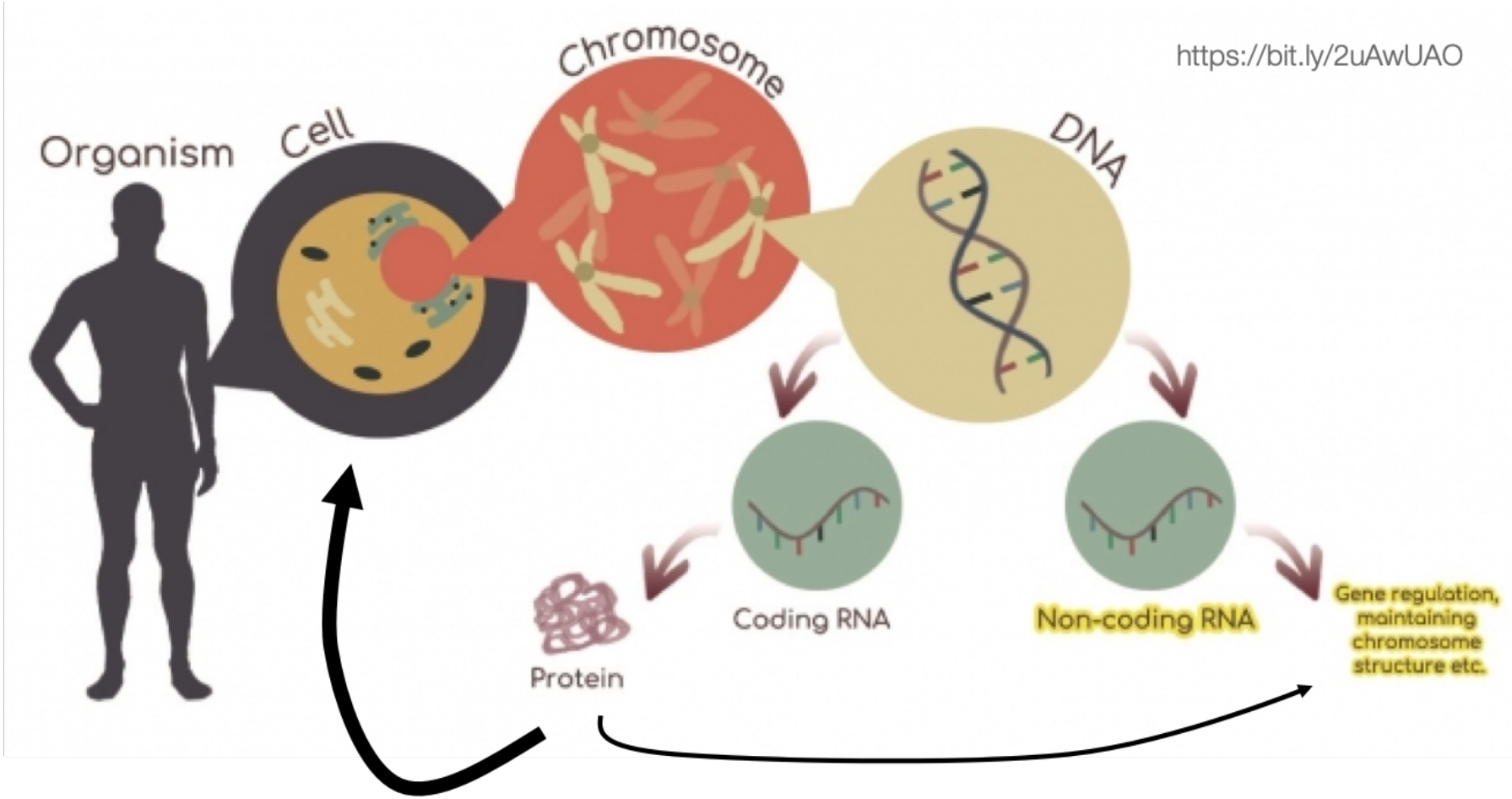
Slides at https://bit.ly/2T3sjRg[1]

February 18, 2020

[1] https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2020/

1. Why study RNA?
2. Different types of RNA – different library preps
3. Gene expression quantification
4. References

Why study RNA?



DNA is just the blueprint, it is not an effector molecule.

DNA is just the blueprint, it is not an effector molecule

GENOMICS

- DNA sequence of an organism
- genetic basis of phenotypic differences
- sites of DNA-protein or DNA-RNA interactions
- sites of open vs. closed chromatin

TRANSCRIPTOMICS

= characterization of **gene products**

- identification of specific RNAs
- quantification of RNAs
- RNA-protein interactions
- RNA structure

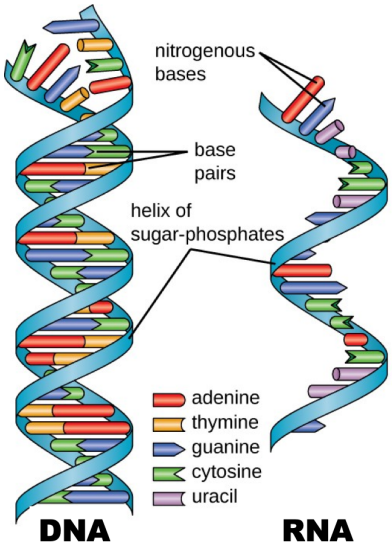
In order to understand the **functional consequences** (capacity) of a DNA sequence, we need to study **its products**, i.e. RNA and proteins.

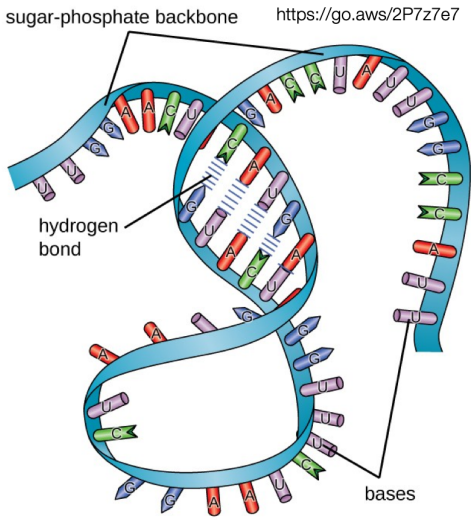
Different types of RNA – different library preps

DNA and RNA have different properties

DNA

- usually double-stranded
- very stable
- mutations are heritable
- same amount in (almost) all cells
- same sequence in every cell of an organism



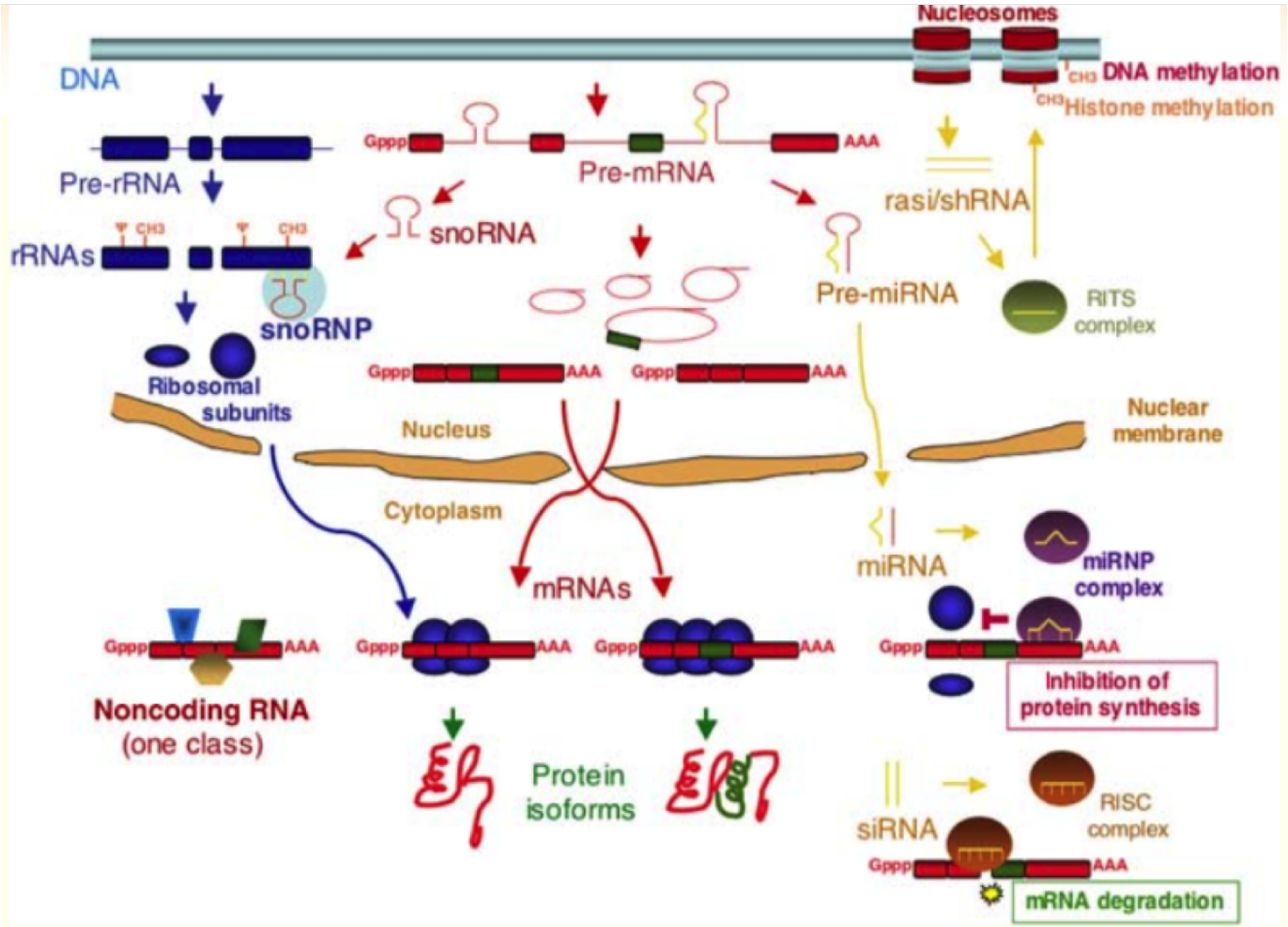


RNA

- generally single-stranded, but with the capacity for complementary base pairing ⇒ ability to form myriad different shapes
- usually fairly short-lived (minutes to hours)
- easily degraded/damaged without protection
- mutations are not passed on
- individual transcript *amounts* differ greatly depending on the gene, the cell type, the developmental status, the environment etc.
- transcript *sizes* range from 10-20bp to several kb

Different types of RNA

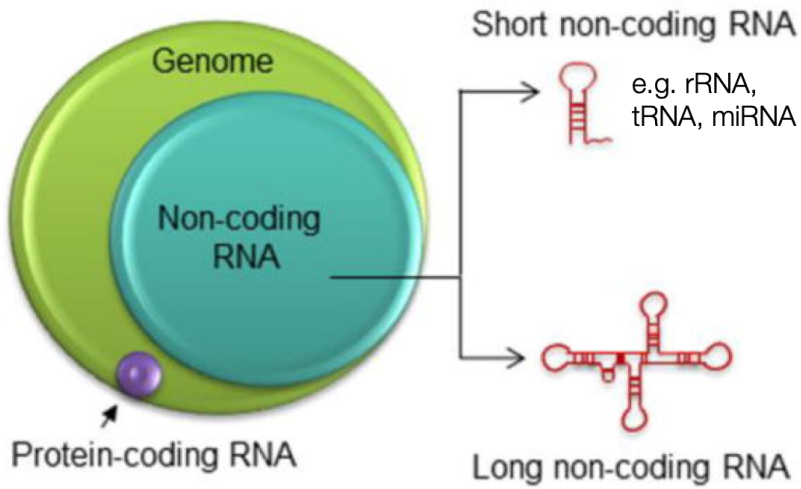
There are **numerous different types of *functional* RNA molecules** *in addition* to messenger RNA, which does *not* carry out a function of its own except transporting the DNA code (genetic information) into the cytoplasm where it can be translated into proteins.



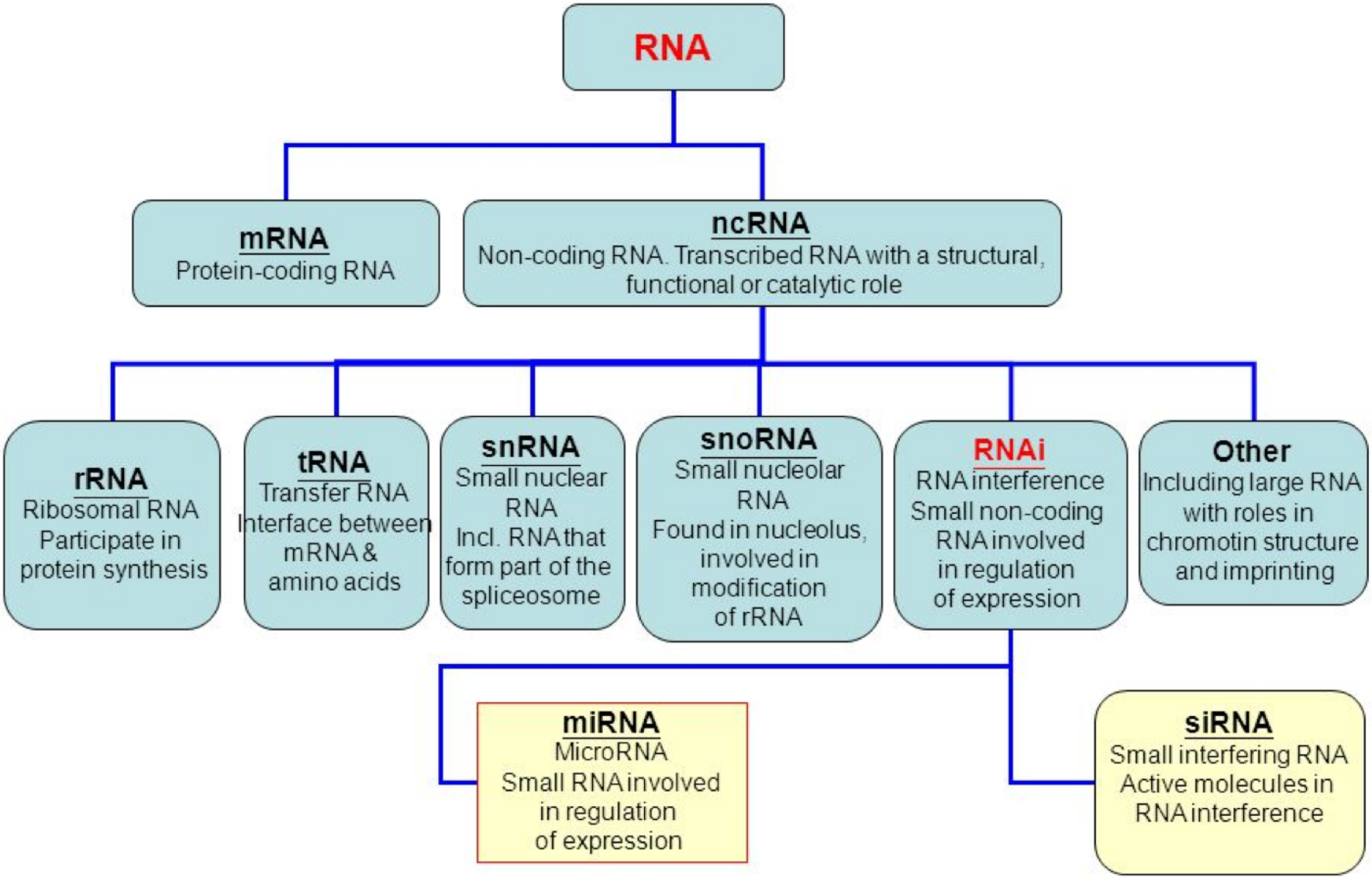
Different types of RNA

- ca. 75% of the human genome can be **transcribed** (= copied into RNA)
- but <3% of the genome is subsequently **translated** into proteins
- **genes** can therefore be **coding** (⇒ final product: protein) or **non-coding** (⇒ final product: RNA)
- non-coding RNAs cover a wide range of functions including **protein assembly** (⇒ ribosomal RNA, transfer RNA) and **gene expression regulation**

See Wilkes et al. [2017], Bartoszewski and Sikorski [2018] and Dai et al. [2020] for an introduction into the diverse RNA families and their functions.



Different types of RNA (there are more!)



Typical applications of RNA-seq

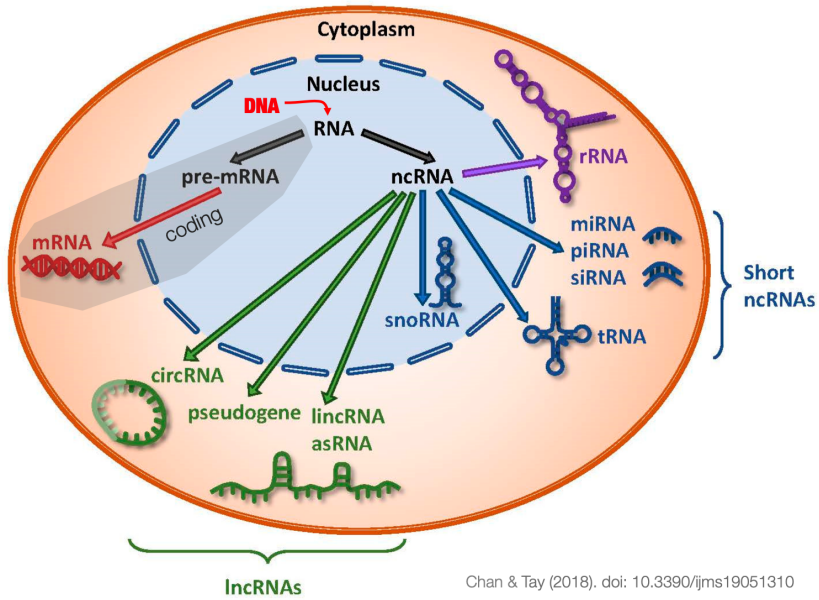
- **identification** of transcripts – *which portions of the genome are expressed?*
  - identification of splice variants
  - transcriptome assembly
  - detection of gene fusion events
- **quantification** of transcripts
  - comparison of different cell types/conditions/diseases and their effect on individual mRNA quantities
  - allele-specific expression

Illumina technology is best suited for the**quantification of known** transcripts; its short reads are not a good match for the identification of novel transcripts in very complex transcriptomes such as the ones found in mammals.

Sequencing prep protocol depends on the RNA properties

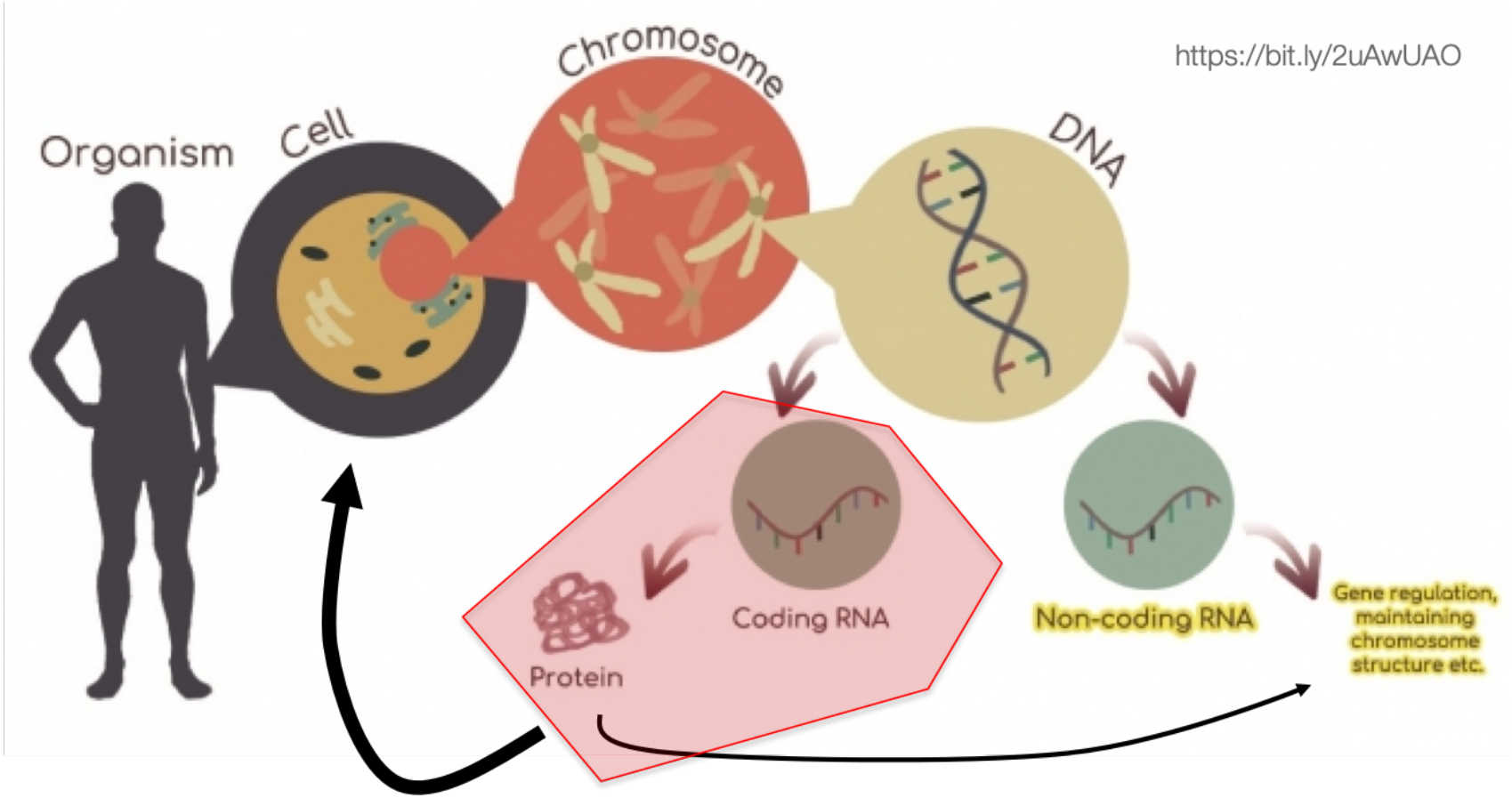
It is not a one-size-fits-all situation!

- **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

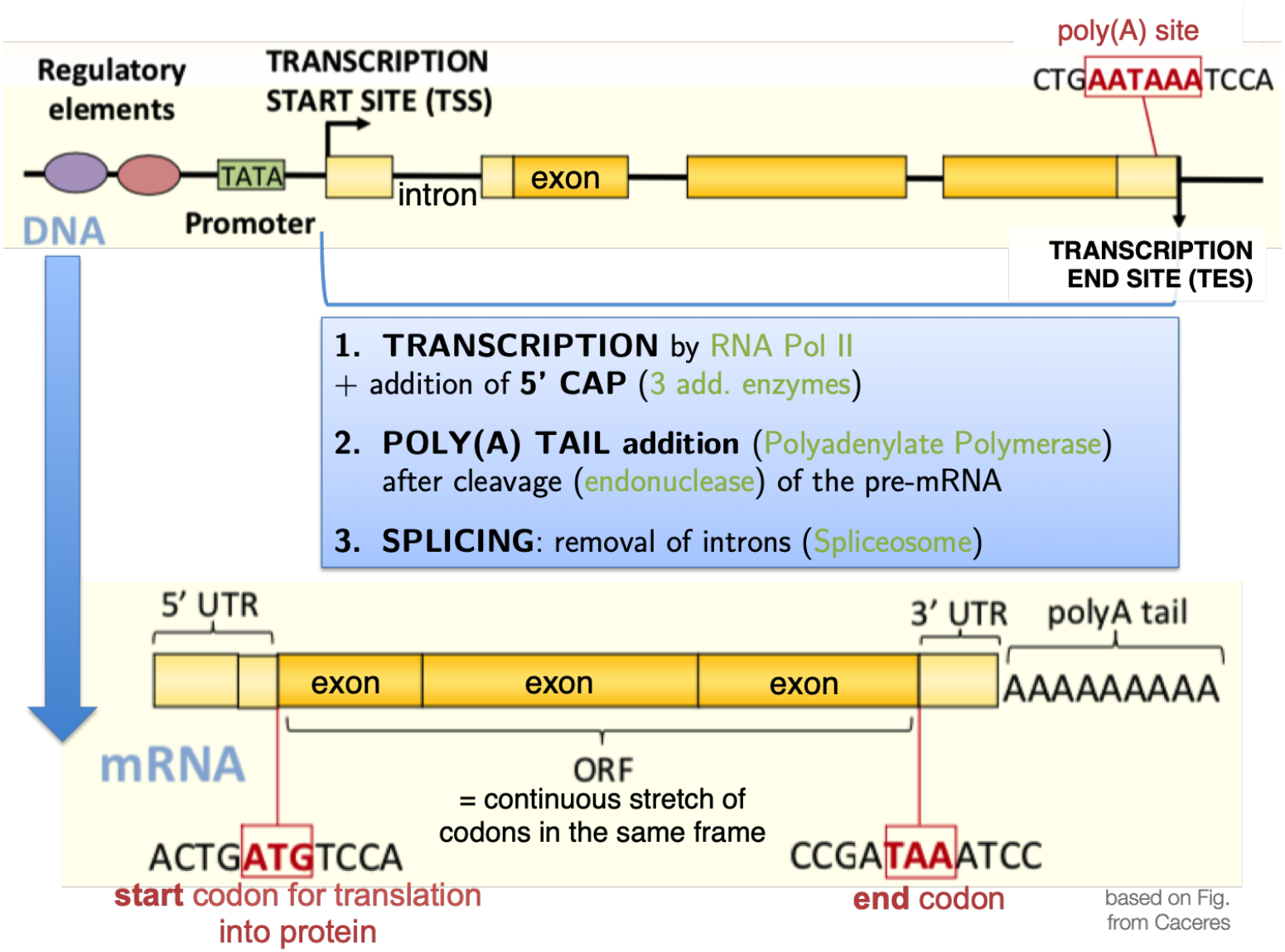


Focus today: messenger RNA

mRNA amounts are used as a proxy for the amounts of their corresponding proteins within a given tissue.



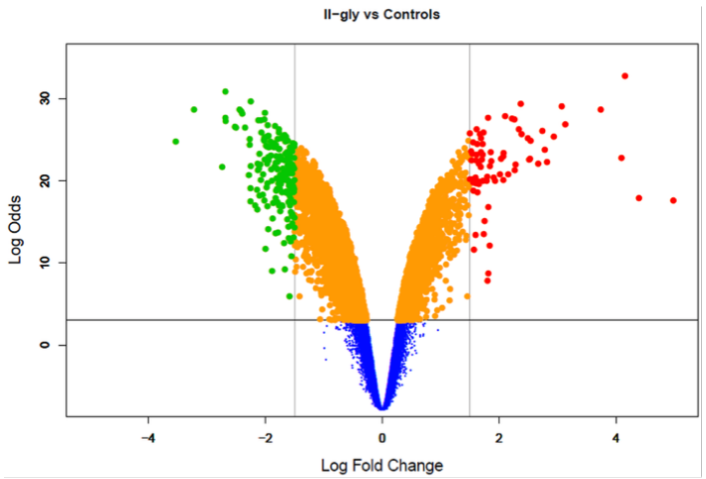
Focus today: messenger RNA

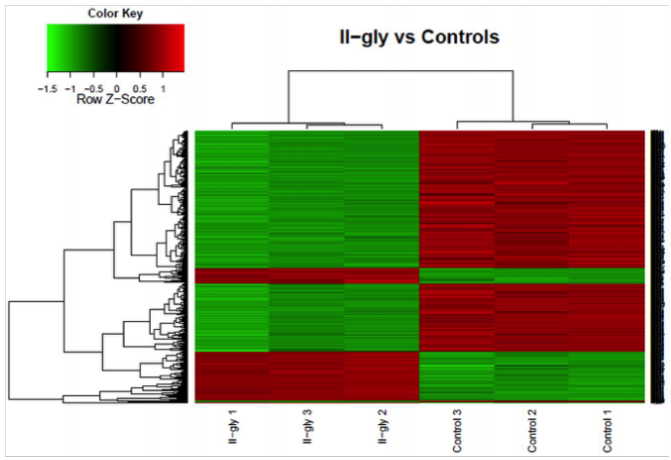


Focus today: messenger RNA

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**





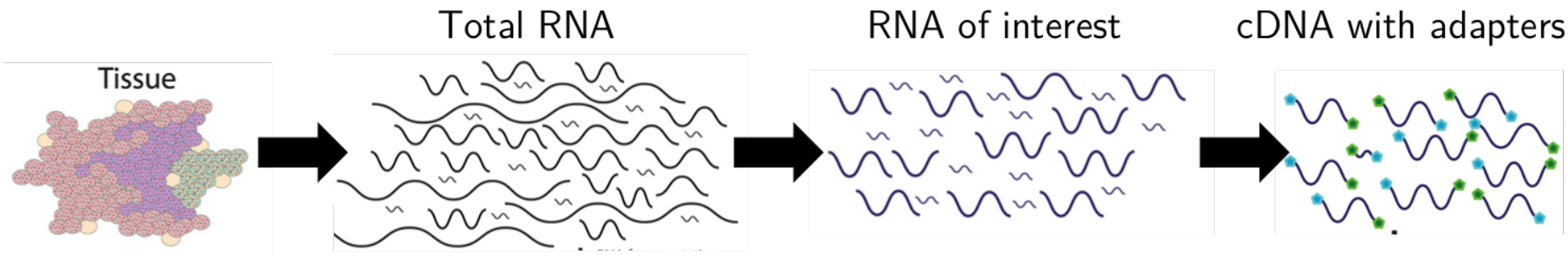
Valencia-Cruz et al. (2013). doi: 10.1371/journal.pone.0054664

Typical questions addressed with bulk RNA-seq

- Does a certain treatment induce gene expression changes? And if it does, which genes are most strongly affected?
- How does the gene expression profile of a cancer cell differ from a healthy cell?
- Which genes are turned on/off during the course of embryonic development?
- Which genes differ in mice that have been engineered to lack a certain gene? E.g., which genes – in addition to the one that's been "knocked-out" – may be depleted or overcompensating for the loss?
- Which genes are activated in response to an environmental stimulus, e.g. heat shock or alcohol poisoining?
- How does the gene expression profile change in the same tissue in an aging individual?
- ...

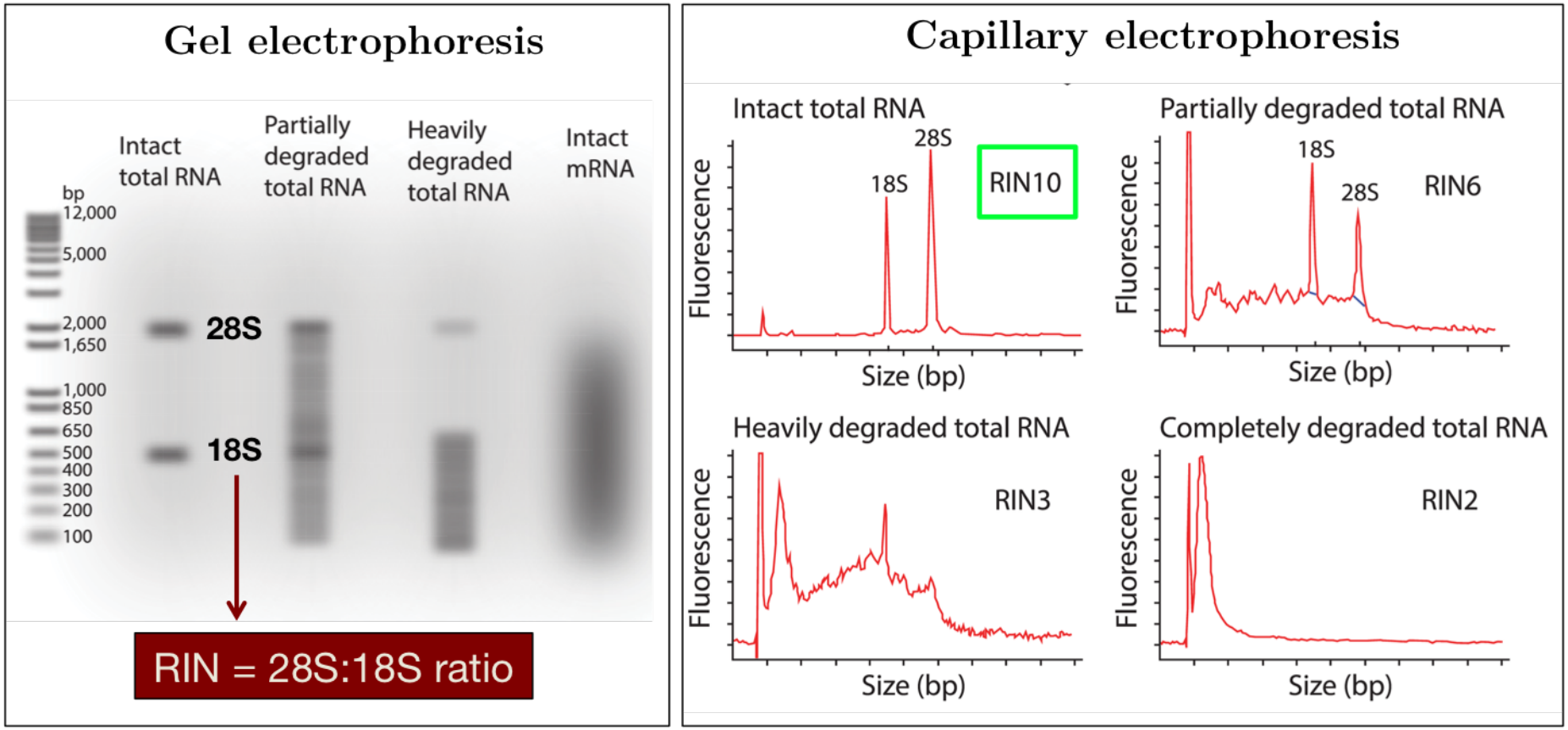
General steps of RNA-seq preparation

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!

QC of RNA extraction

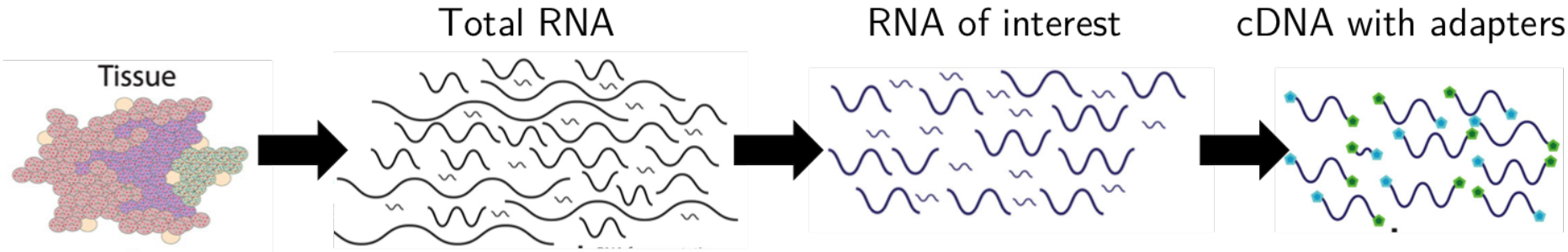


Griffith et al. (2015). doi: 10.1371/journal.pcbi.1004393

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

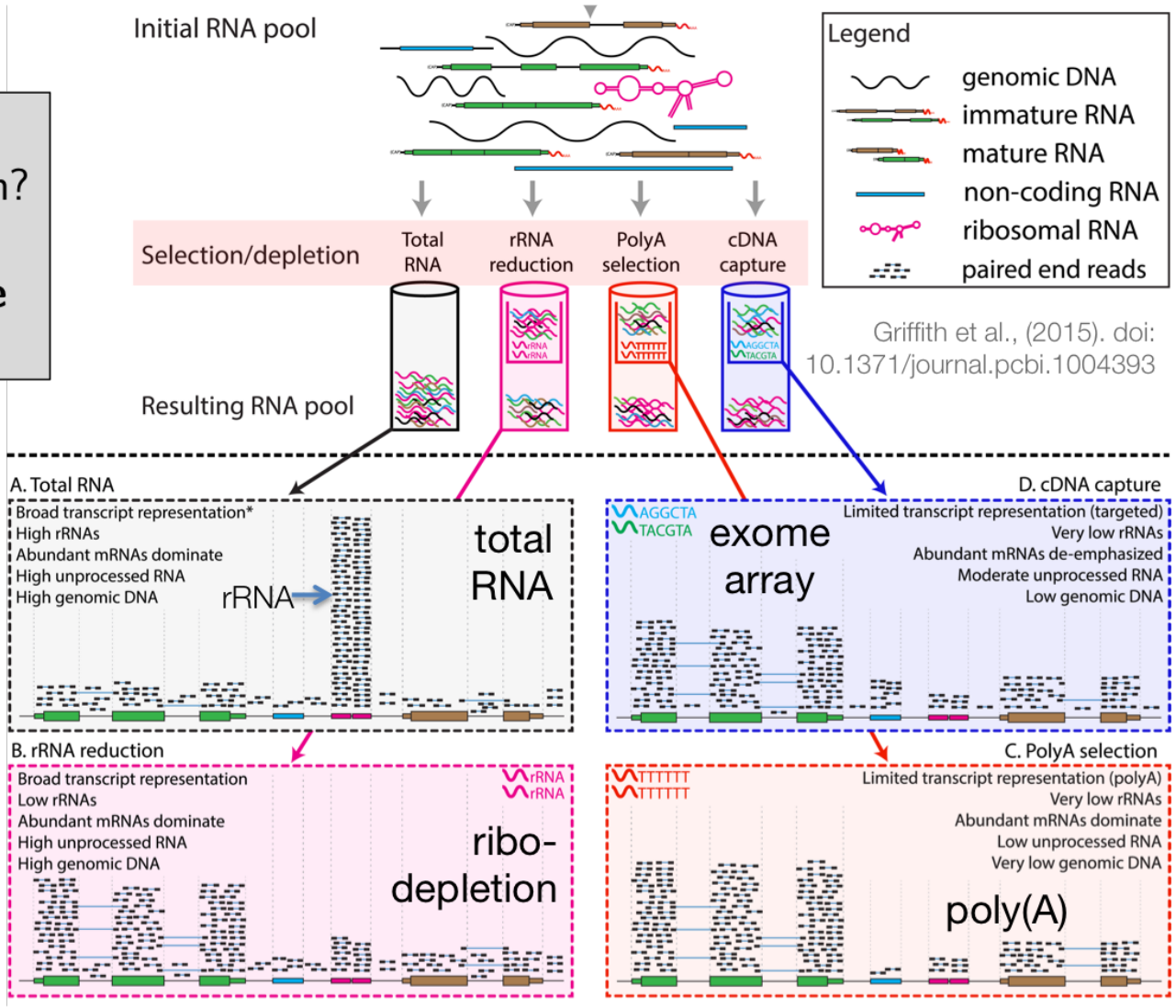


Every step has consequences – example: mRNA enrichment strategies

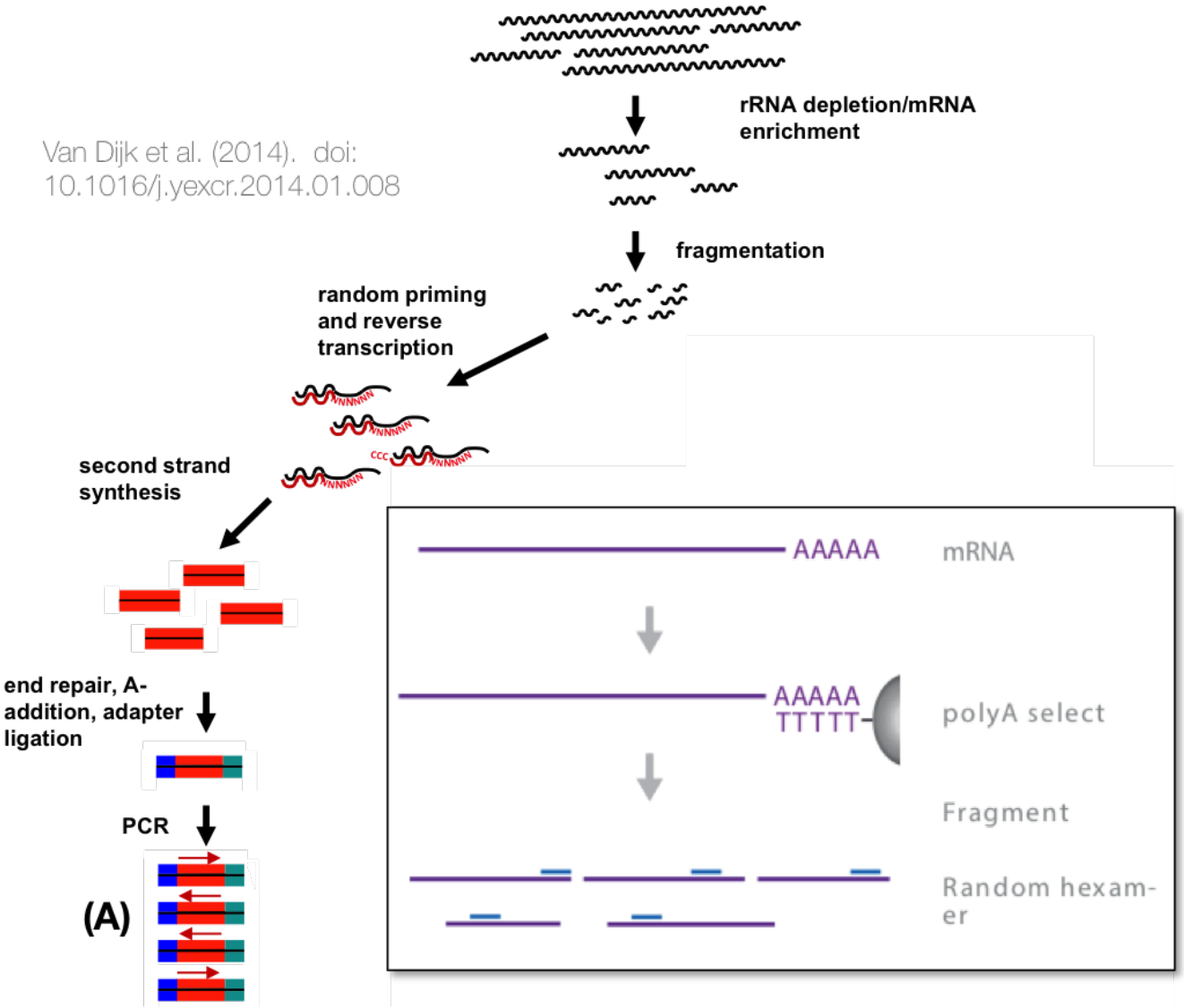
which **transcripts** are you interested in?

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

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

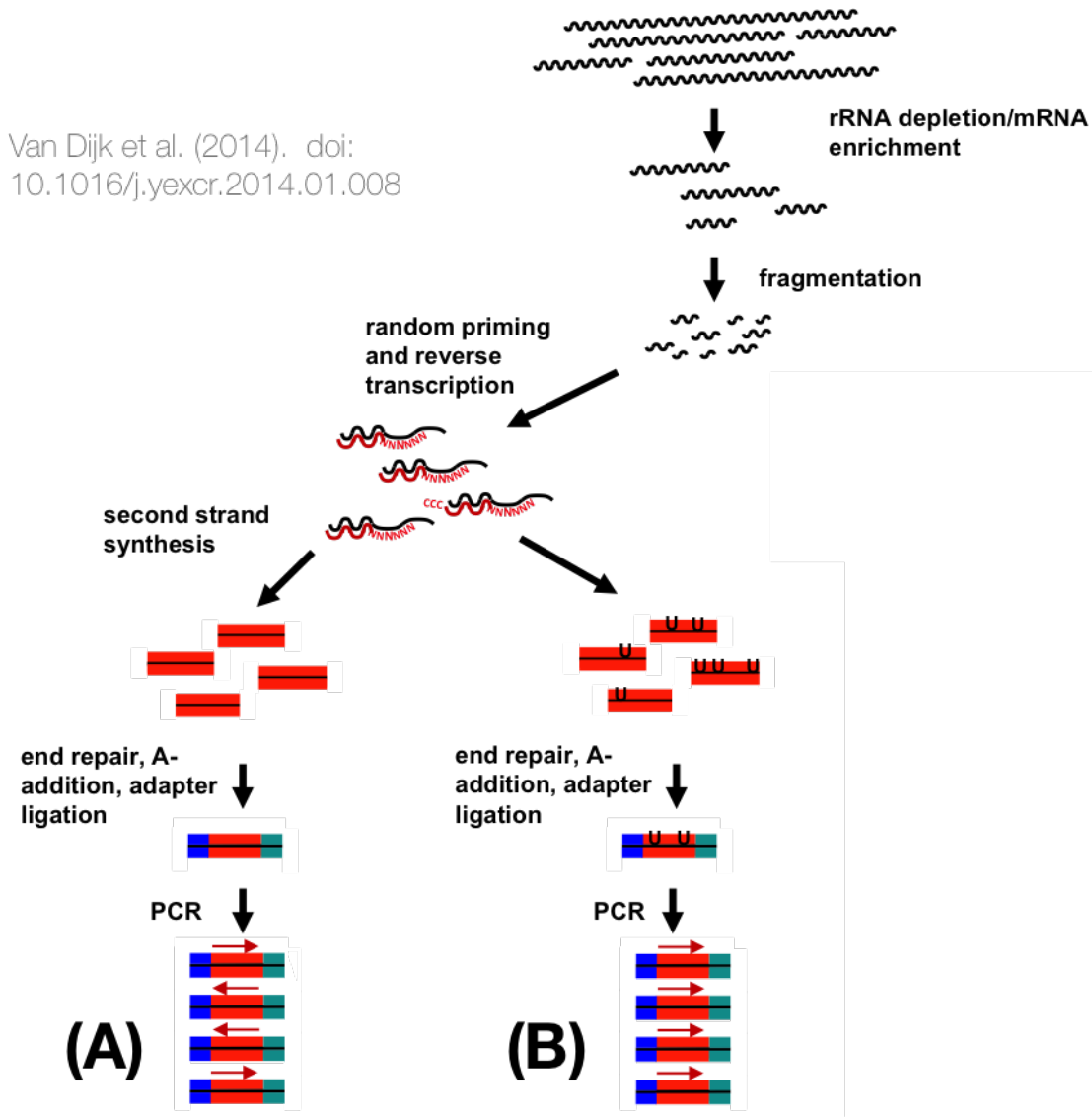


The most common library preparation methods



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

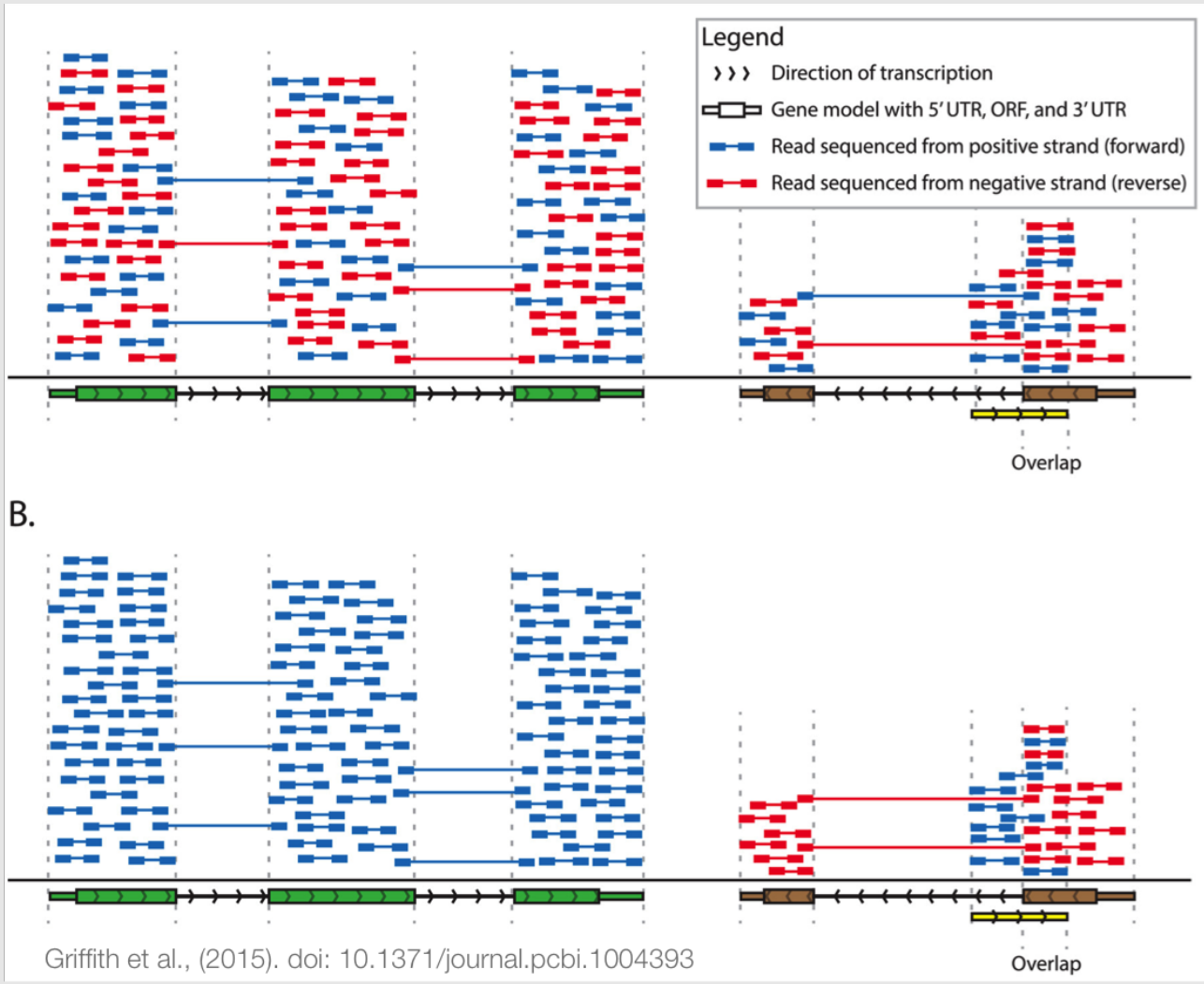
The most common library preparation methods



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

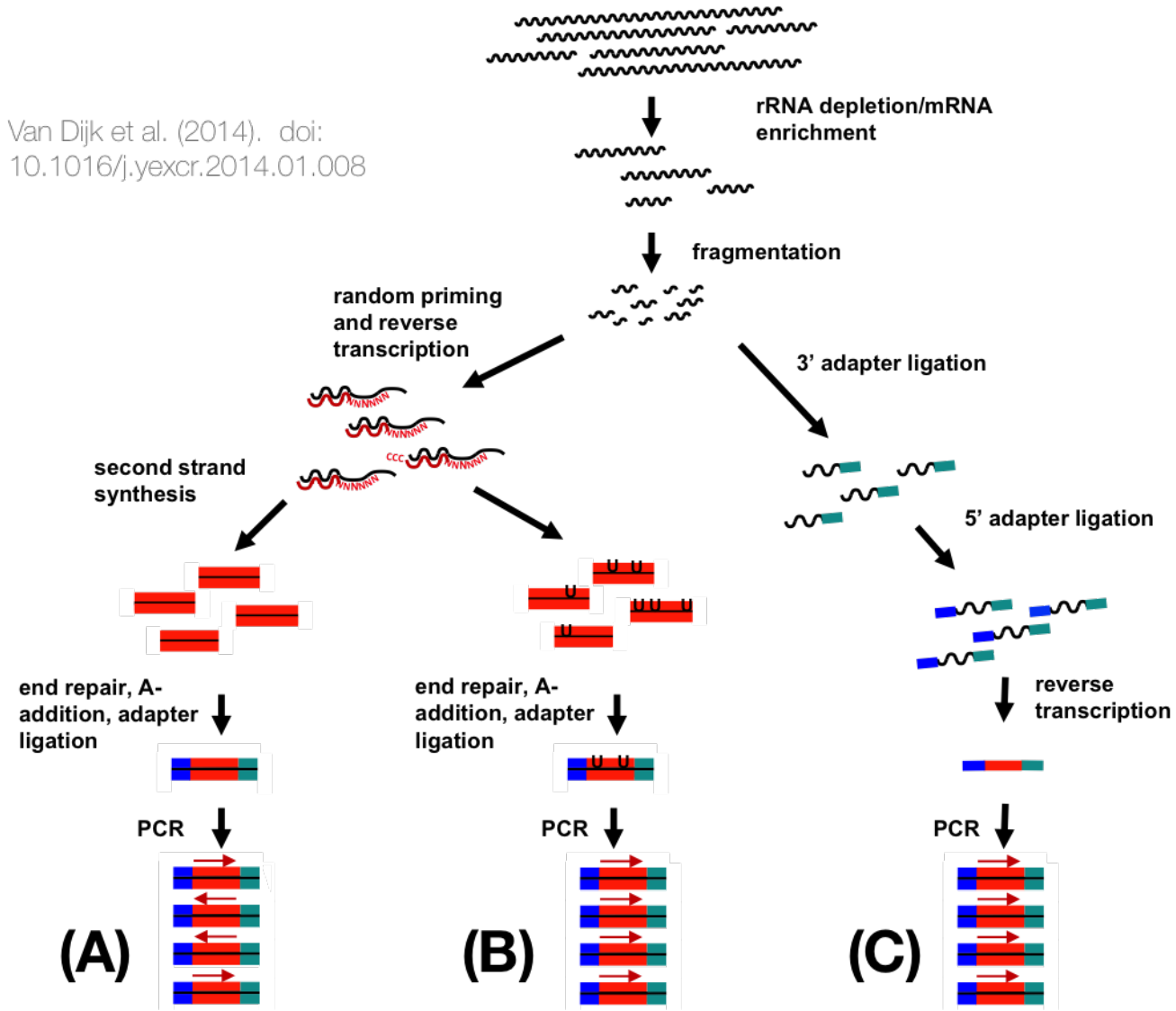
The most common library preparation methods

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!

Every step has consequences

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

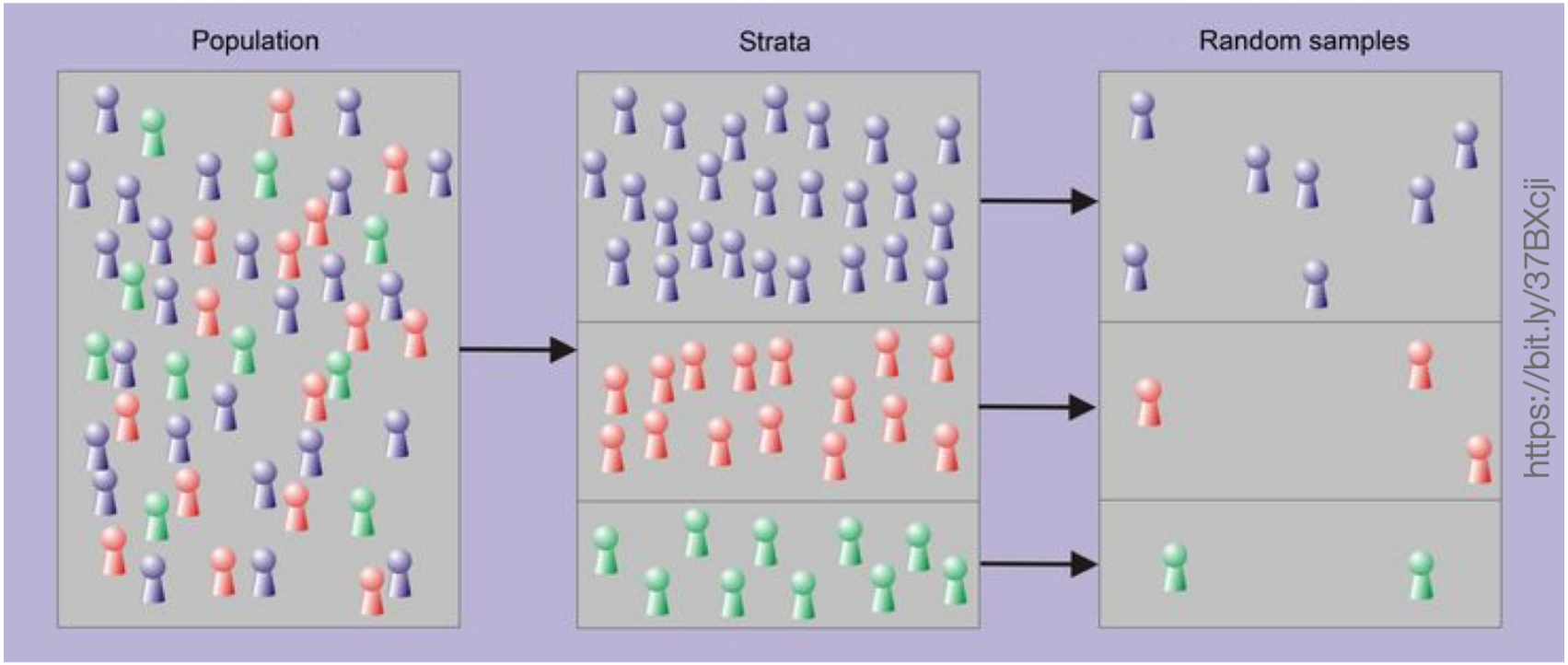
There are many papers comparing different aspects of different RNA-seq approaches, e.g.

- *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

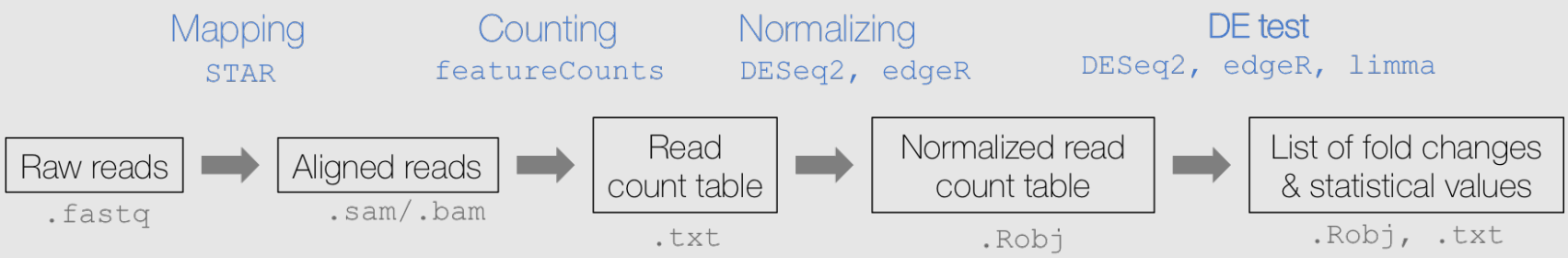
Sampling



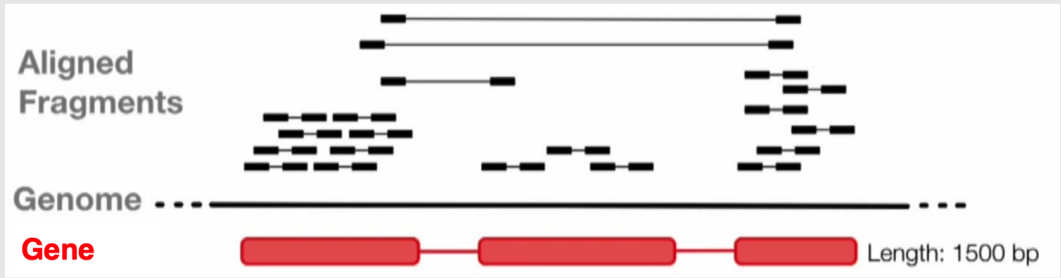
Basic assumption of RNA-seq-based transcript quantification:
The number of reads representing an individual transcript is reflecting its (relative!) abundance in the original transcript pool.
This means that every tx is assumed to have the same chance of being captured (without degradation), sequenced and aligned.

Quantification of gene expression following sequencing

Typical bioinformatics workflow for bulk RNA-seq data



1. **Read mapping**
   - with splice-aware alignment tools! e.g. `STAR` [Dobin et al., 2013, Dobin and Gingeras, 2016, Ballouz et al., 2018]
   - using reference genome and reference transcriptome information
2. **Count** reads that overlap with annotated genes



Quantification of gene expression

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`.

Quantification of gene expression

1. Aligning reads using STAR

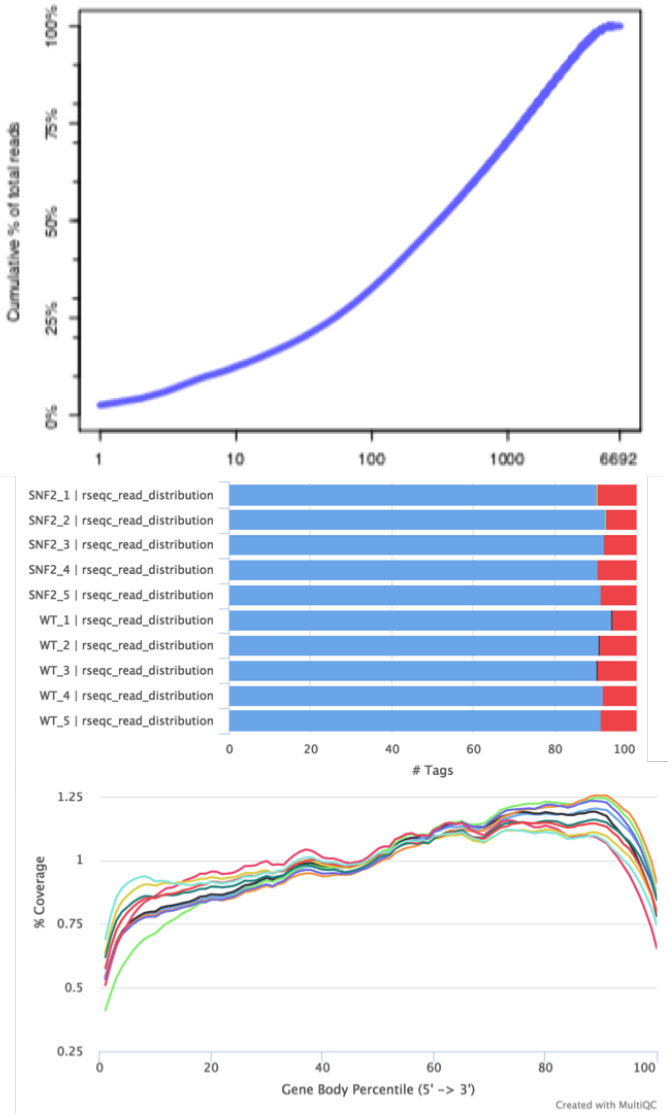
```
# Make the script executable:
$ chmod 755 align_Gierlinski.sh

# 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
```

Typical biases of aligned reads of RNA-seq experiments

- 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)



QC of aligned reads

- **How many reads aligned?**
  - ⇒ aligner output (e.g., `Log.final.out` = STAR's log file)
- **How well did the reads align?**
  - ⇒ `samtools flagstat`, RSeQC's `bam_stat`
  - these provide summaries of the `FLAG` field values
- **Did we capture mostly exonic RNA?**
  - ⇒ RSeQC's `read_distribution.py`, QoRTS
- **Do we see a pronounced 3'/5' bias?**
  - ⇒ RSeQC's `geneBody_coverage.py`, QoRTS

RSeQC package

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

- publication: Wang et al. [2012]
- documentation: http://rseqc.sourceforge.net
- 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

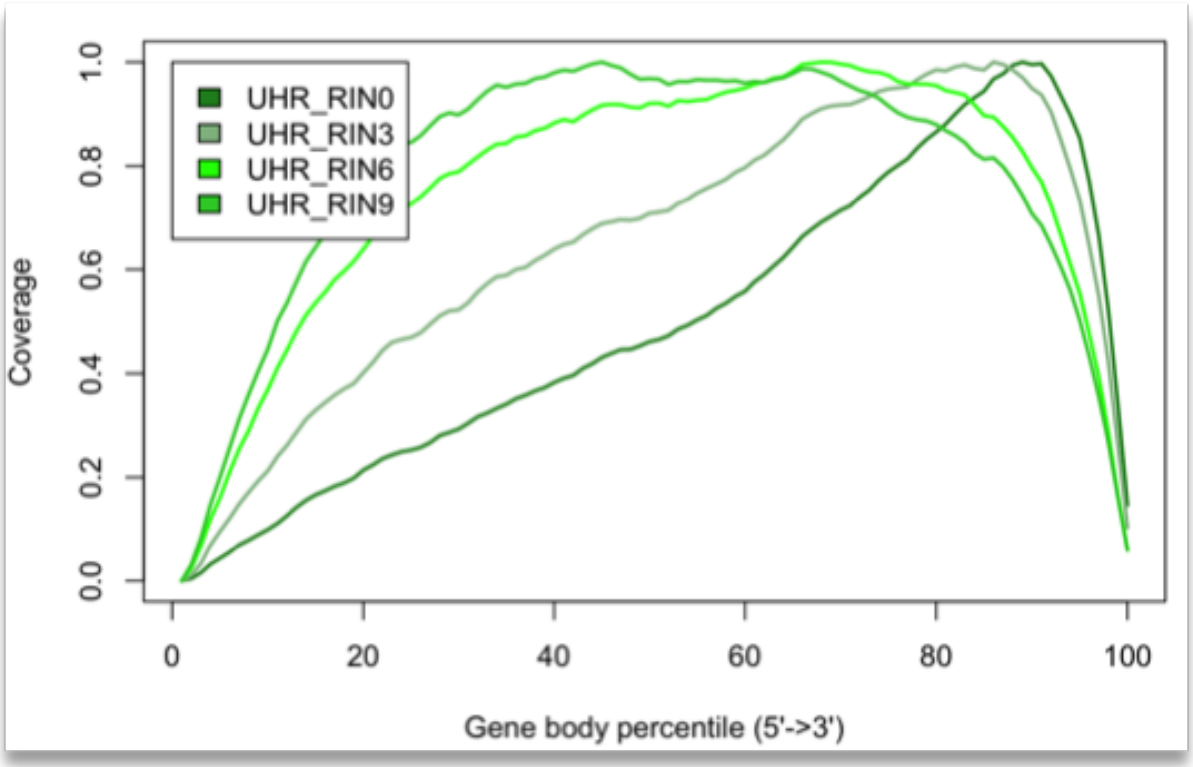
How many reads fall into exons? Based on annotation file (BED!)

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

# 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

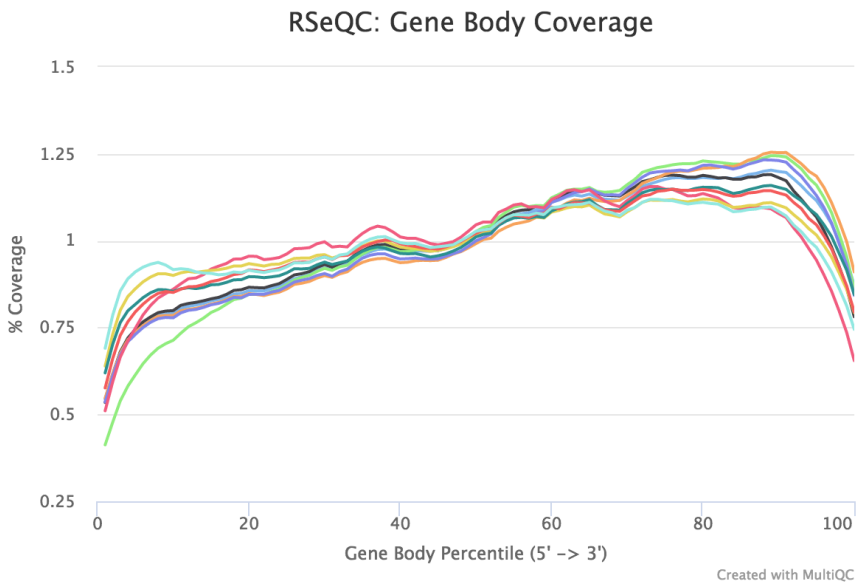
FastQC
(QoRTs)

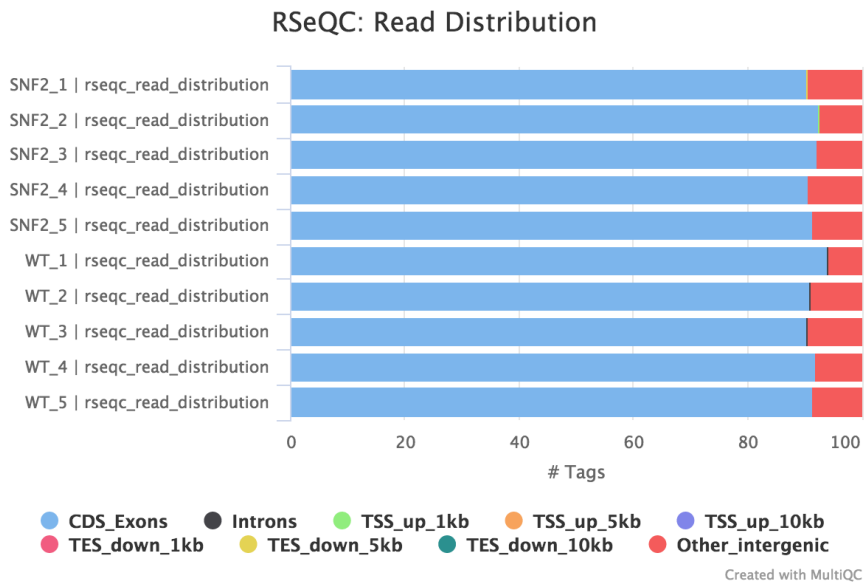
- **aligned reads QC (bam)**
  - % (uniquely) aligned reads
  - % exonic vs. intronic/intergenic
  - gene diversity
  - gene body coverage

aligner's log files
samtools flagstat
RSeQC
QoRTs
MultiQC
...

Summary of RNA-seq alignment QC

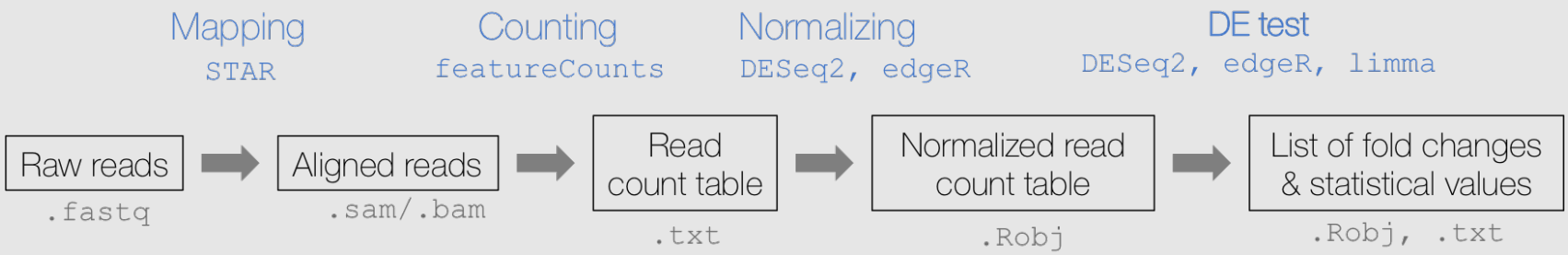
1. Did you capture a diverse set of mRNAs? (or RNAs of the type that you expect)?
2. Are the gene bodies covered similarly across different samples?
3. Is there evidence for contaminations, either from highly abundant, irrelevant transcripts or from genomic DNA?



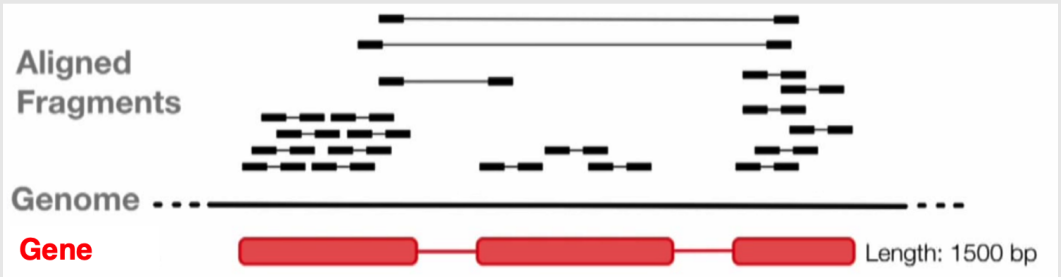


Quantification of gene expression following sequencing

Typical bioinformatics workflow for bulk RNA-seq data



1. **Read mapping**
   - with splice-aware alignment tools! e.g. `STAR`
   - using reference genome and reference transcriptome information
2. **Count** reads that overlap with annotated genes
   - the more often a transcript was present in a cell population, the more reads we should have mapped there



Quantification of gene expression following sequencing

2. Counting read-gene overlaps with `featureCounts` post-alignment

Several tools are available for this task; `featureCounts` is the fastest one [Liao et al., 2014].

Minimal input for counting read-gene overlaps:

- `GTF` file defining the loci from which RNA could have originated

```
(base) [frd2007@buddy ANGSD_2019]$ head RNA-seq/refGenome_S_cerevisiae/sacCer3.gtf
chrI    sacCer3_sgdGene start_codon     130799  130801  0.000000        +       .       gene_id "YAL012W"; transcript_id "YAL012W";
chrI    sacCer3_sgdGene CDS     130799  131980  0.000000        +       0       gene_id "YAL012W"; transcript_id "YAL012W";
chrI    sacCer3_sgdGene stop_codon      131981  131983  0.000000        +       .       gene_id "YAL012W"; transcript_id "YAL012W";
chrI    sacCer3_sgdGene exon    130799  131983  0.000000        +       .       gene_id "YAL012W"; transcript_id "YAL012W";
chrI    sacCer3_sgdGene start_codon     335     337     0.000000        +       .       gene_id "YAL069W"; transcript_id "YAL069W";
chrI    sacCer3_sgdGene CDS     335     646     0.000000        +       0       gene_id "YAL069W"; transcript_id "YAL069W";
chrI    sacCer3_sgdGene stop_codon      647     649     0.000000        +       .       gene_id "YAL069W"; transcript_id "YAL069W";
chrI    sacCer3_sgdGene exon    335     649     0.000000        +       .       gene_id "YAL069W"; transcript_id "YAL069W";
chrI    sacCer3_sgdGene start_codon     538     540     0.000000        +       .       gene_id "YAL068W-A"; transcript_id "YAL068W-A";
chrI    sacCer3_sgdGene CDS     538     789     0.000000        +       0       gene_id "YAL068W-A"; transcript_id "YAL068W-A";
```

- `BAM` file with the genome coordinates of each sequenced read

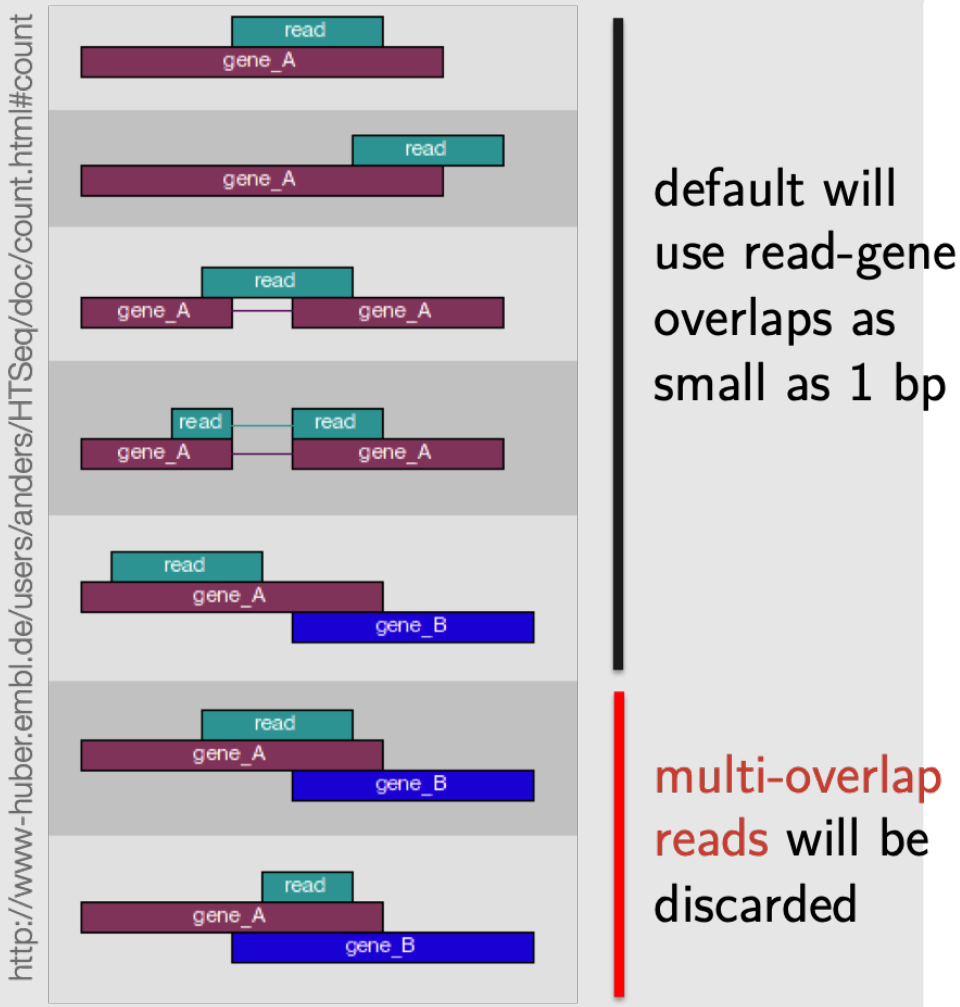
```
(base) [frd2007@buddy ANGSD_2019]$ samtools view alignment/WT_1_Aligned.sortedByCoord.out.bam | head
ERR458496.427513        16      chrI    3782    255     51M     *       0       0       CAGTAAAGGCTTGGTAGTAACCATAATATTACCCAGGTACGA
AACGCTAAG       JJJJIJJJJIJJJJJJJJIJJJJJJJJGHBJJJJJJJHHHHHFFFFFCCB     NH:i:1  HI:i:1  AS:i:50 nM:i:0
ERR458493.243111        0       chrI    3873    255     51M     *       0       0       TGAAAATATTCTGAGGTAAAAGCCATTAAGGTCCAGATAACC
AAGGGACAA       @@<DDDDEH<<<CEFHAEIHHGGCHIGEEGG?@FF@<?ECGGGIIHIIIII     NH:i:1  HI:i:1  AS:i:50 nM:i:0
ERR458494.816646        16      chrI    3972    255     51M     *       0       0       TAATGAGCTAGTGATCCGGAAAGCTACTTTATGATGTTTCAA
GGCCTGAAG       @BBEBG;IGFDD@C:BHC<4A?94C<CH<EEEFABBFE<CDCBBD;DD=?1     NH:i:1  HI:i:1  AS:i:50 nM:i:0
ERR458497.736781        16      chrI    3976    255     51M     *       0       0       GAGCTAGTGATCCGGAAAGCTACTTTATGATGTTTCAAGGCC
TGAAGTTTG       IJJJJJJJJJJJJJJJJJJIIJJJJJJJJJJJJJJJJHHHHHFFFFFCCC     NH:i:1  HI:i:1  AS:i:50 nM:i:0
```

Quantification of gene expression following sequencing

2. Counting read-gene overlaps with `featureCounts` post-alignment

- **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` (`-g` option)

See http://bioinf.wehi.edu.au/featureCounts/ and Chapter 7 of SubreadUsersGuide.pdf for details of `featureCounts` usage!



Quantification of gene expression following sequencing

2. Counting read-gene overlaps with `featureCounts` post-alignment

```
# featureCounts is part of the subread suite
$ spack load subread
# we can use all BAM files in the folder alignment via *.bam
$ featureCounts -a sacCer3.gtf \
                -o featCounts_Gierlinski_genes.txt \
               ../alignment/*.bam
```

The default `featureCounts` settings include **summarization at the meta-feature level** identified via "gene id" in the GTF file.

Quantification of gene expression following sequencing

2. Counting read-gene overlaps with `featureCounts` post-alignment

2 result files:

(1) `.txt.summary`: how many reads could be assigned to a feature

Status	SNF2_1	SNF2_2	SNF2_3	SNF2_4	SNF2_5
Assigned	9518261	8025575	8099295	9933479	6389328
Unassigned_Unmapped	0	0	0	0	0
Unassigned_MappingQuality	0	0	0	0	0
Unassigned_Chimera	0	0	0	0	0
Unassigned_FragmentLength	0	0	0	0	0
Unassigned_Duplicate	0	0	0	0	0
Unassigned_MultiMapping	0	0	0	0	0
Unassigned_Secondary	0	0	0	0	0
Unassigned_Nonjunction	0	0	0	0	0
Unassigned_NoFeatures	747388	458993	493370	750124	433271
Unassigned_Overlapping_Length	0	0	0	0	0
Unassigned_Ambiguity	586529	527766	544541	612430	427882

This file can easily be included in any `MultiQC` report.

Quantification of gene expression following sequencing

2. Counting read-gene overlaps with `featureCounts` post-alignment

(2) `.txt`: contains the actual read counts per feature

```
$ head -1 featCounts_Gierlinski_genes.txt.
# Program:featureCounts v1.6.2; Command:"featureCounts" "-a" "sacCer3.gtf" "-o"
# "featCounts_Gierlinski_genes.txt" "SNF2_1" "SNF2_2" "SNF2_3" "SNF2_4" "SNF2_5"
# "WT_1" "WT_2" "WT_3"
```

Geneid	Chr	Start	End	Strand	Length	SNF2_1	SNF2_2	SNF2_3
YAL012W	chrI	130799	131983	+	1185	7351	7180	7648
YAL069W	chrI	335	649	+	315	0	0	0
YAL068W-A	chrI	538	792	+	255	0	0	0
YAL068C	chrI	1807	2169	-	363	2	2	2
YAL067W-A	chrI	2480	2707	+	228	0	0	0
YAL067C	chrI	7235	9016	-	1782	103	51	44

This is the file you're going to use for downstream processing and analyses.

References

Figures taken from the following publications:

[Cáceres, 2012, Chan and Tay, 2018, Dündar et al., 2019, Griffith et al., 2015, Parasramka et al., 2016, Valencia-Cruz et al., 2013, van Dijk et al., 2014]

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.

Mario Cáceres. Functional genomics and transcriptomics, 2012. URL http://bioinformatica.uab.cat/base/documents/mastergp/Transcriptomics_2012.pdf.

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

Xiaofeng Dai, Shuo Zhang, and Kathia Zaleta-Rivera. RNA: interactions drive functionalities. *Molecular Biology Reports*, 47(2):1413–1434, 2020. doi: 10.1007/s11033-019-05230-7.

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.

Alexander Dobin and Thomas R. Gingeras. Optimizing RNA-seq mapping with STAR. In *Methods in Molecular Biology*, volume 1415, pages 245–262. Humana Press, New York, NY, 2016. doi: 10.1007/978-1-4939-3572-7_13.

Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29 (1):15–21, 2013. doi: 10.1093/bioinformatics/bts635.

Friederike Dündar, Luce Skrabanek, and Paul Zumbo. Introduction to differential gene expression analysis using rna-seq, 2019. 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.

Yang Liao, Gordon K Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–30, April 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt656.

Mansi A. Parasramka, Sayantan Maji, Akiko Matsuda, Irene K. Yan, and Tushar Patel. Long non-coding RNAs as novel targets for therapy in hepatocellular carcinoma. *Pharmacology and Therapeutics*, pages 67–78, 2016. doi: 10.1016/j.pharmthera.2016.03.004.

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.

Alejandra Idan Valencia-Cruz, Laura I. Uribe-Figueroa, Rodrigo Galindo-Murillo, Karol Baca-López, Anllely G. Gutiérrez, Adriana Vázquez-Aguirre, Lena Ruiz-Azuara, Enrique Hernández-Lemus, and Carmen Mejía. Whole Genome Gene Expression Analysis Reveals Casiopeína-Induced Apoptosis Pathways. *PLoS ONE*, 2013. ISSN 19326203. doi: 10.1371/journal.pone.0054664.

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.