Day 2: Identifying the transcripts that were sequenced

1. Experimental Design
2. FastQC results
3. Reference genome & transcript annotation
4. Alignment
   - STAR
   - BAM/SAM files
5. QC of alignment step

EXPERIMENTAL DESIGN

How to avoid spurious signals and drowning in noise

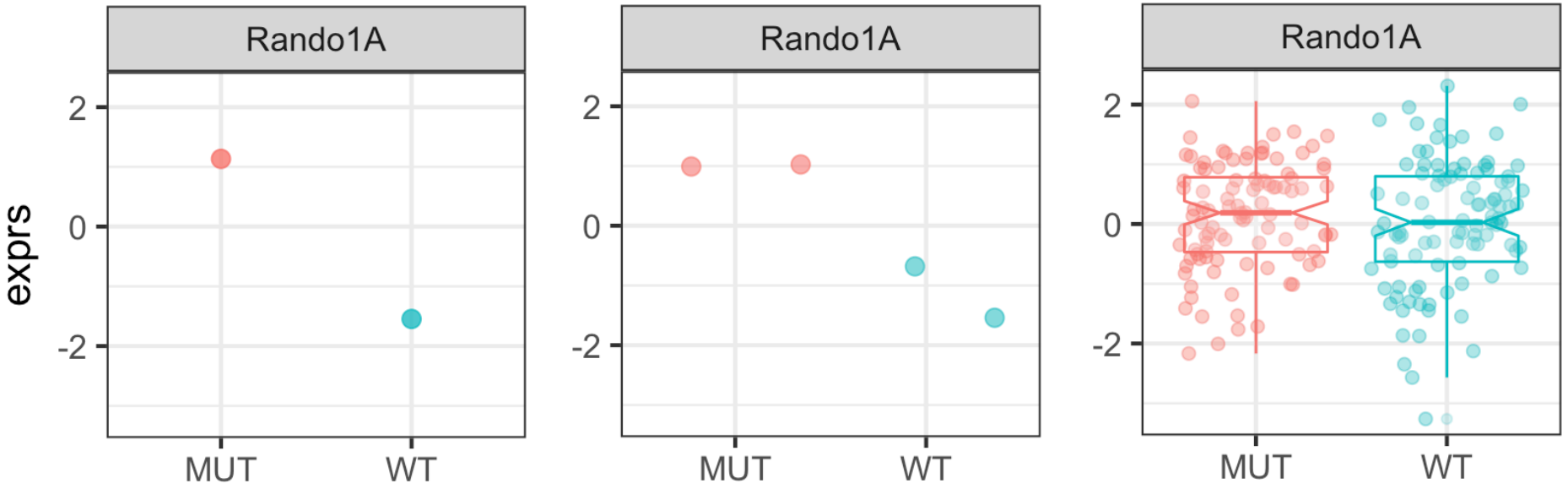
Why do we need replicates?

Goal: Identify differences in expression for every gene.

…and "differences" should preferably be due to our experiment, not noise!

"Samples are our windows to the population, and their statistics are used to estimate those of the population."
Martin Krzywinski & Naomi Altman

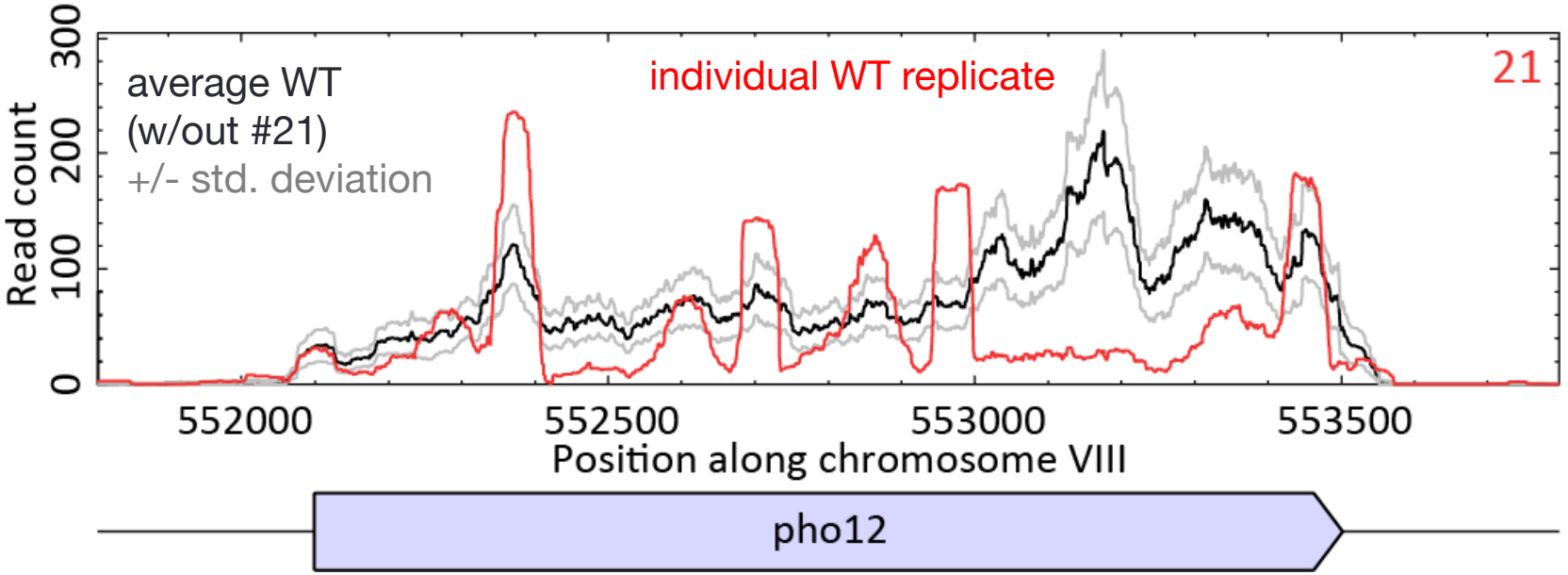
```
testdat <- data.frame(exprs = rnorm(200),
                      condition = c("WT","MUT"),
                      gene_name = "Rando1A")
```



doi:10.1038/nmeth.2613

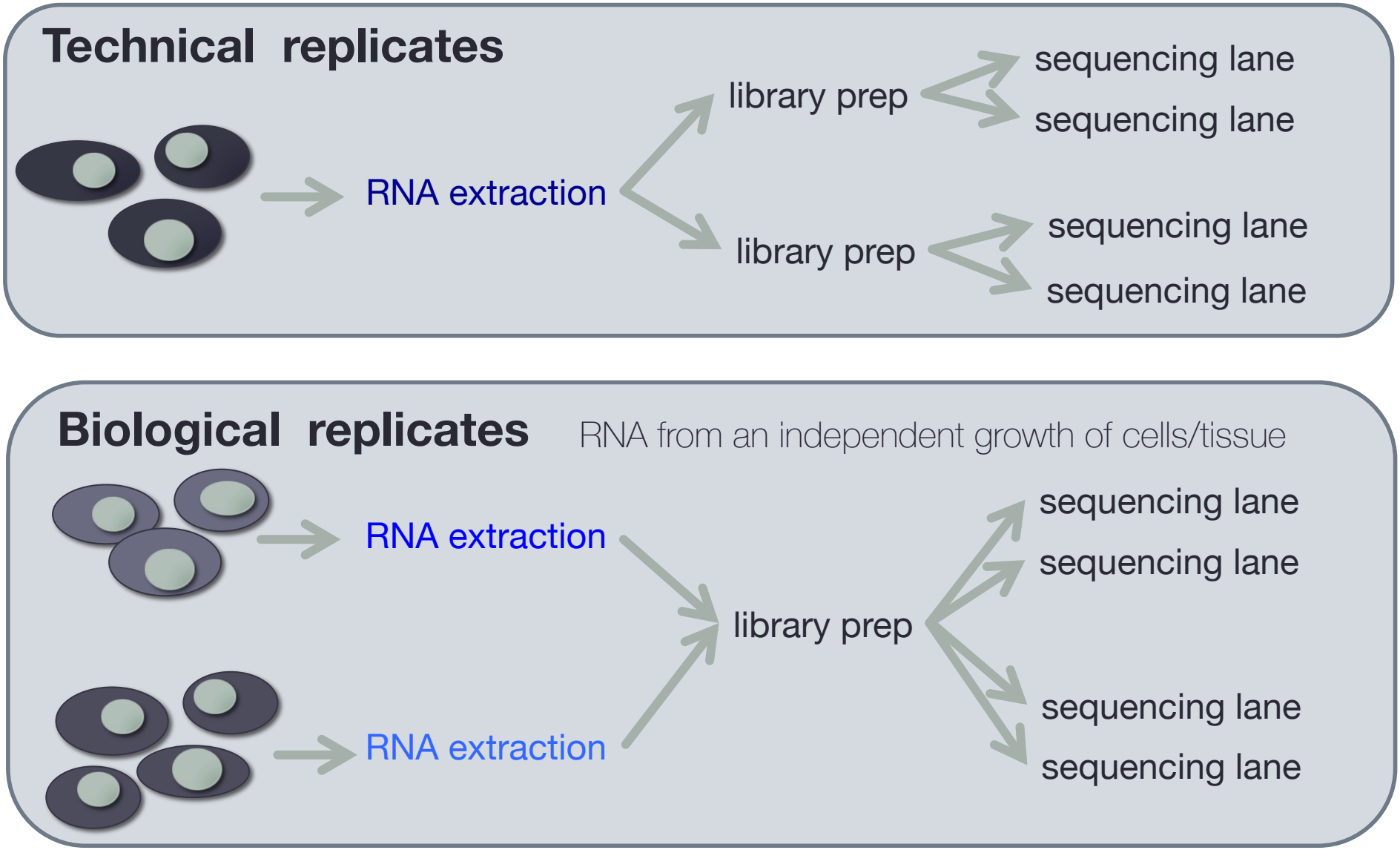
Invest in replicates!

- recommended: **6 biological replicates per condition** for DGE of strongly changing genes (logFC >= 2) [based on insights from the fairly simple yeast transcriptome]



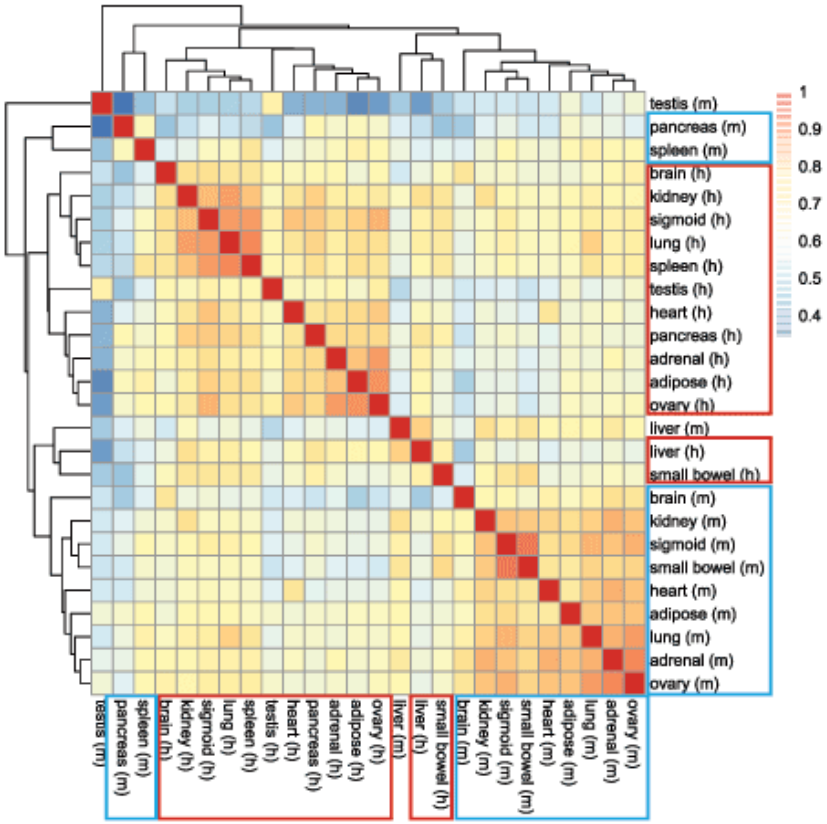
The most effective way to improve detection of differential expression in low expression genes is to add more biological replicates, rather than adding more reads (see Rapaport et al., 2013).

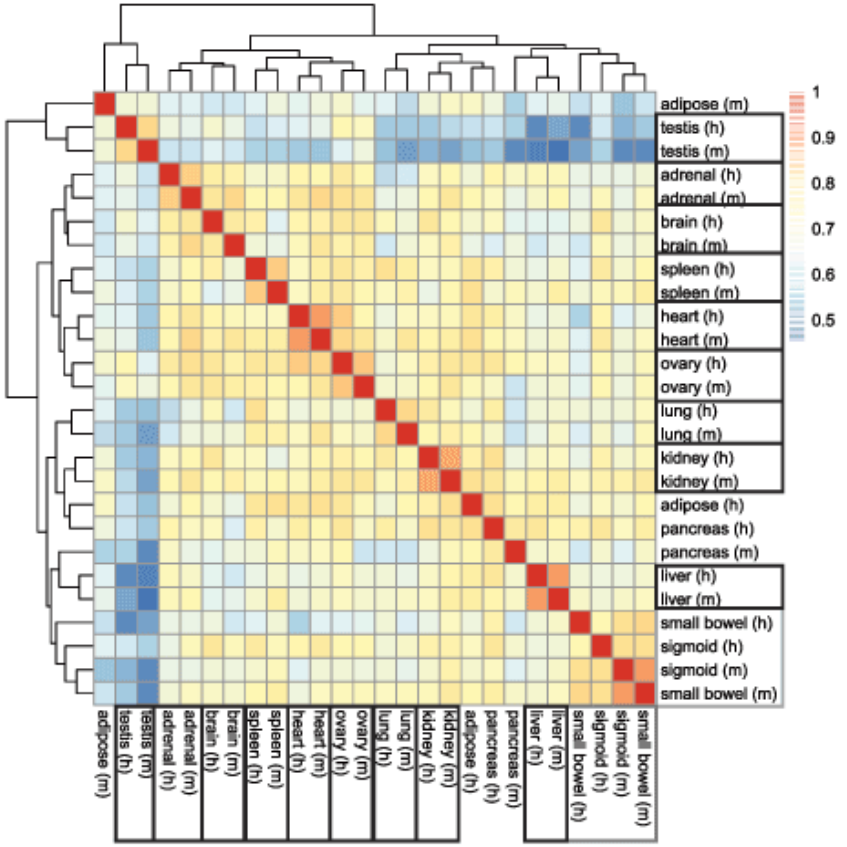
Replicate types



also see course notes and Blainey et al. (2014) Nature Methods, 1(9) 879–880.

Batch effects can happen everywhere





"Overall, our results indicate that there is ***considerable RNA expression diversity between humans and mice****, well beyond what was described previously, likely reflecting the fundamental physiological differences between these two organisms. "*

*"****Once we accounted for the batch effect*** *(…), the comparative gene expression data no longer clustered by species, and instead, we observed* ***a clear tendency for clustering by tissue****."*

Lin, Lin, and Snyder (2014). PNAS 111:48

Gilad & Mizrahi-Man (2015). F1000Research 4:121

ENCODE's* study design was not optimal

Most human samples were sequenced separately from the mouse samples:

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

Many tissues were not sex-matched

Tissue	Human	Mouse
adipose	FEMALE	MALE
adrenal	MALE	FEMALE
brain	FEMALE	MALE
heart	FEMALE	FEMALE
kidney	MALE	FEMALE
liver	MALE	FEMALE
lung	FEMALE	FEMALE
ovary	FEMALE	FEMALE
pancreas	FEMALE	FEMALE
sigmoid colo	MALE	FEMALE
small bowel	FEMALE	FEMALE
spleen	FEMALE	MALE
testis	MALE	MALE

not all variables can be controlled for
human data: deceased organ donors
mouse data: 10-week-old littermates

and that's ok, but you've got to be mindful of these limitations when making bold claims

A very good read (including the reviews and comments) that discusses many scientific as well as ethical issues: https://f1000research.com/articles/4-121/v1

* not just ENCODE: see e.g. Leek et al. (2010) Nat Rev Gen 11(10) 733-739 or Jaffe & Irizarry (2014) Genome Biol 15(R31) 1–9

Avoiding bias

Completely randomized design

STRESS	A	B	A	A	B	A	B	A	A	B	B	B
DIET	1	2	1	2	2	1	1	2	2	1	2	1

Restricted randomized design

GENOTYPE	A	A	A	A	A	A	B	B	B	B	B	B
DIET	1	2	1	2	2	1	1	2	1	1	2	2

Blocked & randomized design

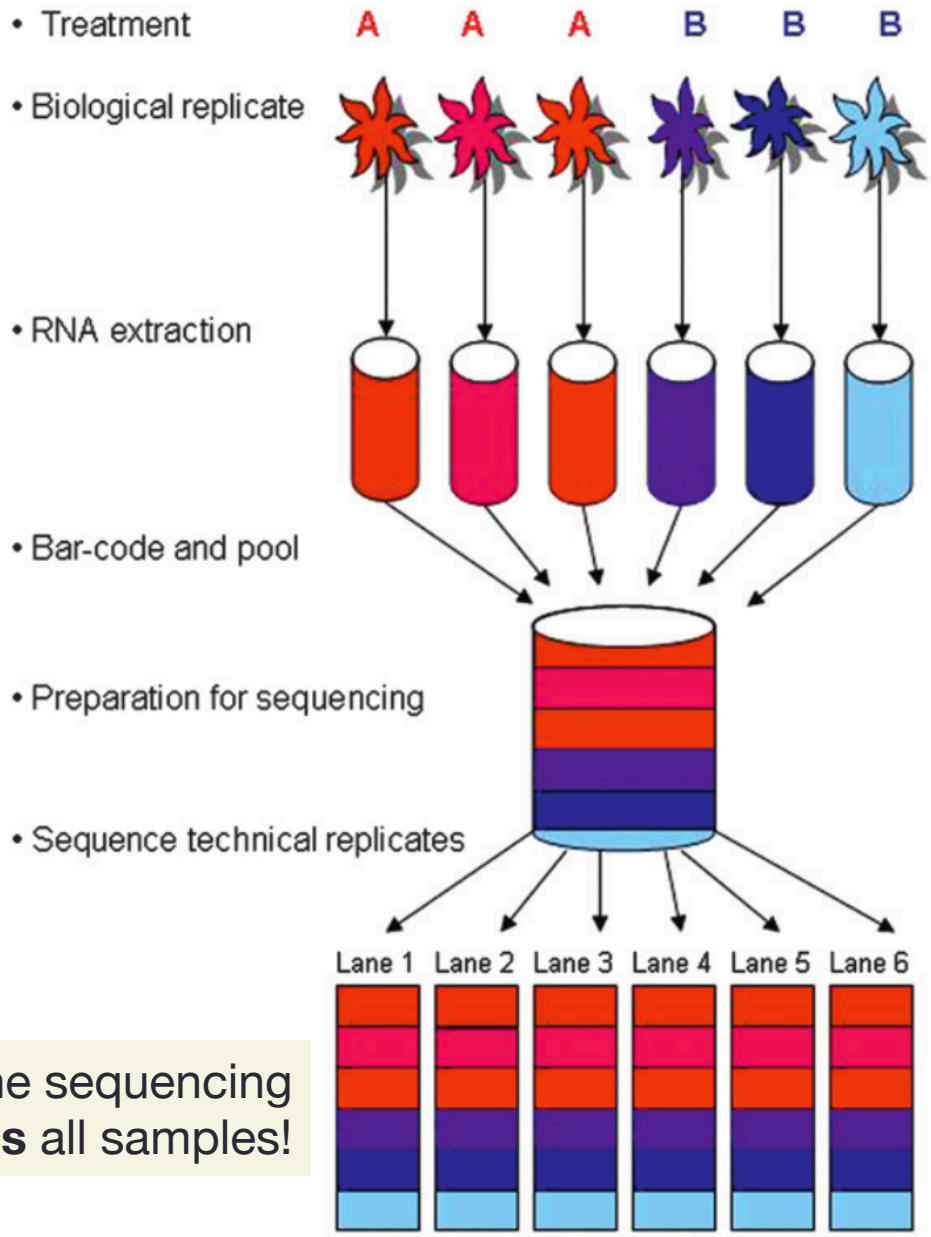
GENOTYPE	A	A	B	B	A	A	B	B	A	A	B	B
DIET	1	2	1	2	1	2	1	2	1	2	1	2
WEIGHT	•	•	•	•	●	●	●	●	⬤	⬤	⬤	⬤

Block what you can,
randomize what you cannot.

*What factors are of **interest**? Which ones might introduce noise?*
Which nuisance factors do you absolutely need to account for?

Typical RNA-seq set-up

- keep the **technical nuisance** factors (harvest date, RNA extraction kit, sequencing date…) to a **minimum**
- cover only as much of the **biological variation as needed** (but keep possible limitations for the final conclusions in mind)



Make sure the sequencing core **multiplexes** all samples!

How deep is deep enough?

for DGE (logFC~ 2) in mammals:
20 – 50 mio SR, 75 bp

Goals that require **more**, **longer,** and possibly **paired-end** reads:

- quantification of **lowly expressed** genes
- identification of genes with **small changes** between conditions
- investigation of **alternative splicing**/isoform quantification
- identification of **novel transcripts**, chimeric transcripts
- *de novo* **transcriptome assembly**

Remember: The addition of replicate samples provides substantially greater detection power of DE than increased sequence depth. (Rapaport et al., 2013)

Summary

- RNA-seq analysis is not a completely solved issue – but **DE analysis on a gene level** is decently mature and the field seems to gravitate towards some sort of standard
- no analysis tool can enforce (or replace!) common sense and knowledge about the biology behind the experiment
- crap in, crap out
- more replicates are often better investments than more reads

QUALITY CONTROL OF RAW READS

FastQC results

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/
```

FastQC & MultiQC

randomly selected 8 biological replicates for each condition (WT, SNF2)

```
mkdir raw_reads_QC/fastqc_results

for GENOTYPE in WT SNF2
do
    for i in 1 2 5 6 13 21 25 28 # random selection
    do
        echo Running FastQC for: ${GENOTYPE} Sample No ${i}
        # make a folder for every sample's FastQC results
        mkdir raw_reads_QC/fastqc_results/${GENOTYPE}_${i}
        # run FastQC
        ~/mat/software/FastQC/fastqc ~/precomputed/rawReads_yeast_Gierlinski/${GENOTYPE}_${i}/*stq.gz \
            -o raw_reads_QC/fastqc_results/${GENOTYPE}_${i} -q
    done
done

cd raw_reads_QC/fastqc_results/

# run MultiQC to summarize all the FastQC results into one document
~/mat/software/anaconda2/bin/multiqc . --dirs --interactive # --dirs will use the folder names as
sample names in the output
```

http://chagall.med.cornell.edu/RNASEQcourse/multiqc_report.html

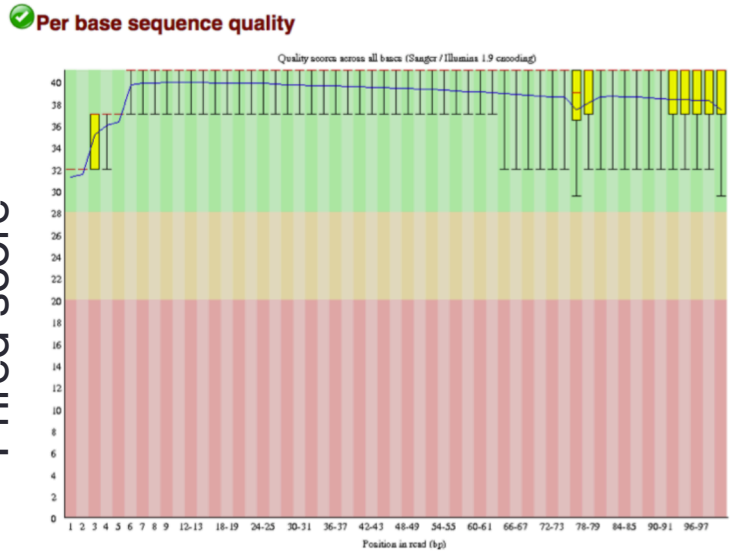
Two basic questions of QC

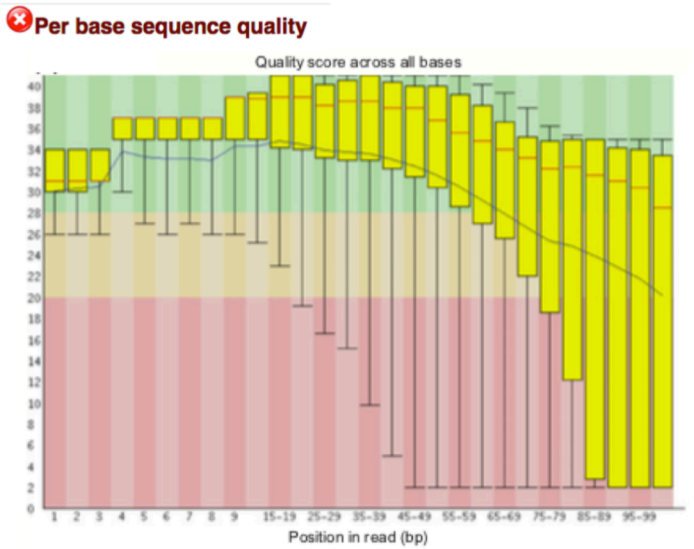
- How successful was the actual **sequencing**?
  - consistently high base call confidence

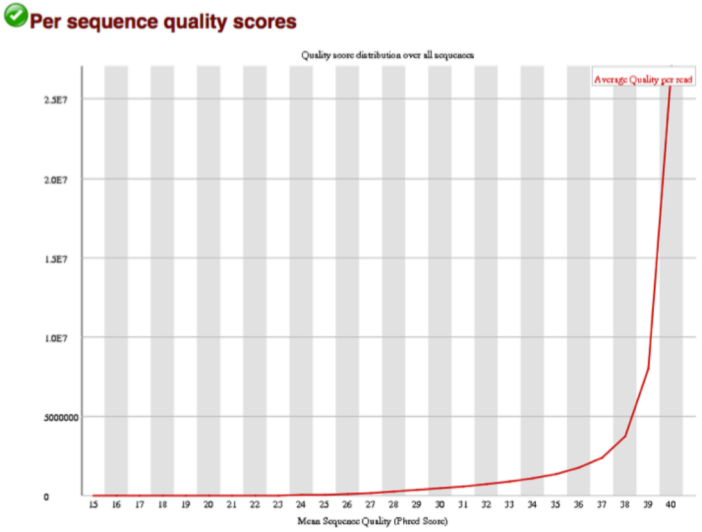
- Did our **library pep** generate a **faithful representation** of the DNA/RNA molecules in our samples?
  - ideally, the entire universe of transcripts has been sufficiently sampled (diverse library)
  - no contaminations (rRNA, foreign DNA, adapters, primers, …)
  - no bias towards fragments of certain GC contents/sizes
  - no degradation [cannot be assessed without alignment]

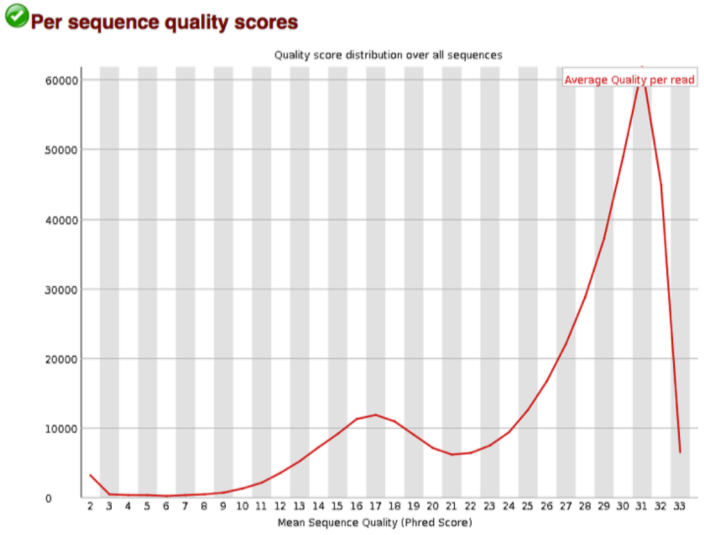
Sequencing quality per cycle

Phred score



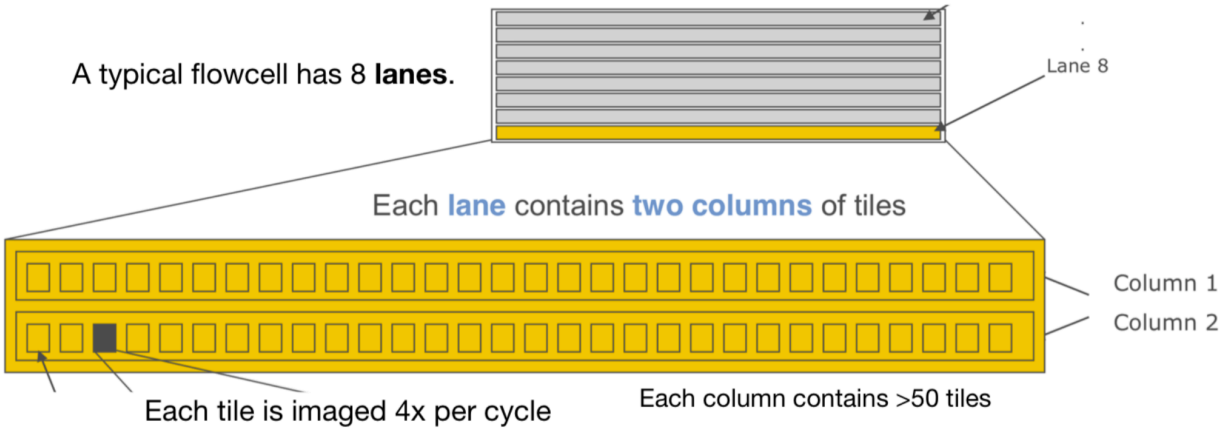




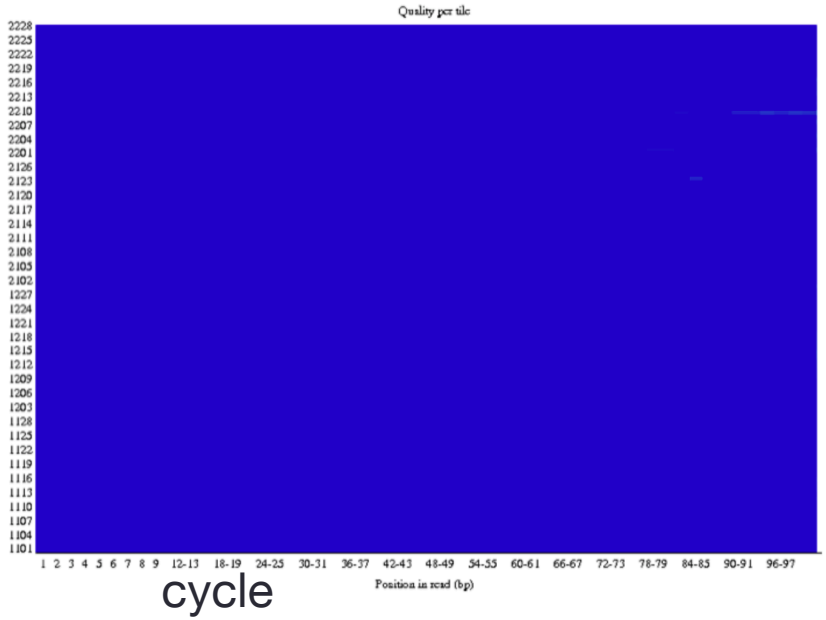


noise/uncertainty = fluorophore intensity not as clear as expected

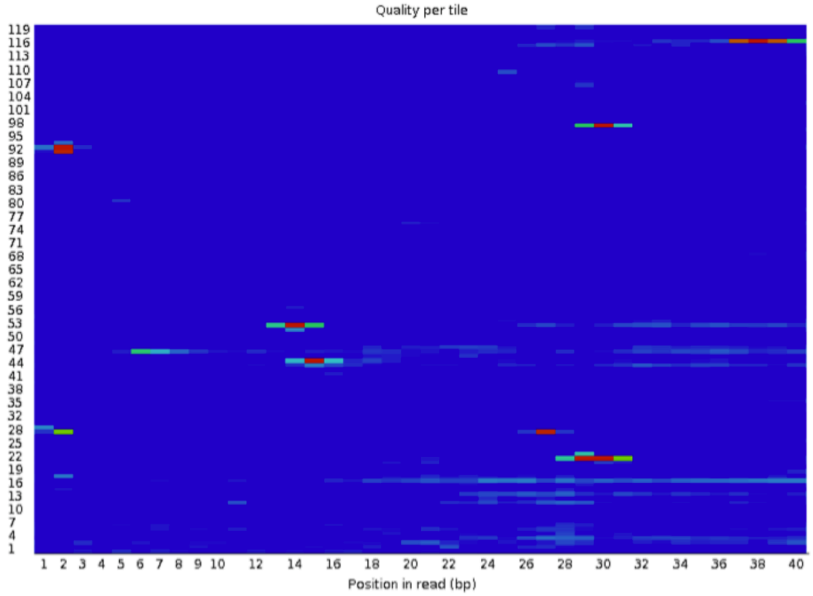
Physically localized error rates



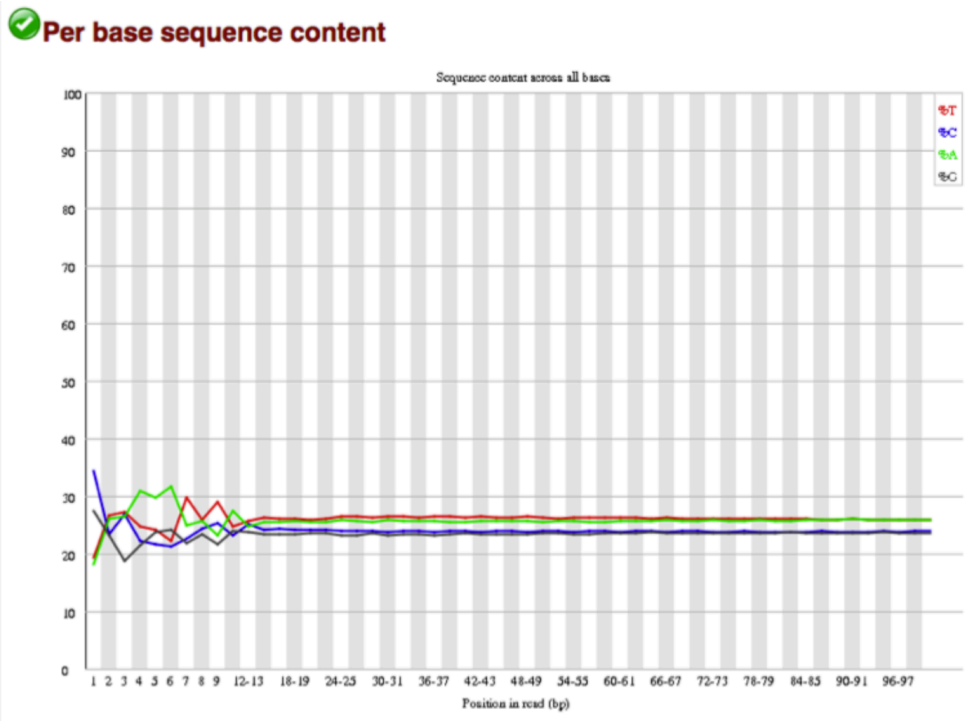
Per tile sequence quality



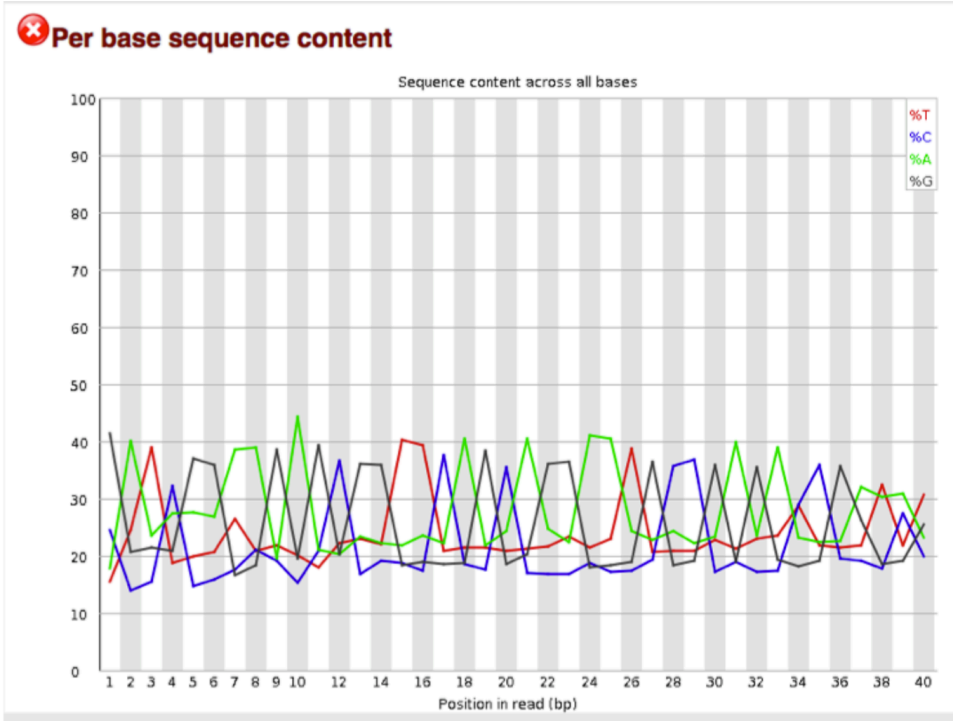
Per tile sequence quality



Sequence composition



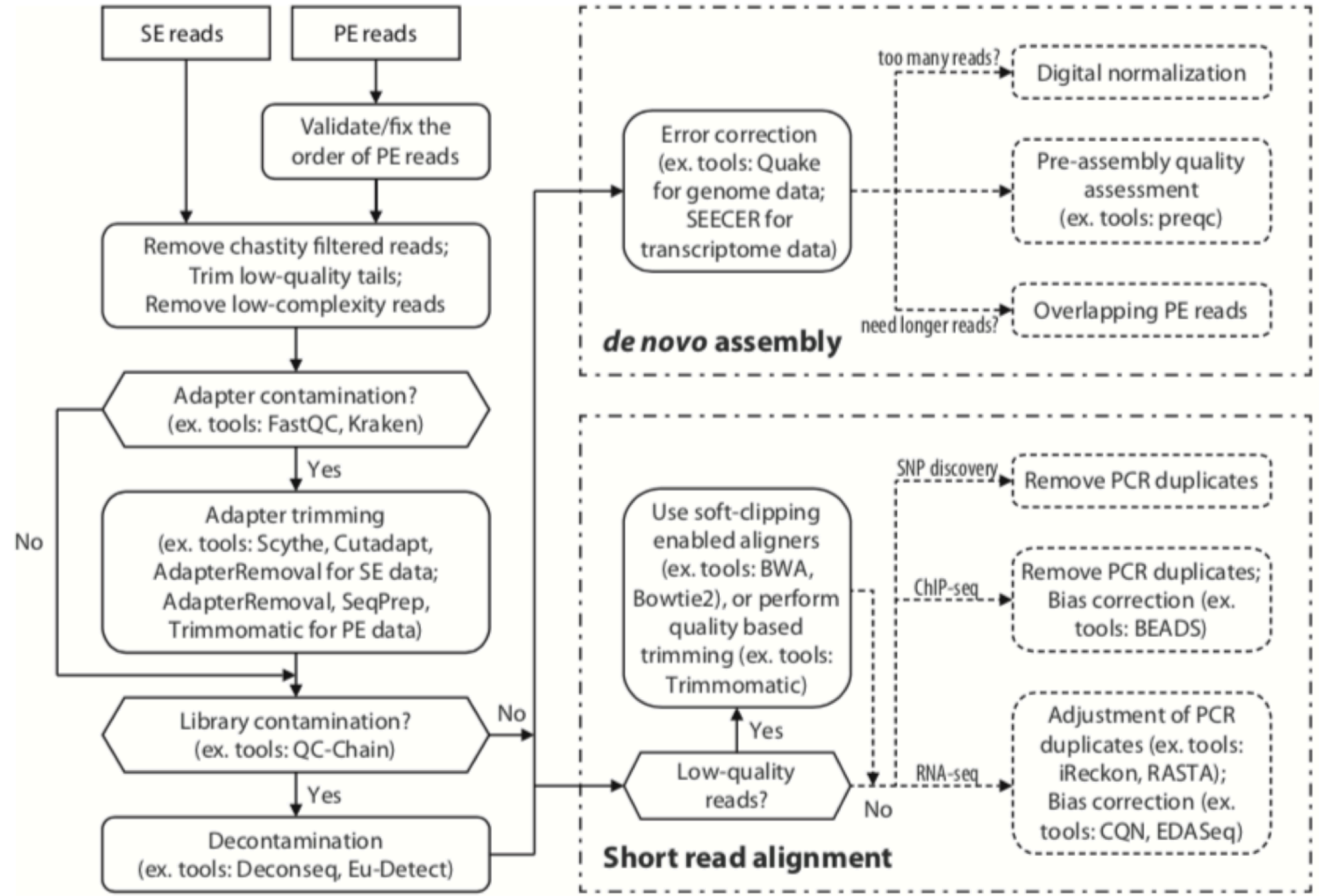
“normal” RNA-seq pattern
→ random hexamer priming not sufficiently random



highly irregular pattern
often indicative of adapter contamination

More QC details

- Zhou, X., & Rokas, A. (2014). **Prevention, diagnosis and treatment of high-throughput sequencing data pathologies.** *Molecular Ecology*, *23*(7), 1679–1700. https://doi.org/10.1111/mec.12680
- https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq
- https://sequencing.qcfail.com/
- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/



Zhou, X., & Rokas, A. (2014). doi: 10.1111/mec.12680

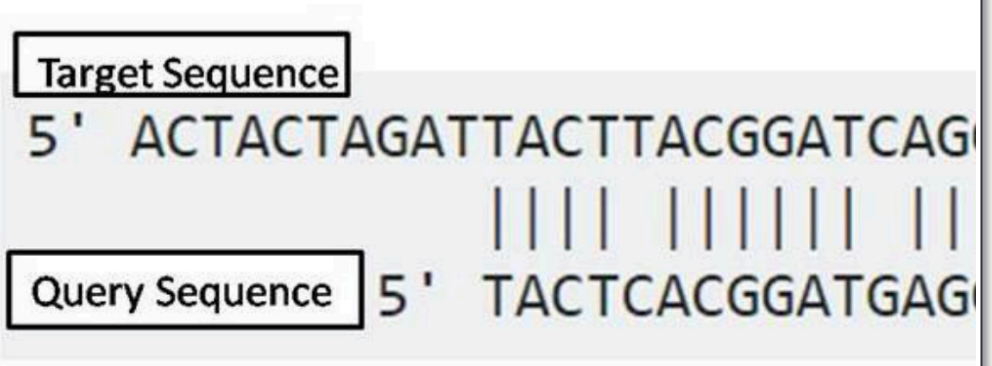
READ MAPPING

Finding out where the reads came from

Different philosophies of transcript quantification

alignment followed by **counting** of reads overlapping with genes/exons

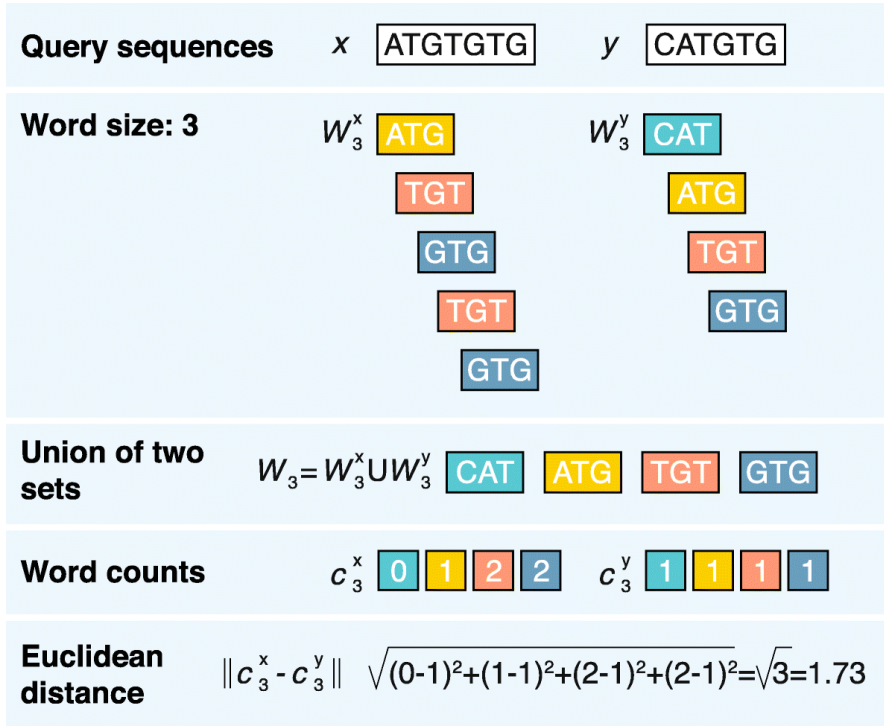
e.g. `STAR + featureCounts`



Both approaches absolutely rely on excellent reference sequences.

estimating expression levels of individual isoforms/genes based on **alignment-free k-mer matching**

`salmon, kallisto`



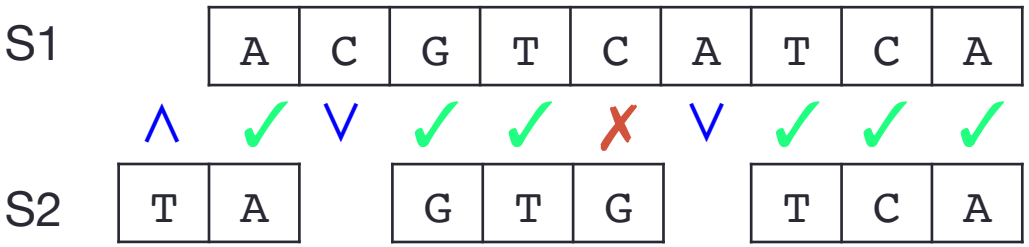
Zielezinski et al. (2017). Genome Biology. doi: 10.1186/s13059-017-1319-7

Read alignment basics

S1: A C G T C A T C A

S2: T A G T G T C A

Alignment = lining up the letters of two (or more) strings so that each letter in S1 either matches a gap or another letter in S2.



edit distance
= number of changes that are needed to match S1 and S2

To find the best alignment, we need:

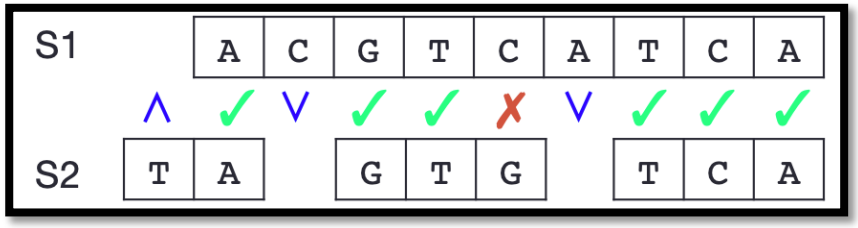
choices made by the programmer of a given tools

- **scoring function** for the edit distance
- efficient alignment-solving **algorithm**

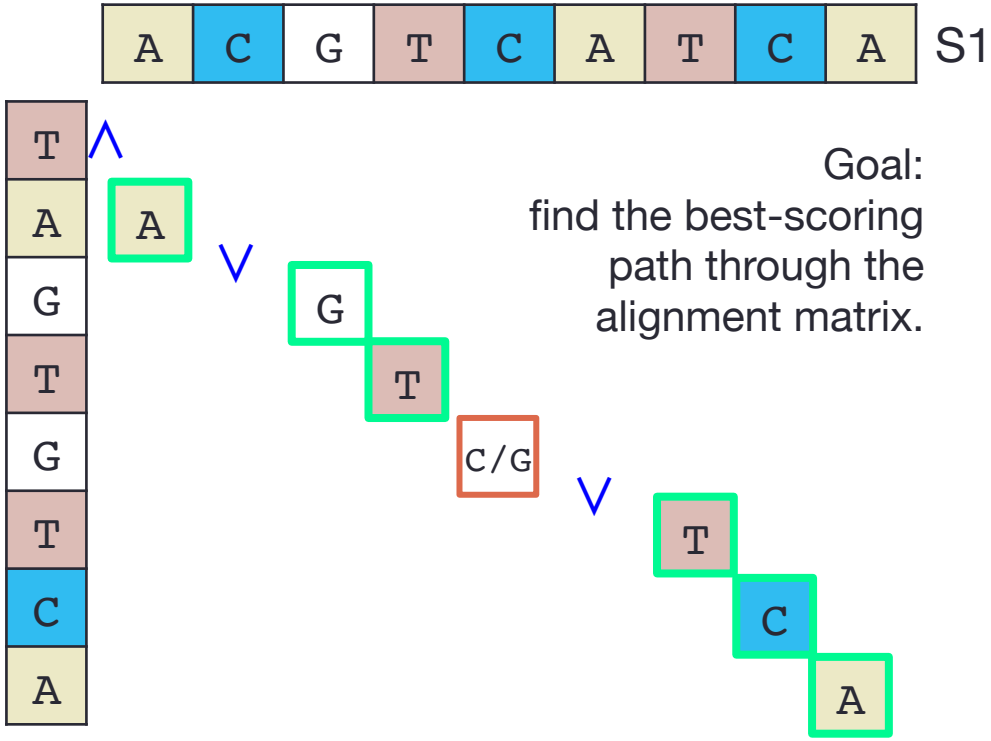
Needleman-Wunsch | Smith-Waterman | BLAST

https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-096-algorithms-for-computational-biology-spring-2005/lecture-notes/lecture5_newest.pdf

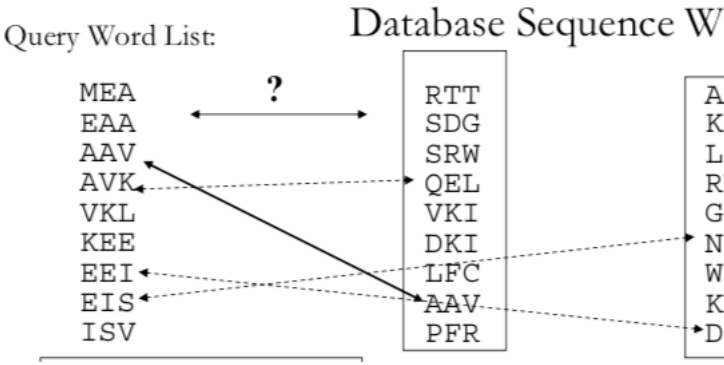
LINEAR REPRESENTATION



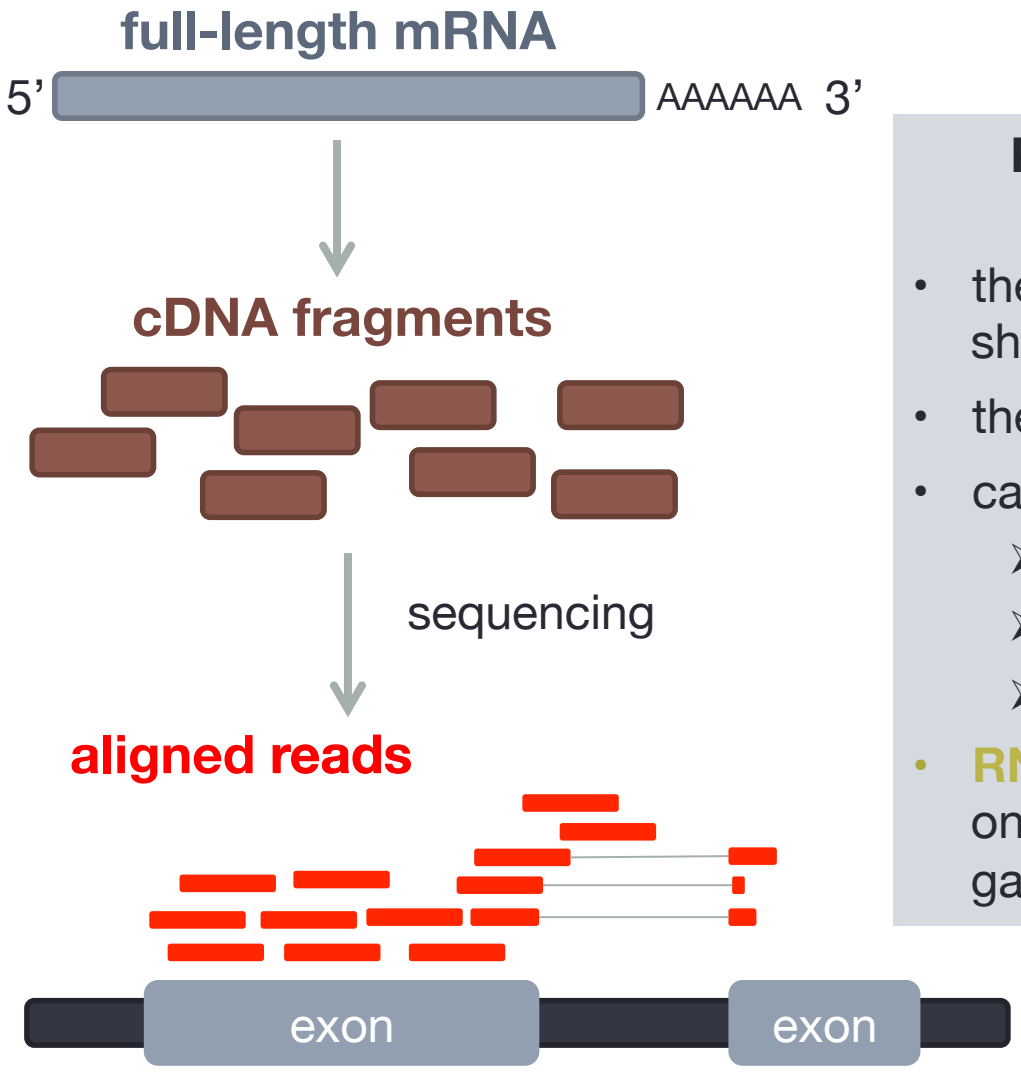
MATRIX REPRESENTATION



Goal:
find the best-scoring path through the alignment matrix.



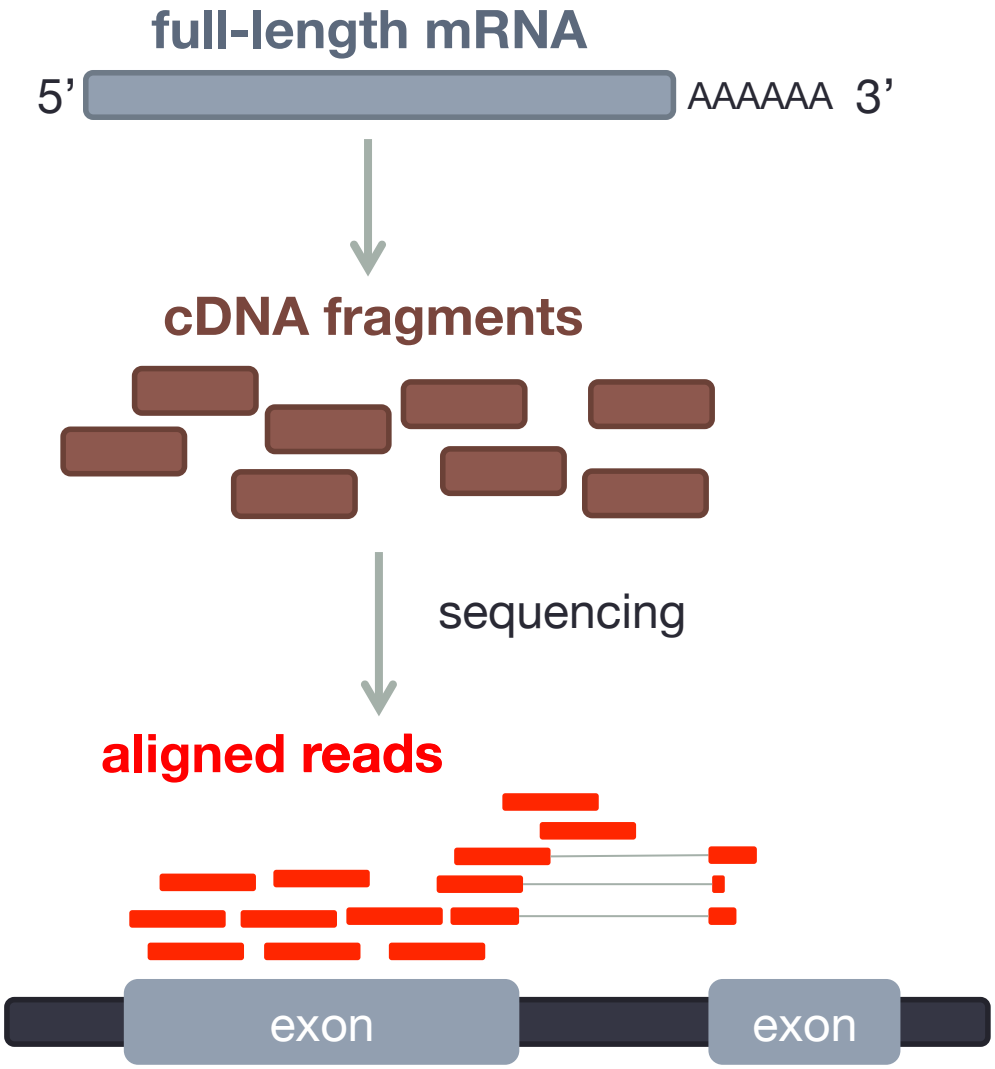
Aligning short RNA-seq reads



Particular challenges of Illumina sequencing:

- the query sequences (= reads) are very short
- there are millions of them!
- cannot expect 100% exact matches
  - seq. errors
  - biological variation
  - reference errors
- **RNA-seq**: some cDNA fragments can only be aligned if one allows for gigantic gaps (= **introns**)

Aligning short RNA-seq reads



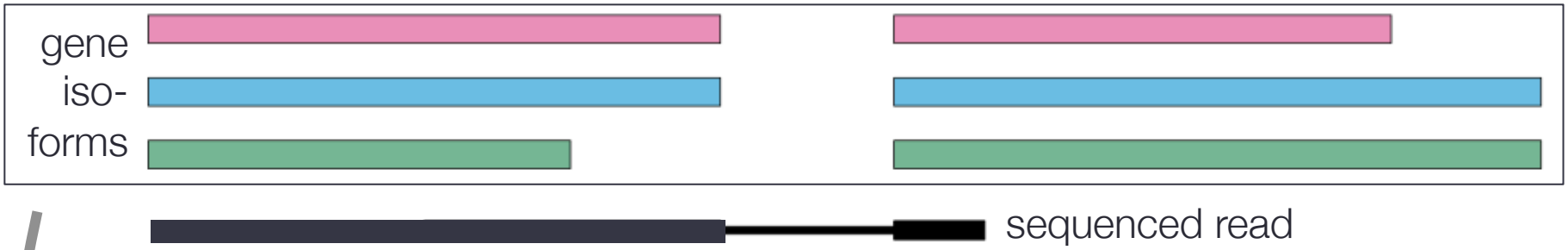
Spliced alignment tools usually need:

1) **reference genome** for the **alignment**
2) **annotation** to inform decisions about where to allow **gaps** in the alignment

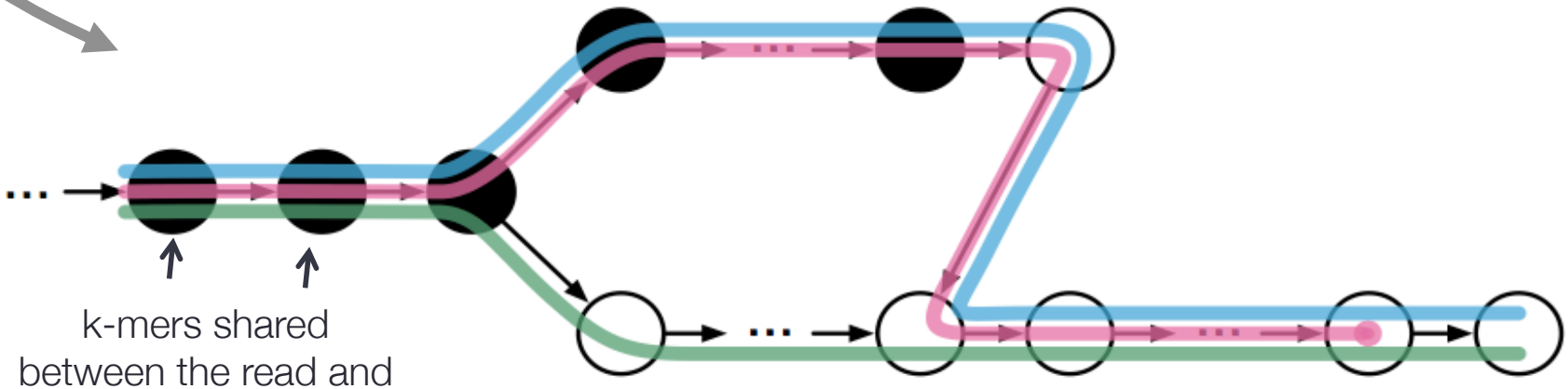
greatest downside of alignment approach: it’s resource-intensive!

… and the result is **not inherently quantitative** (it’s just read coordinates, really)!

Pseudo-alignment = alignment-free k-mer matching



sequences are **split into k-mers**, which can then be represented as nodes

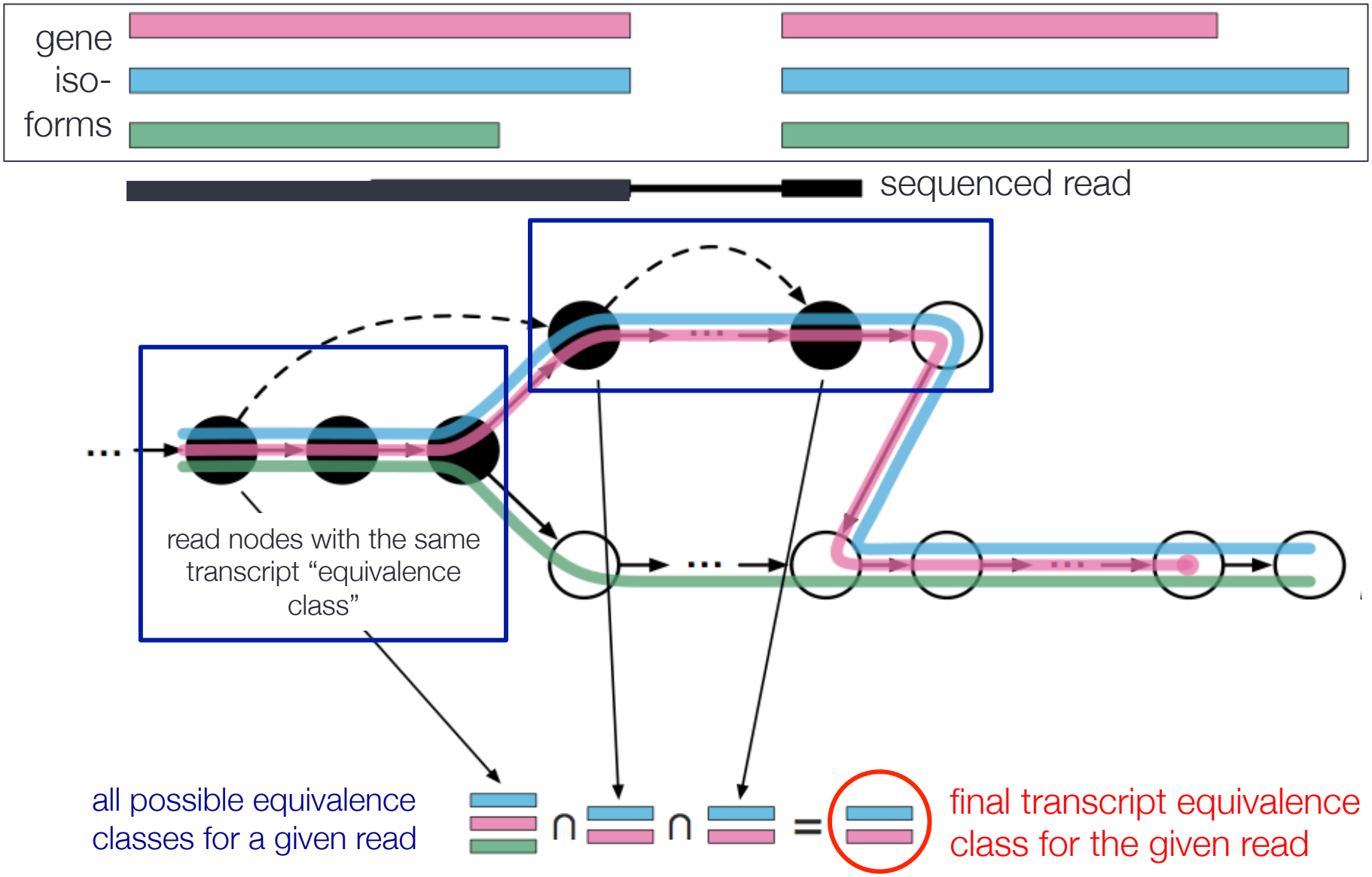


k-mers shared between the read and a transcript set

transcript set k-mers without matches in the read

Bray et al. (2016). doi: 10.1038/nbt.3519

Kallisto's pseudoalignment

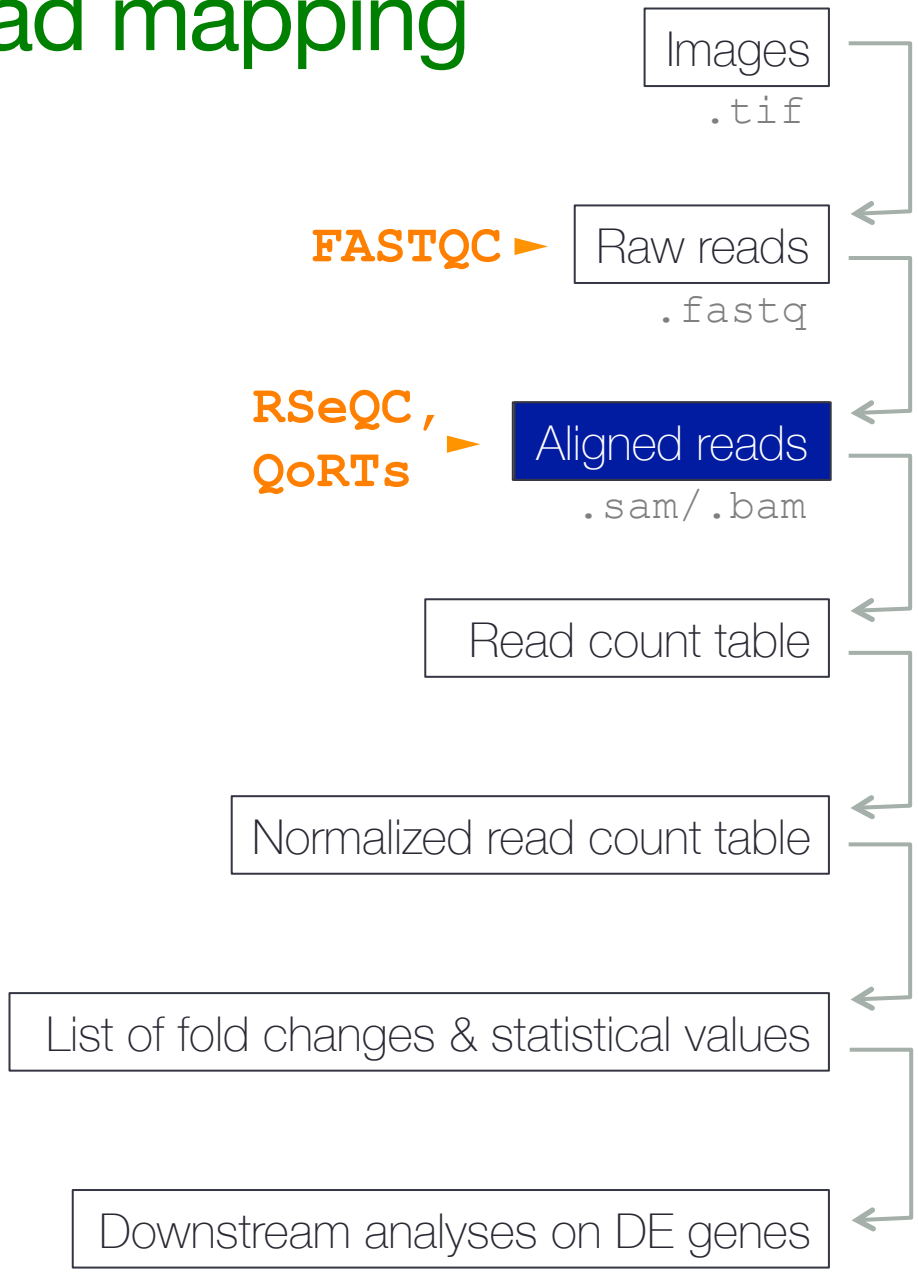


Bray et al. (2016). doi: 10.1038/nbt.3519

Alignment vs. lightweight mapping

	Alignment	Pseudo-alignment
Example workflow	`STAR + featureCounts`	`salmon`
Read mapping based on:	**Where** does a read match best?	Which **collection of unique k-mer**'s does a given read match best?
Reference needed:	**Genome** sequence + exon boundaries	**cDNA** sequences
Mapping result	Genome coordinates (`BAM`)	Table of expression level estimates (txt)
Expression quantification:	Counting how many reads overlap a gene.	Summing the values assigned to each collection of unique k-mers (equivalence class)
Output:	Read counts (integers)	Estimated transcript abundances (numeric)
Speed	++ & +++	++++

Read mapping



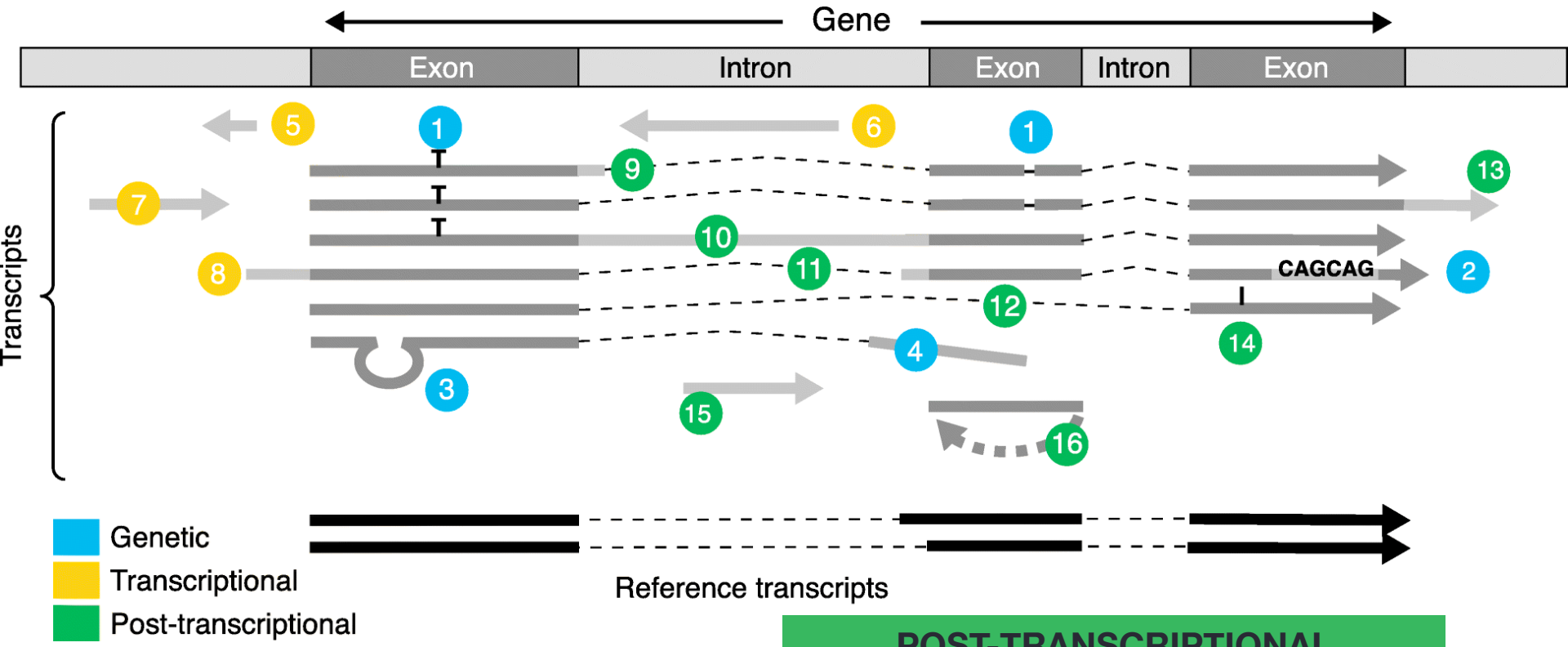
1. Theoretical background
2. Reference & annotation files
3. Files with aligned reads: SAM/BAM
4. Looking at aligned reads

Reference sequences

- reference sequences (genome, cDNA, …) were originally produced with Sanger sequencing
- most reference sequences will undergo continuous refinement (→ “genome versions”)
- **RefSeq** & **Ensembl** are two pan-species databases with homogenous computational annotation workflows
- reference *genomes* are longer, but less ambiguous than reference *transcriptome* sequences!

Reference sequences are provided in `FASTA` (!) format.
Compressed versions of `FASTA` are typically `2bit` or `fa.gz`.

Most individual RNA variations do not find their way into the reference sequences



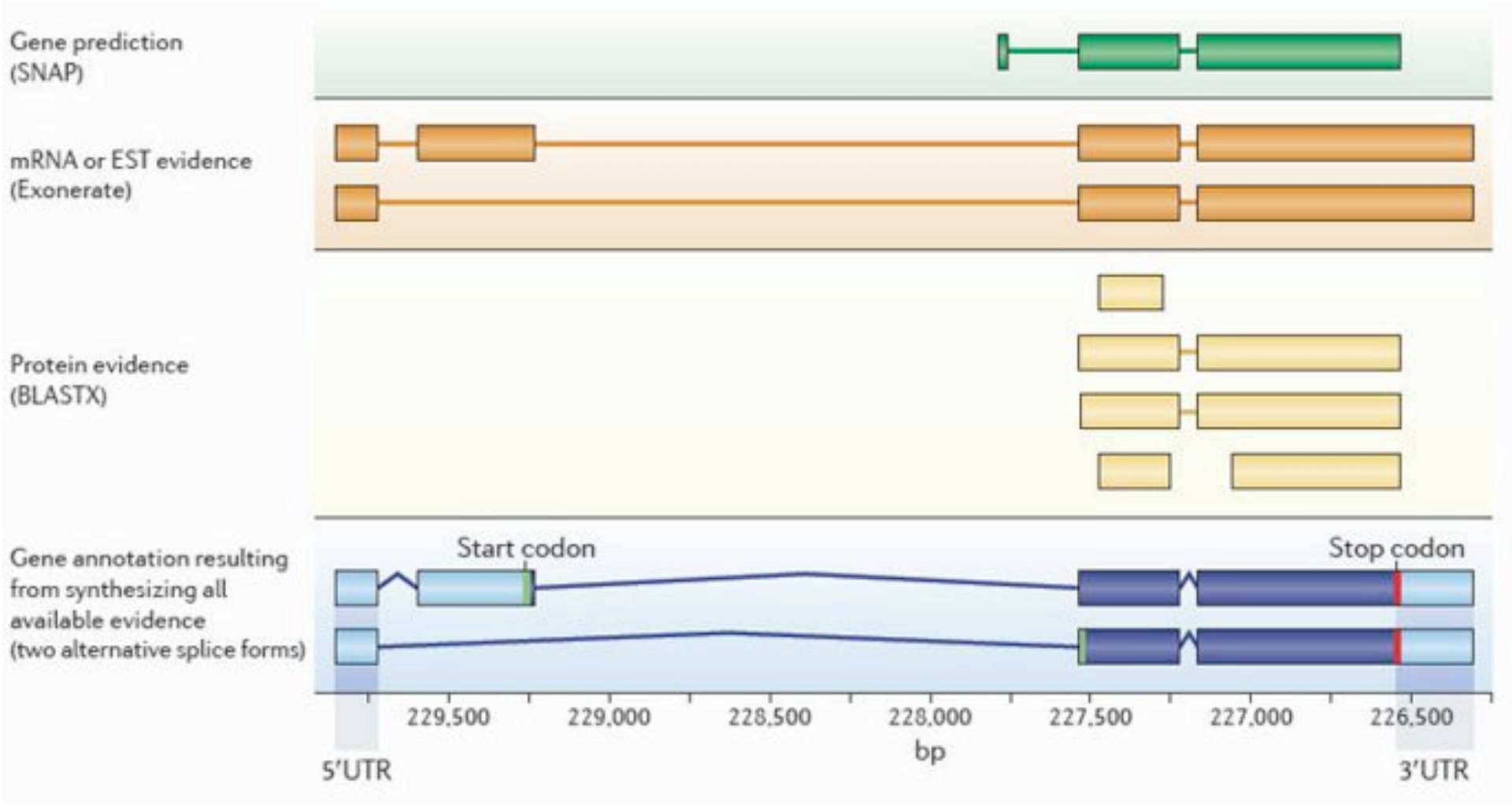
GENETIC
1 SNP or short InDel
2 microsatellite variation
3 transposition
4 gene fusion

TRANSCRIPTIONAL
5 bidirectional TSS
6 antisense transcript
7 enhancer RNA
8 alternative TSS

POST-TRANSCRIPTIONAL
9 alternative 5' splice site (SS)
10 alternative 3' SS
11 alternative 3' SS
12 skipped exon
13 alternative poly(A) site
14 editing and modification
15 processed pre-mi/ snoRNA
16 circular RNA

Morillon & Gautheret (2019) Genome Biology, 20(1), 112. doi: 10.1186/s13059-019-1710-7

Gene annotation

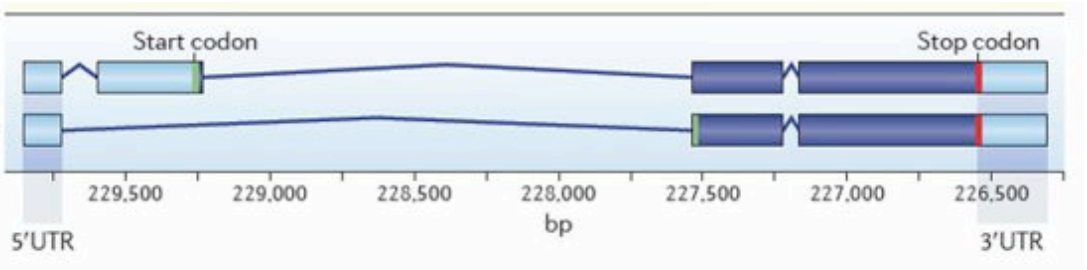


Gene annotations generally include UTRs, alternative splice isoforms and have attributes such as evidence trails.

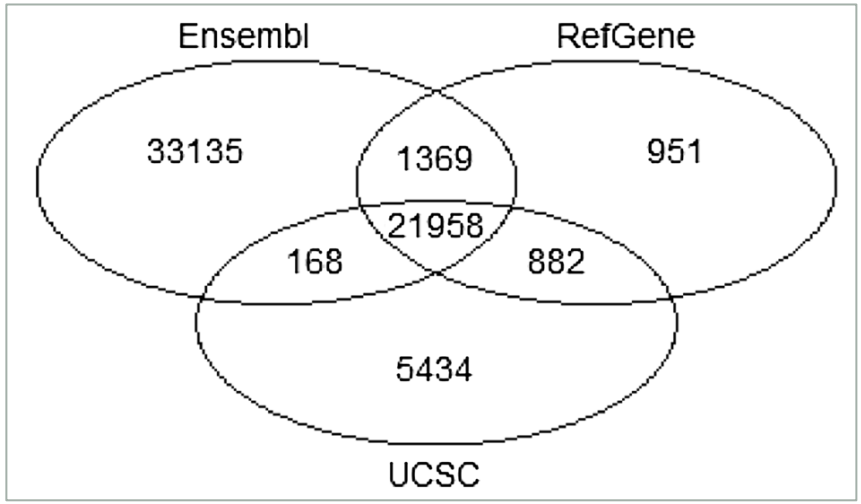
Yandell (2012). A beginner's guide to eukaryotic genome annotation. doi: 10.1038/nrg3174

Annotation: defining transcript structures

- Automated vs. manual curation ("evidence-based")
  - heterogeneous types of evidence: expressed sequence tags (ESTs), RNA-seq data, protein homologies, CDS predictions



Annotation is **dynamic**! (sequence, coordinates, types of elements)



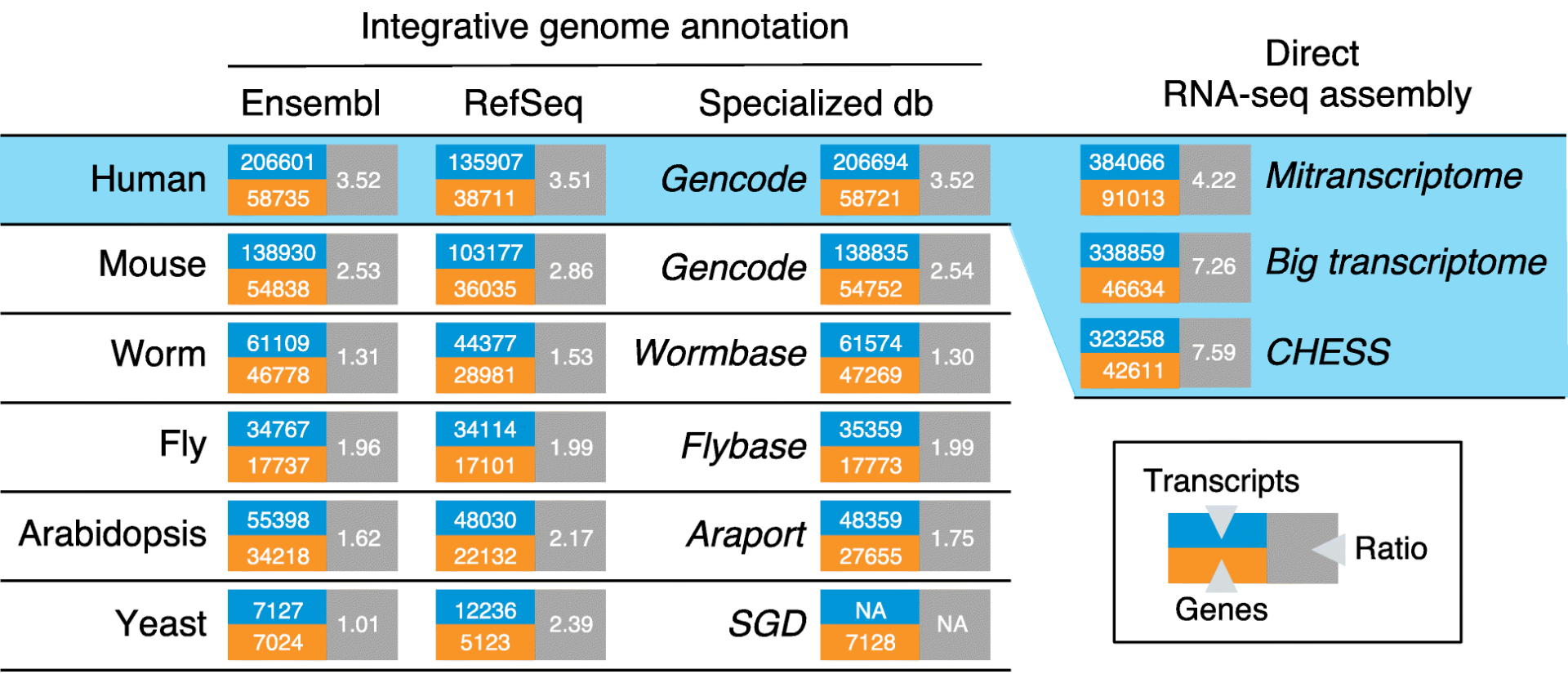
RefSeq ncbi.nlm.nih.gov/refseq

UCSC Known Genes genome.ucsc.edu

Ensembl/Gencode gencodegenes.org

1/3 protein-coding genes
> 17,000 non-coding RNAs
> 15,000 pseudogenes

Zhao & Zhang (2015) BMC Genomics. doi:10.1186/s12864-015-1308-8



Morillon & Gautheret (2019) Genome Biology, 20(1), 112. doi: 10.1186/s13059-019-1710-7

Which annotation should one use?

*"More sensitive annotations, such as **Ensembl** (…) **should be preferred** over more specific annotations, such as RefSeq (…) if the aim is to obtain accurate expression estimates."*

Janes et al. (Briefings in Bioinformatics, 2015). doi: 10.1093/bib/bbv007

*"We observe that **RefSeq Genes produces the most accurate fold-change measures** with respect to a ground truth of RT-qPCR gene expression estimates. "*

Wu et al. (BMC Bioinfo, 2013). doi: 10.1186/1471-2105-14-S11-S8

*"In practice, there is **no simple answer to this question**, and it depends on the purpose of the analysis. (…) When choosing an annotation database, researchers should keep in mind that **no database is perfect** and **some gene annotations might be inaccurate or entirely wrong**."*

Zhao & Zhang (BMC Genomics, 2015). doi:10.1186/s12864-015-1308-8

Storing annotation information

- representing genome coordinates + description/name
  - intron–exon structures, start and stop codons, UTRs, alternative transcripts
- various formats (all are plain text files): GFF2, GFF3, GTF, BED, SAF…

see the course notes for details

GTF ("GFF2.5")

1. reference coordinate
2. source
3. annotation type
4. start position
5. end position
6. score
7. strand
8. frame/phase
9. attributes: <TYPE VALUE>; <TYPE VALUE>; <TYPE VALUE>

GFF2

```
# GFF-version 2
IV      curated exon    5506900 5506996 . + .   Transcript B0273.1
IV      curated exon    5506026 5506382 . + .   Transcript B0273.1
IV      curated exon    5506558 5506660 . + .   Transcript B0273.1
IV      curated exon    5506738 5506852 . + .   Transcript B0273.1

# GFF-version 3
ctg123  .  exon  1300  1500  .  +  .  ID=exon00001
ctg123  .  exon  1050  1500  .  +  .  ID=exon00002
ctg123  .  exon  3000  3902  .  +  .  ID=exon00003
ctg123  .  exon  5000  5500  .  +  .  ID=exon00004
ctg123  .  exon  7000  9000  .  +  .  ID=exon00005
```

GFF3

GTF

```
# example for the 9th field of a GTF file
    gene_id "Em:U62.C22.6"; transcript_id "Em:U62.C22.6.mRNA"; exon_number 1
```

0 vs. 1 based conventions

one-based, fully-closed

ATG location: 7 - 9 or [7,9]
Cut site: 11^12 or (11,12)
Interval length = stop - start + 1

GFF format

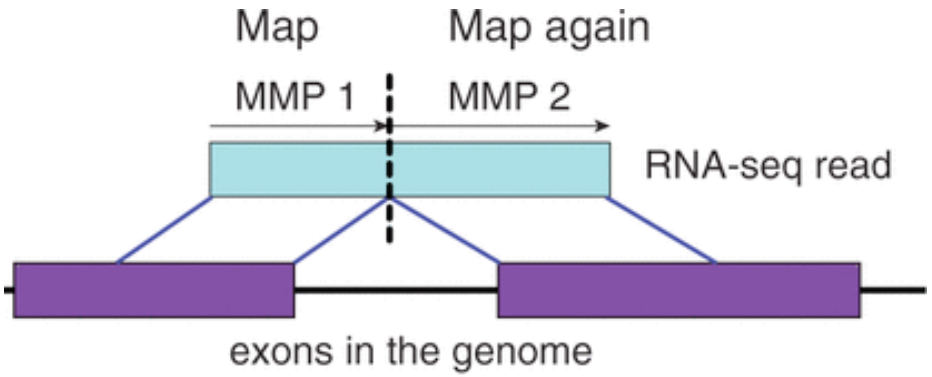
zero-based, half open

ATG location: 6 - 9 or [6,9)
Cut site: 11-11 or [11,11)
Interval length = stop - start

BED format

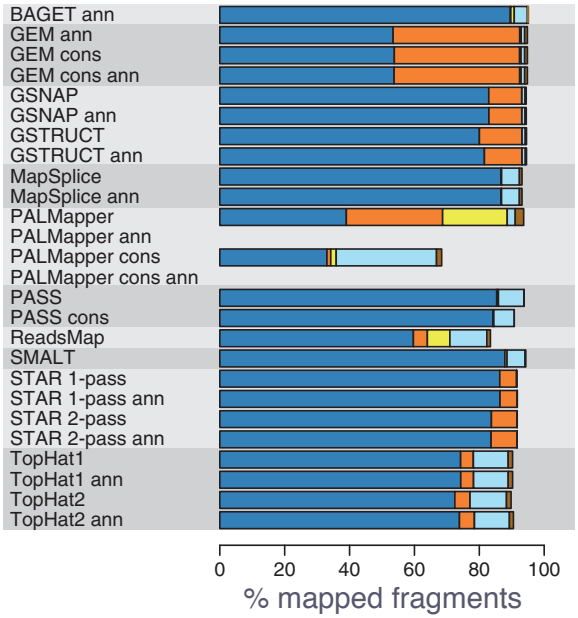
http://alternateallele.blogspot.com/2012/03/genome-coordinate-cheat-sheet.html

Spliced Transcriptome Alignment to Reference (STAR)



- accurate & sensitive
- very fast
- memory intensive! (use it on the server!)

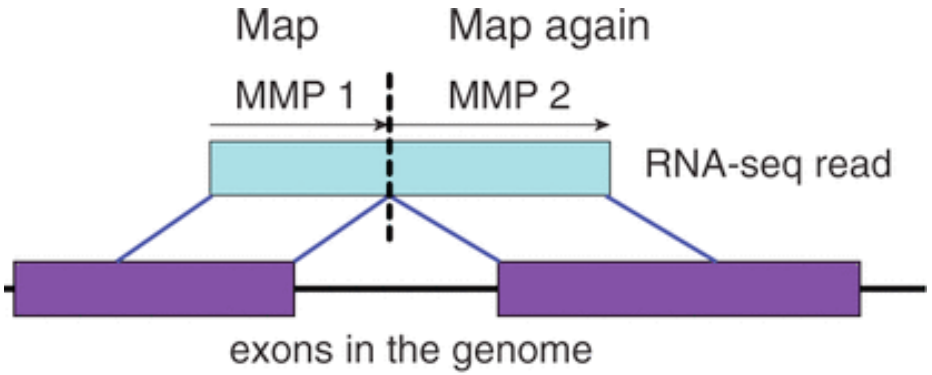
Spliced alignment programs



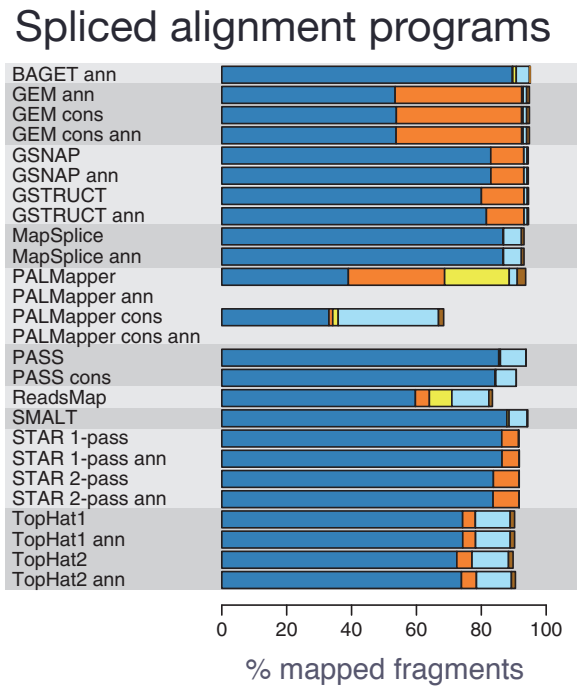
Engström et al. (2013) Nature Methods, 10(12), 1185–1191. doi:10.1038/nmeth.2722

- MMP = maximal mappable prefix (aka *maximum matching portion*)
- reads are split when a continuous alignment is not possible
- the remaining unmappable portion is then aligned again
- finally, aligned portions of the original full-length reads are stitched together

STAR spliced alignment



- accurate & sensitive
- very fast
- memory intensive!



Engström et al. (2013) Nature Methods, 10(12), 1185–1191. doi:10.1038/nmeth.2722

STAR has myriad options! Tune them to meet your needs

Current Protocols in Bioinformatics (Sept 2015)
DOI: 10.1002/0471250953.bi1114s51
and
STARmanual.pdf

2 main STAR modules

1. generate **genome index**

```
--runMode genomeGenerate
--genomeFastaFiles sacCer3.fa
--sjdbGTFfile sacCer3.gtf
```

needs to be done just 1x per transcriptome!

2. **align**

   2.1. align to *reference* & identify novel splice junctions

   2.2 re-run alignment including the novel splice junctions

```
$runSTAR –genomeDir STARindex/ \
         --readFilesIn $FASTQ_FILES \
         --readFilesCommand zcat \
```

```
--twopassMode
```

must be done for every sample

Let's align the reads for WT_1!