Lesson 8

Expression data analysis part I

Quick querying of TSV files with awk

- awk is a simple tool for text manipulation
- Can be used to select specific records from a TSV file (e.g. SAM/VCF/other)

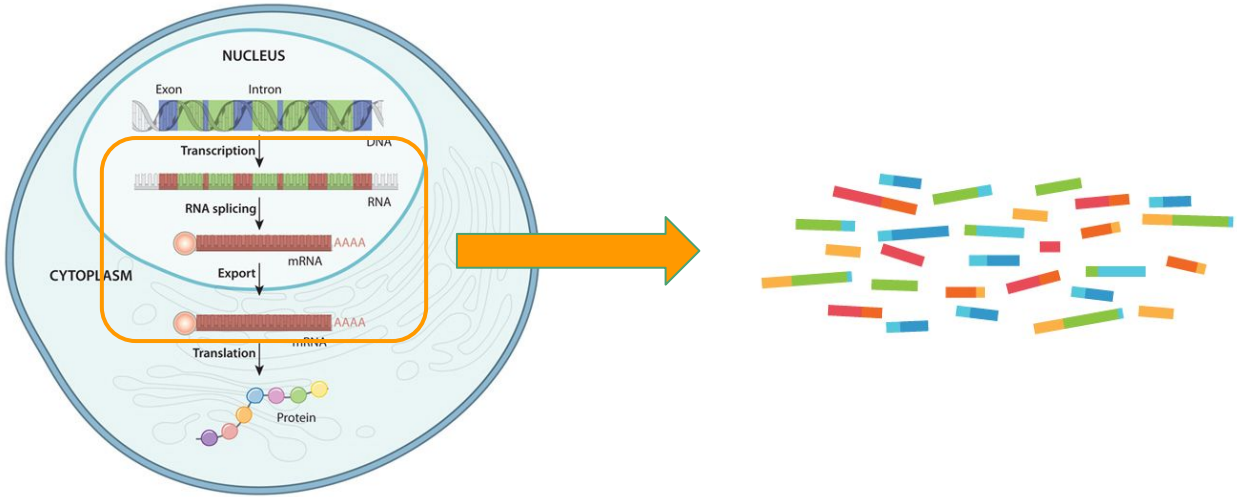
```
# Extract VCF variants on chromosome II
$ awk '$1 == "ChrII"' SRR1569760_vs_S288C_reference.vcf
# Extract VCF variants with REF allele G
$ awk '$4 == "G"' SRR1569760_vs_S288C_reference.vcf
# Extract VCF variants with QUAL < 20
$ awk '$6 < 20' SRR1569760_vs_S288C_reference.vcf
# Boolean expressions
awk '$1 == "ChrII" && ($6 < 20 || $5 == "G")'
SRR1569760_vs_S288C_reference.vcf
```

By the end of this lesson you will...

- Understand some common RNA-seq uses and protocols
- Be familiar with the basic workflow of gene expression data analysis
    - Specifically differential gene expression analysis
- Know how to map RNA-seq reads to a reference genome using STAR
- Be able to QA mapping results using Qualimap

What is RNA-seq?

- Sequencing of RNA using NGS technology
- Allows assessment of presence and quantity or RNA in a sample
- Take a snapshot of expression in a sample



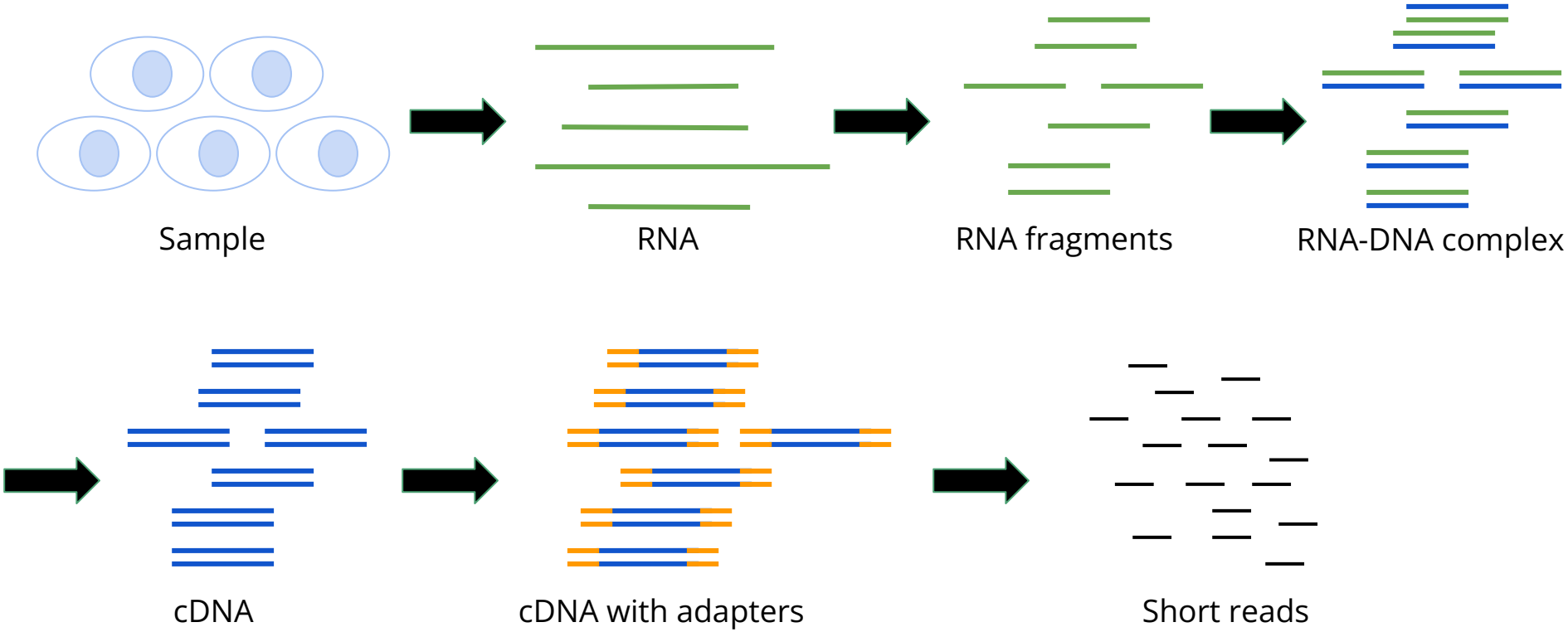
Why study the transcriptome?

- Indication of cell physiology
- Dynamic - responds to the environment
  - Changes over time
  - Responds to external stimuli
  - Controls cellular processes
- Reduced representation of the genome
  - Smaller = cheaper
  - Only the “functional” parts of the genome

Types of expression analysis

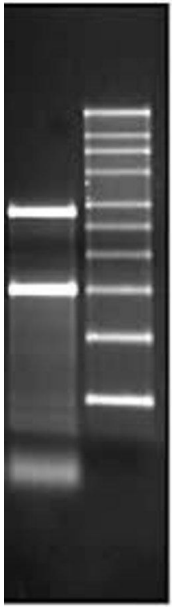
- Expression quantification
- Differential gene expression
- Assemble whole transcriptomes
- Detect new transcripts
- Detect splicing variants
- Detect allele-specific expression
- Gene co-expression
- Single-cell analysis

RNA-seq - basic protocol

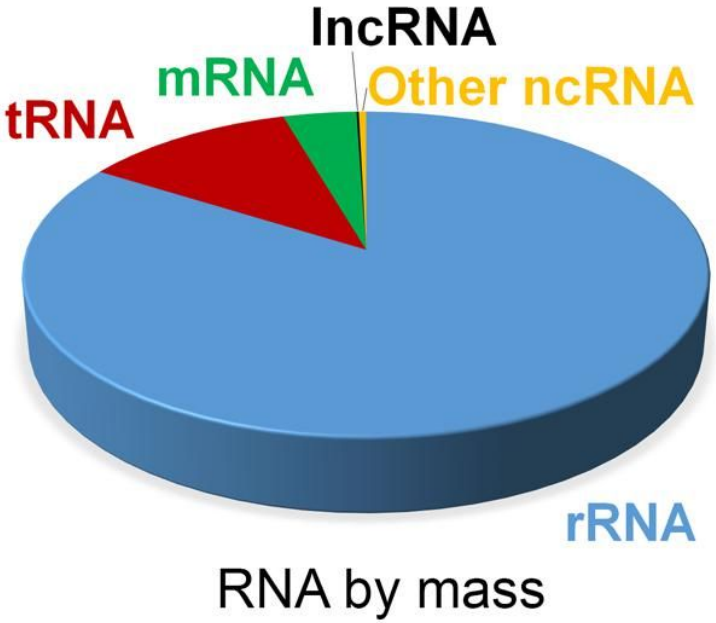


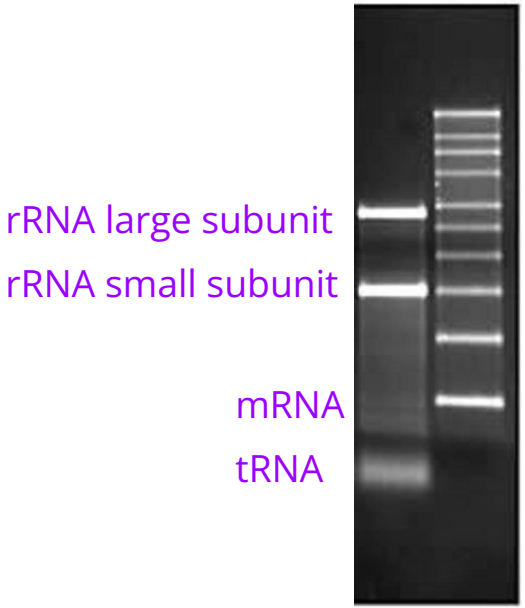
Total RNA

Which band contains the mRNA?



Total RNA composition (mammalian cell)



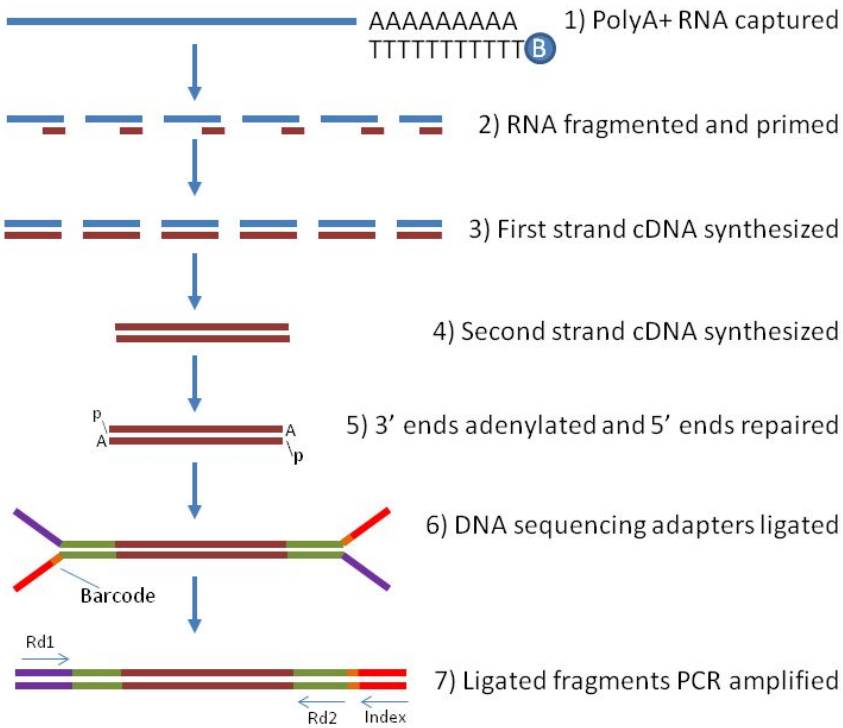


Palazzo, Alexander F., and Eliza S. Lee. "Non-coding RNA: what is functional and what is junk?." *Frontiers in genetics* 6 (2015): 2.

Enriching for mature mRNA

- **Poly-A selection** method
  - use poly-dT baits to bind mRNAs and discard the rest
  - Removes rRNA, tRNA and others
  - Enriches for mature mRNA (containing poly-A tail)
  - Not all mRNAs have poly-A tails
    - Histones mRNA in Metazoans
    - Mitochondrial mRNA
- **rRNA depletion** method
  - Use baits designed specifically for rRNA
  - Does not remove tRNA
  - Only option in bacteria (no poly-A tail)
  - Only option when extracting small RNAs

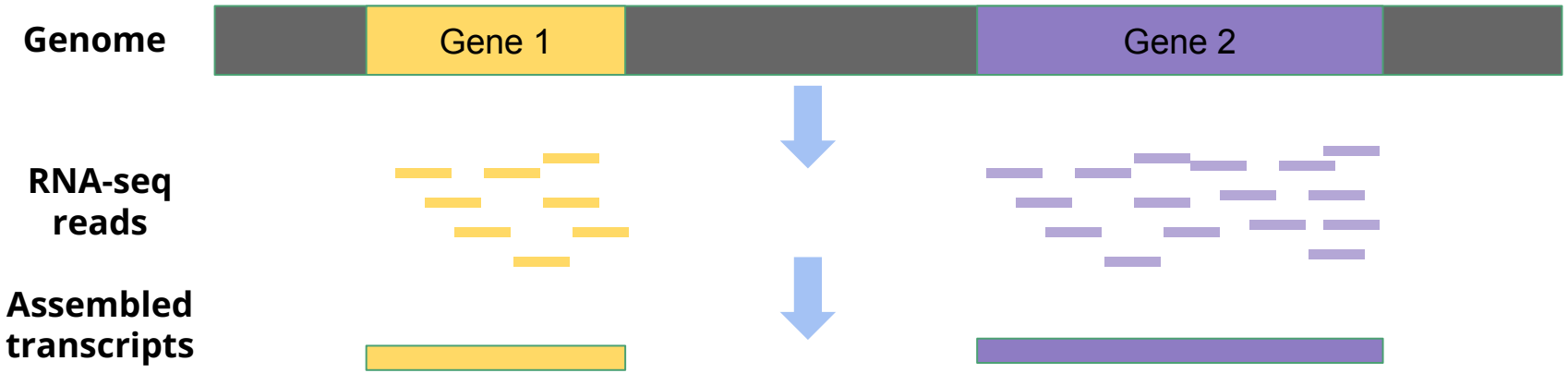
RNA-seq library prep



https://www.labome.com/method/RNA-seq.html

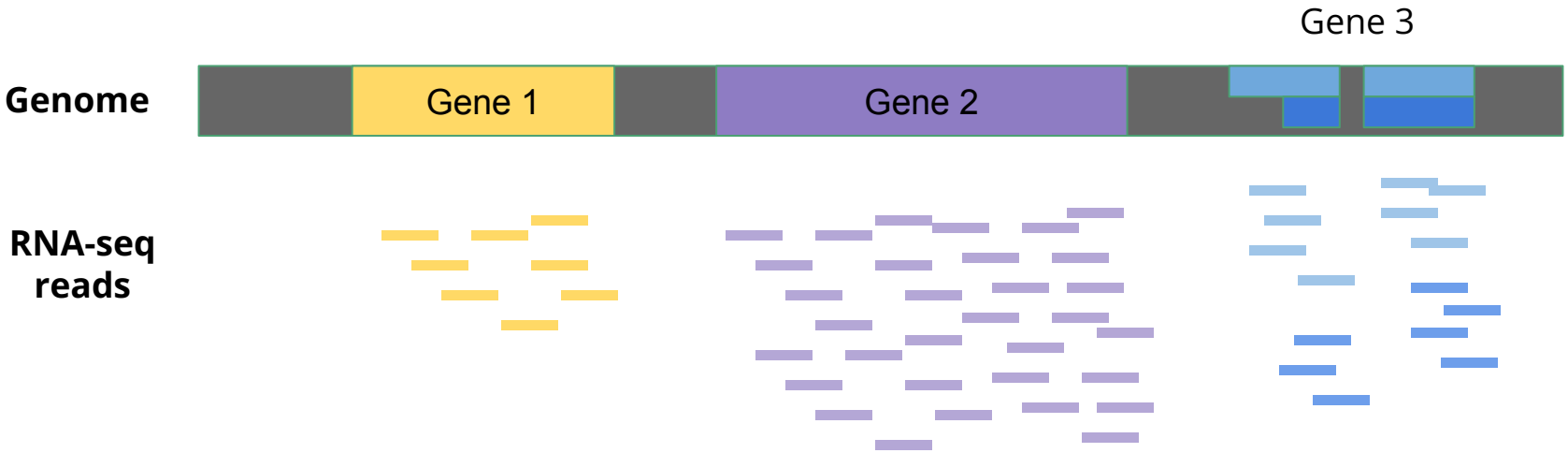
Transcriptome assembly

- Stitch together short RNA-seq reads to recover full transcript sequences
- De novo or genome-guided
- Useful for:
  - Discovery of new genes
  - Discovery of splice variants
  - Low-budget alternative for genome assembly



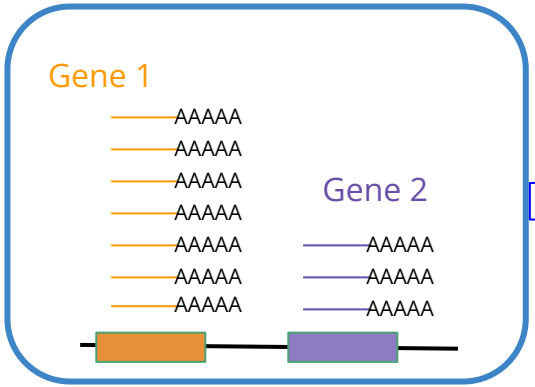
Transcriptome assembly vs. genome assembly

- Non-uniform sequencing coverage (depends on expression level
- Splice variants from the same gene are highly similar
- “Fragmented” by nature



Measuring gene expression levels with RNA-seq

Higher expression → more transcripts → more RNA-seq reads



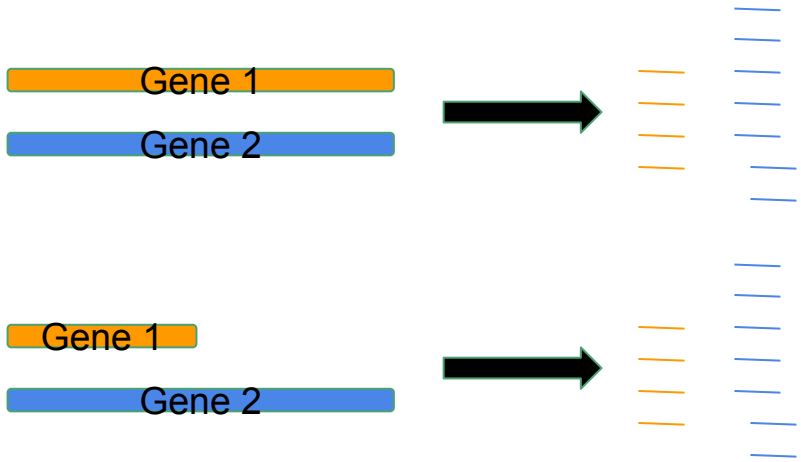
RNA-seq

Count

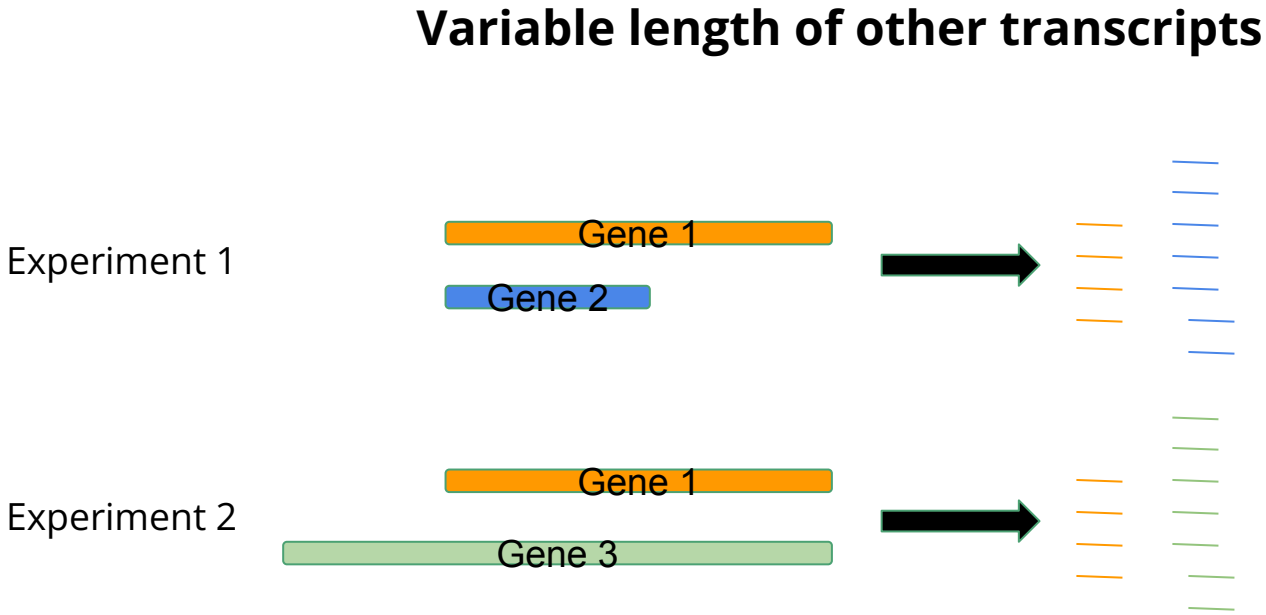
Gene	Reads
Gene 1	7
Gene 2	3

Are read counts a good measure?

Variable length of transcript of interest

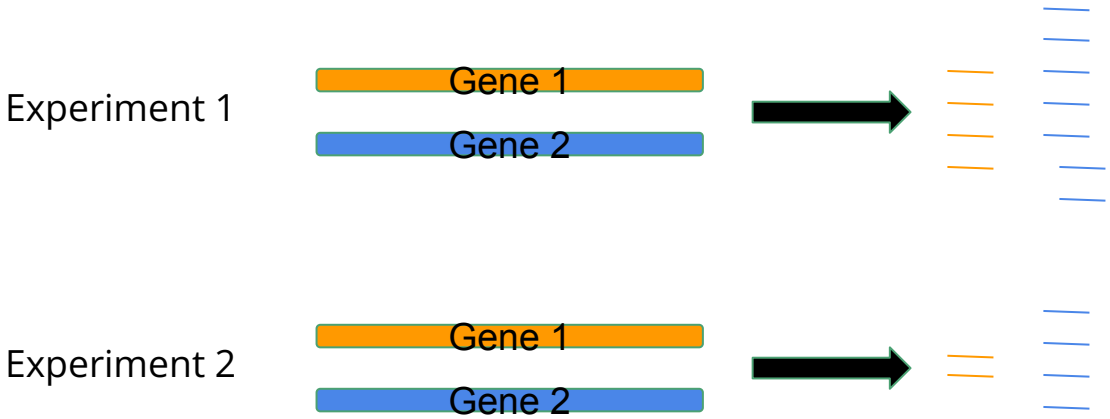


Are read counts a good measure?



Are read counts a good measure?

Variable number of reads between experiments



Read count normalization

- Normalize for transcript length
  RPK - **reads per kilobase** (of transcript)

$$RPK_i = 10^3 \cdot \frac{n_i}{l_i}$$

- Normalize for sequencing depth
  RPKM -
  reads per kilobase (of transcript) **per million** (reads)

n_i - number of reads from transcript *i*
l_i - length (in bp) of transcript *i*

$$RPKM_i = 10^9 \cdot \frac{n_i}{l_i \cdot \sum_j n_j}$$

Let's practice...

$$RPK_i = 10^3 \cdot \frac{n_i}{l_i}$$

$$RPKM_i = 10^9 \cdot \frac{n_i}{l_i \cdot \sum_j n_j}$$

Experiment 1 - total reads: 100,000

Gene	Transcript length	Reads	RPK	RPKM
Gene1	500	10		
Gene2	1000	20		

Experiment 2 - total reads: 1,000,000

Gene	Transcript length	Reads	RPK	RPKM
Gene1	500	10		
Gene2	1000	50		

Let's practice...

$$RPK_i = 10^3 \cdot \frac{n_i}{l_i}$$

$$RPKM_i = 10^9 \cdot \frac{n_i}{l_i \cdot \sum_j n_j}$$

Experiment 1 - total reads: 100,000

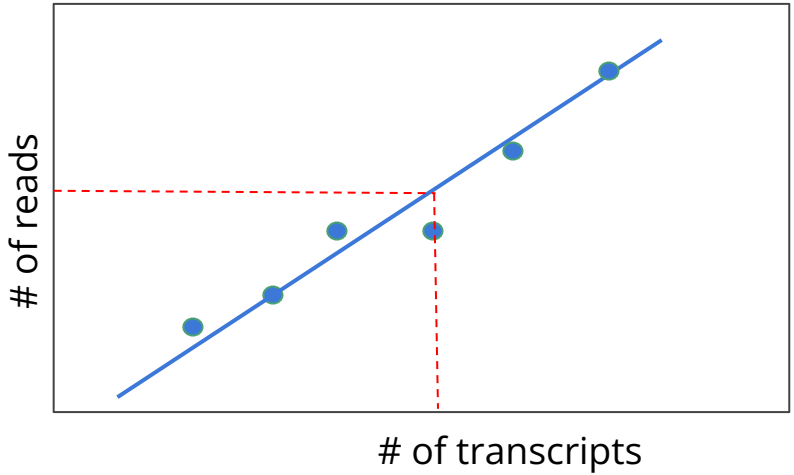
Gene	Transcript length	Reads	RPK	RPKM
Gene1	500	10	20	200
Gene2	1000	20	20	200

Experiment 2 - total reads: 1,000,000

Gene	Transcript length	Reads	RPK	RPKM
Gene1	500	10	20	20
Gene2	1000	50	50	50

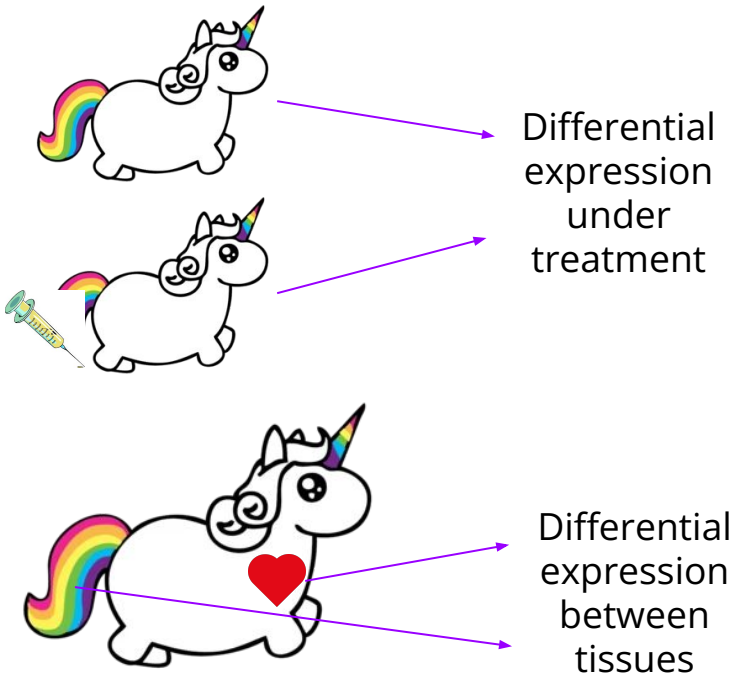
Relative vs. absolute expression

- RNA-seq produces **relative** measures of gene expression
- This is regardless of any normalization method we use
- If we want the absolute expression level we can use **spike-in** controls**:**
  1) Add a set of transcripts from another organism in known quantities to your library
  2) Sequence
  3) Use spike-in controls to calibrate

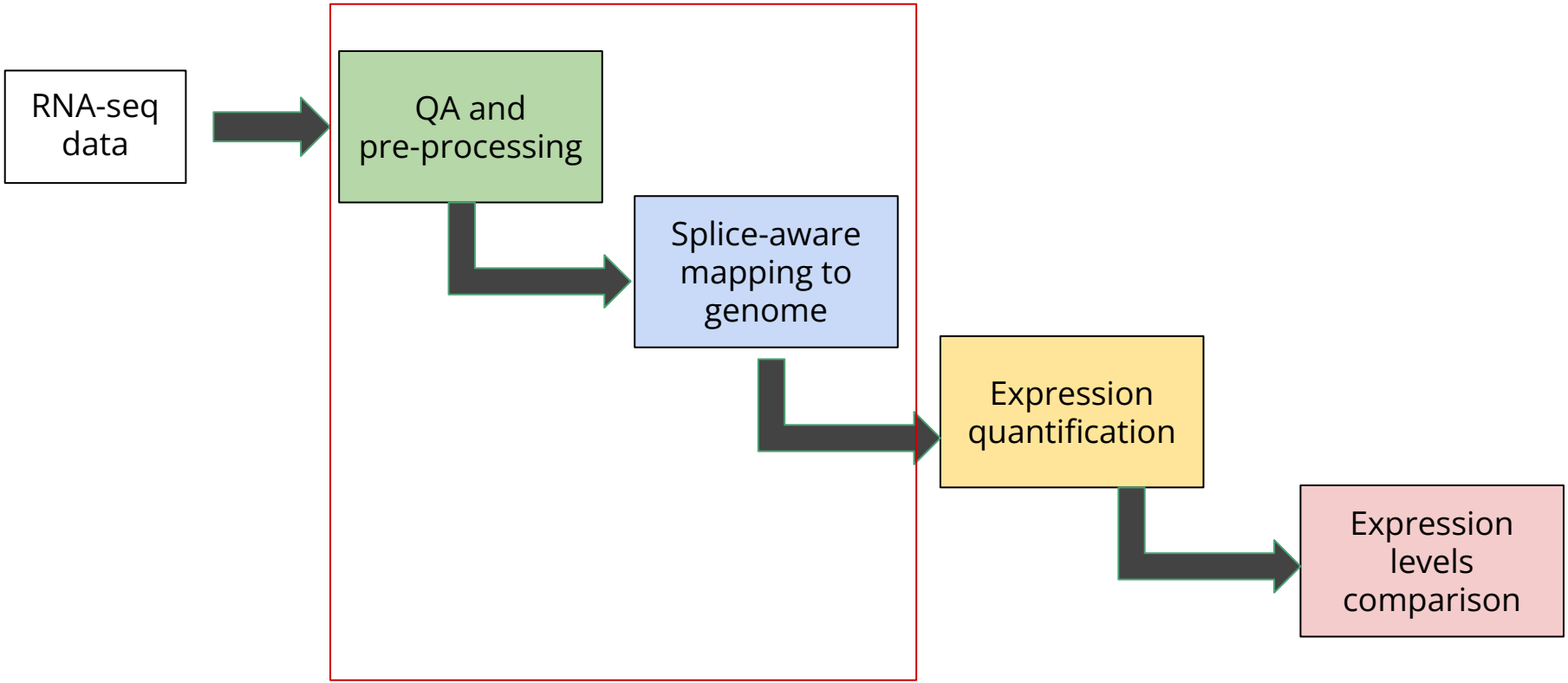


Differential gene expression

- Compare gene expression levels across all genes between two (or more) samples
- Samples of the same cell type
  - Different physiological conditions
  - Different environmental conditions
    - Growth medium
    - Treatment
- Samples of different cell types
  - Different tissues
  - WT vs. mutant
  - Different strains
  - Normal vs. tumor

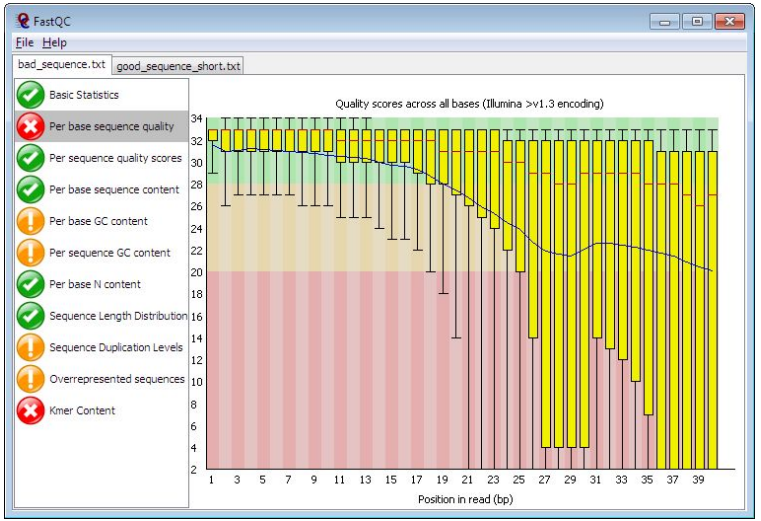


Differential gene expression workflow



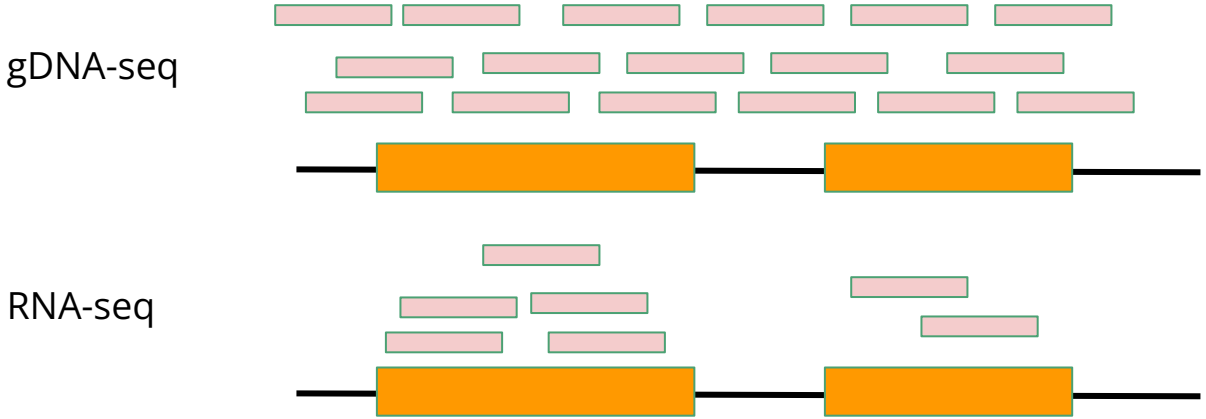
RNA-seq data QA and pre-processing

- Same as for genomic data!
- FastQC for QA stats
- Trimmomatic for adapter and low quality trimming
- Remove duplicate reads



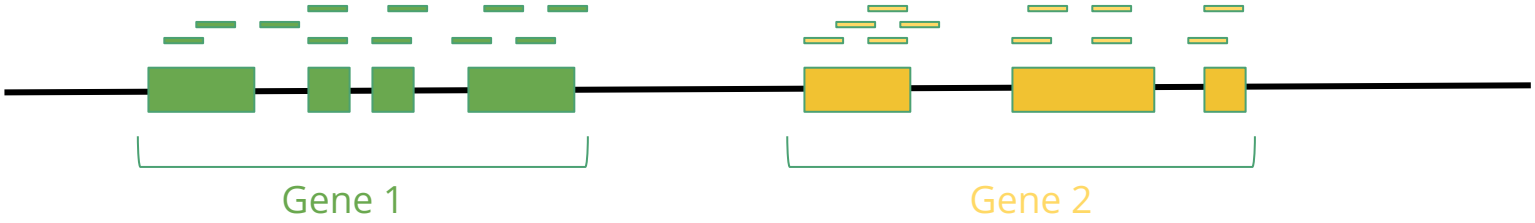
RNA-seq depth...?

- Average depth - how many reads cover each base **on average**
- But with RNA-seq “depth” depends on gene expression levels
- So talking about sequencing depth is **meaningless**
- Instead, we just indicate how many reads were generated



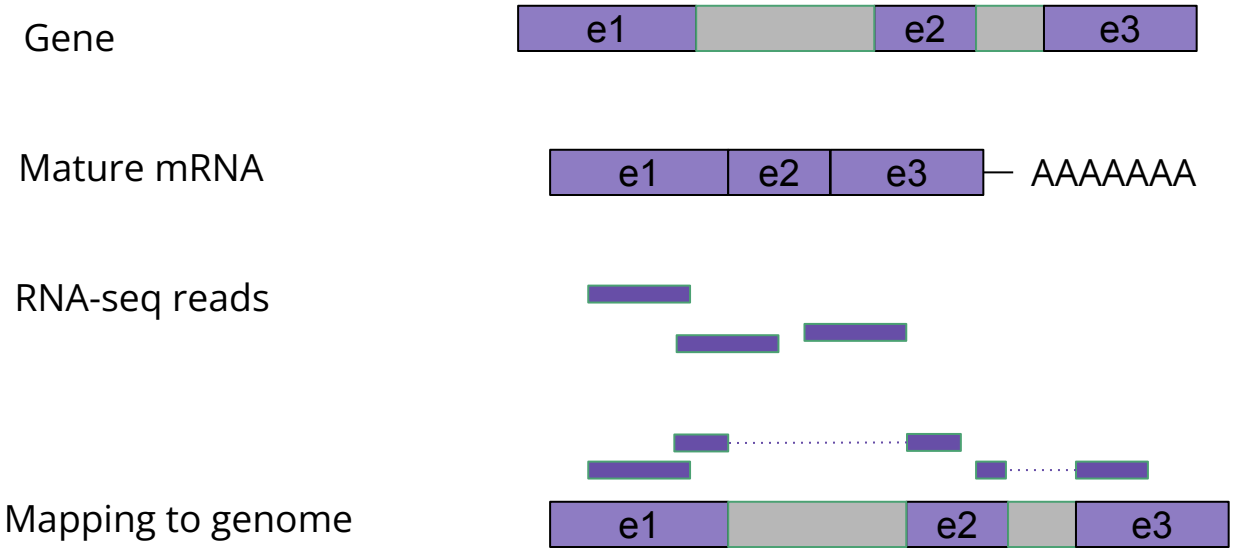
Whose read is it anyway?

- We need to assign RNA-seq reads to specific genes
- The simplest way is by read mapping
- We must have a reference genome and annotation



Splice-aware mapping

- We need to consider intron-exon gene structure
- Allow large gaps in read alignment

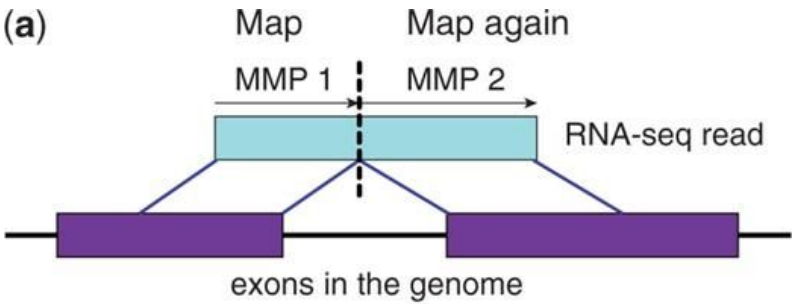


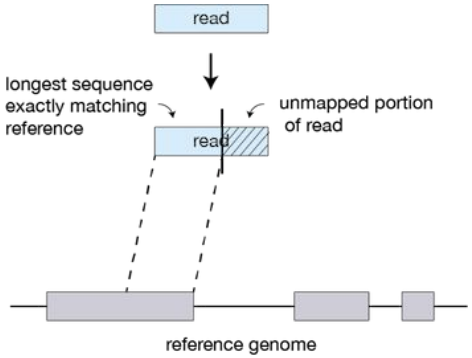
STAR - Mapping RNA-seq reads to a genome

- BWA is not designed for the task
- STAR - a very popular choice
- Very fast and memory-efficient

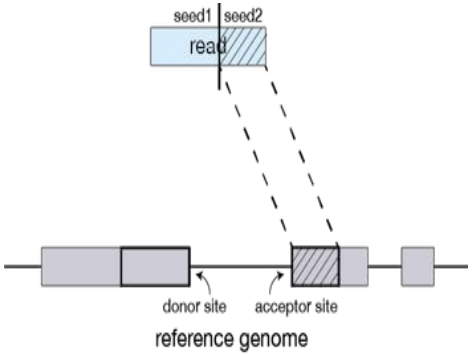
The algorithm works in two steps:

1. Find splice junctions by allowing partial read mapping
2. Stitch together parts of reads mapped to proximal genomic positions

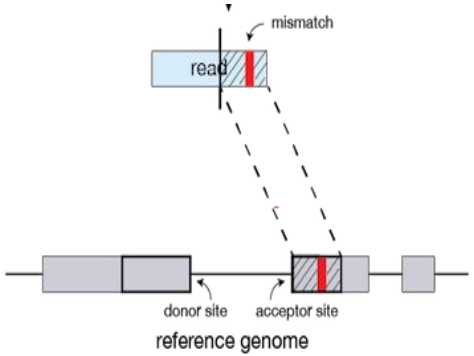




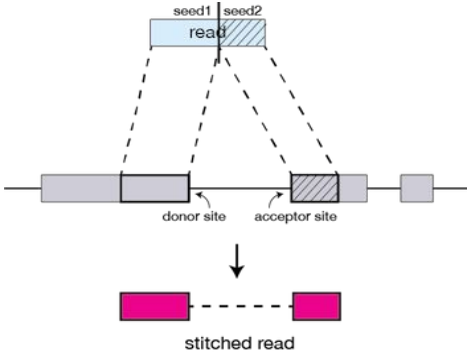
Find best exact mapping



Find best exact mapping
for unmapped part



Try to extend mapping
over mismatches



Cluster and stitch mappings
based on proximity

The GFF format

- **G**eneral **F**eature **F**ormat
- A **hierarchical** format for describing genomic features
- A TSV text file
- **Watch out for different GFF versions - it's a mess!**
  - GFF
  - GFF2 ≃ GTF
  - GFF3

The GFF3 format

Specifications

- Always starts with the line:
  `##gff-version 3`
- 9 mandatory columns
- Coordinates start from 1
- End coordinate is included

	Name	Description
1	seqid	Chromosome/scaffold name
2	source	Program or DB that created the annotation
3	type	Feature type (gene, exon, CDS...)
4	start	Start position on chromosome/scaffold
5	end	End position on chromosome/scaffold
6	score	Some score we assign to the feature (or .)
7	strand	+/- (or .)
8	phase	0,1 or 2 - indicating the coding frame (or .)
9	attributes	Additional information about the feature

GFF3 - example

```
##gff-version 3
ChrI    SGD     chromosome      1       230218  .       .       .       ID=chrI;dbxref=NCBI:NC_001133;Name=chrI
ChrI    SGD     telomere        1       801     .       -       .       ID=TEL01L;Name=TEL01L;Note=Telomeric%20region%20on%20the%2
ChrI    SGD     X_element       337     801     .       -       .       ID=TEL01L_X_element;Name=TEL01L_X_element;dbxref=SGD:S0000
ChrI    SGD     X_element_combinatorial_repeat  63      336     .       -       .       ID=TEL01L_X_element_combinatorial_repeat;N
ChrI    SGD     telomeric_repeat        1       62      .       -       .       ID=TEL01L_telomeric_repeat;Name=TEL01L_telomeric_r
ChrI    SGD     gene    335     649     .       +       .       ID=YAL069W;Name=YAL069W;Ontology_term=GO:0003674,GO:0005575,GO:000
ChrI    SGD     mRNA    335     649     .       +       .       ID=YAL069W_mRNA;Name=YAL069W_mRNA;Parent=YAL069W
ChrI    SGD     exon    335     649     .       +       0       Parent=YAL069W_mRNA;Name=YAL069W_exon;orf_classification=Dubious
ChrI    SGD     CDS     335     649     .       +       0       Parent=YAL069W_mRNA;Name=YAL069W_CDS;orf_classification=Dubious
ChrI    SGD     ARS     707     776     .       .       .       ID=ARS102;Name=ARS102;Alias=ARSI-1;Note=Autonomously%20Replicating
ChrI    SGD     gene    87286   87752   .       +       .       ID=YAL030W;Name=YAL030W;gene=SNC1;Alias=SNC1,SNAP%20receptor%20SNC
ChrI    SGD     mRNA    87286   87752   .       +       .       ID=YAL030W_mRNA;Name=YAL030W_mRNA;Parent=YAL030W
ChrI    SGD     exon    87286   87387   .       +       0       Parent=YAL030W_mRNA;Name=YAL030W_exon;orf_classification=Verified
ChrI    SGD     CDS     87286   87387   .       +       0       030W_CDS;orf_classification=Verified
ChrI    SGD     exon    87501   87752   .       +       exon;orf_classification=Verified
ChrI    SGD     CDS     87501   87752   .       +       DS;orf_classification=Verified
```

What Linux command should we use to extract features of type “gene” located on chromosome II?

Viewing GFF files in IGV

Running STAR

- Inputs:
  - Reference genome - fasta
  - RNA-seq reads - fastq (SE or PE)
  - Optional: genome annotation - GFF/GTF

- Output: reads alignment in sam/bam format

Running STAR

Step 1: create a reference index:

```
STAR --runMode genomeGenerate --genomeFastaFiles
</path/to/genome.fasta> --sjdbGTFfile </path/to/genes.gff>
--genomeDir </path/to/ref_index_dir/>
```

Step 2: align reads to indexed reference:

```
STAR --runMode alignReads --genomeDir
</path/to/ref_index_dir/> --readFilesIn reads_R1.fastq
reads_R2.fastq --outSAMtype BAM SortedByCoordinate
--outFileNamePrefix <name>
```

Example of STAR output


```
SRR1177156.52530710     0       ChrXI   109569  255     9M306N30M12S    *       0       0       AACGCTGAAGCTAAAGGTTTGGATGC
TACTAAATTGTACTGTAGGCACCAT       CCCFFFFFHHHHHIIIIGHIIIIIIIIIIIIIIICFIIIIDDHIHIIIII     NH:i:1  HI:i:1  AS:i:37 nM:i:0
```

QA of RNA-seq mapping with Qualimap

- Takes BAM file from STAR and produces various stats and plots
- Command:

```
qualimap rnaseq -bam </path/to/STAR.bam> -gtf
</path/to/genes.gtf>
```

Summary stats

Summary

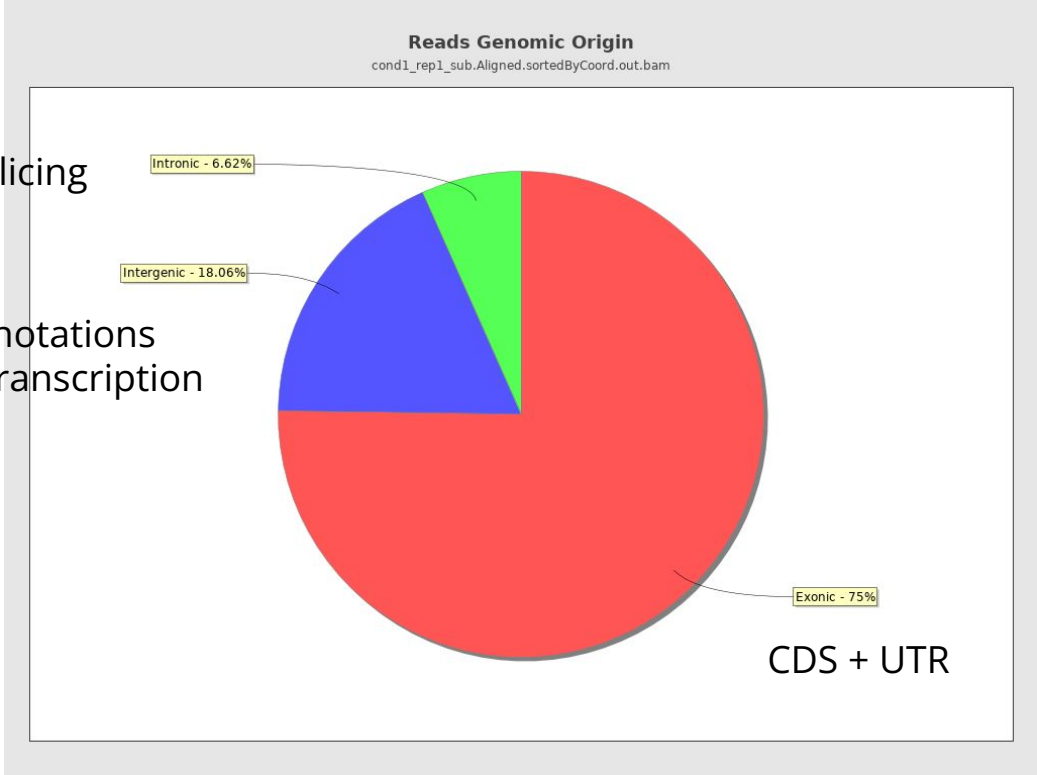
Reads alignment

Number of mapped reads:	3,909,889
Total number of alignments:	3,968,309
Number of secondary alignments:	58,420
Number of non-unique alignments:	102,393
Aligned to genes:	2,910,346
Ambiguous alignments:	747
No feature assigned:	953,759
Missing chromosome in annotation:	1,064
Not aligned:	0
Strand specificity estimation (fwd/rev):	0.01 / 0.99

Reads genomic origin

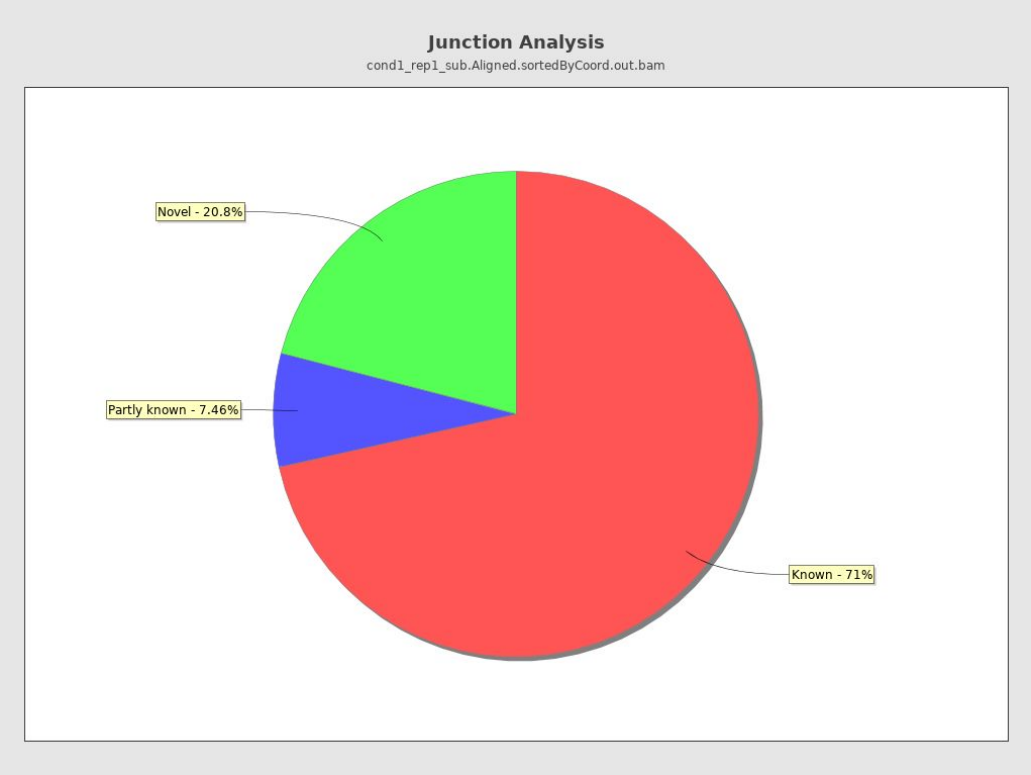
Pre-mRNA
Alternative splicing

Missing annotations
Pervasive transcription



CDS + UTR

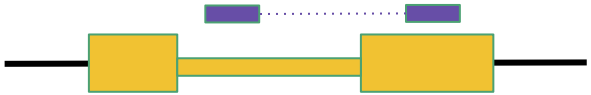
Splice junction analysis



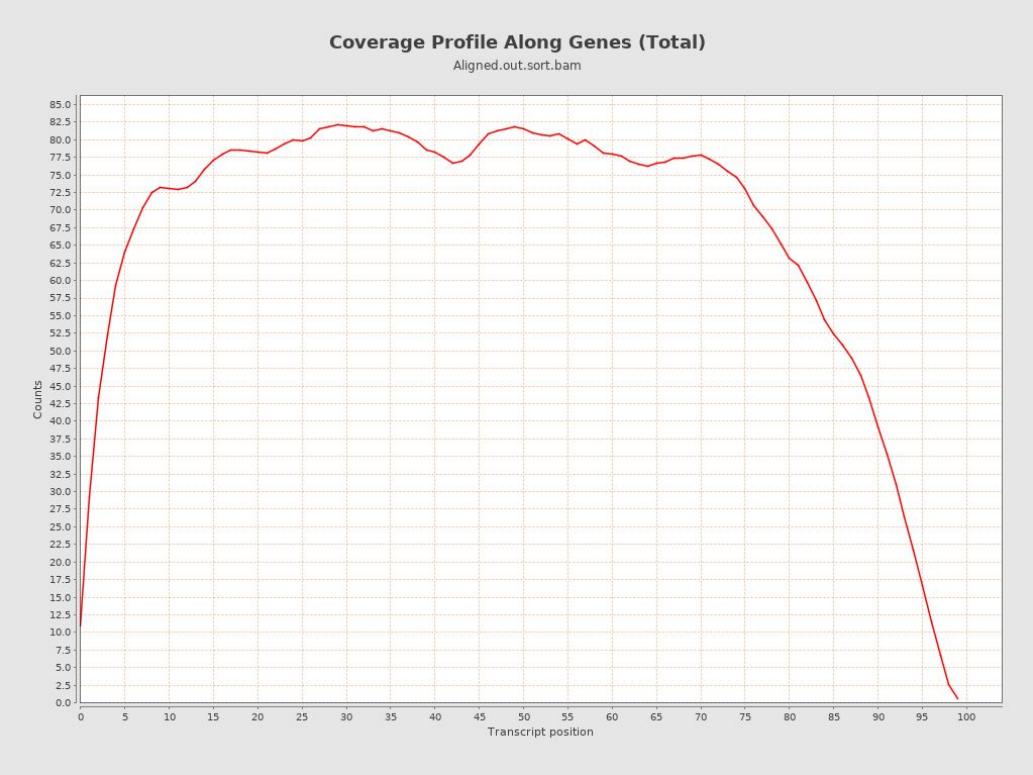
Known

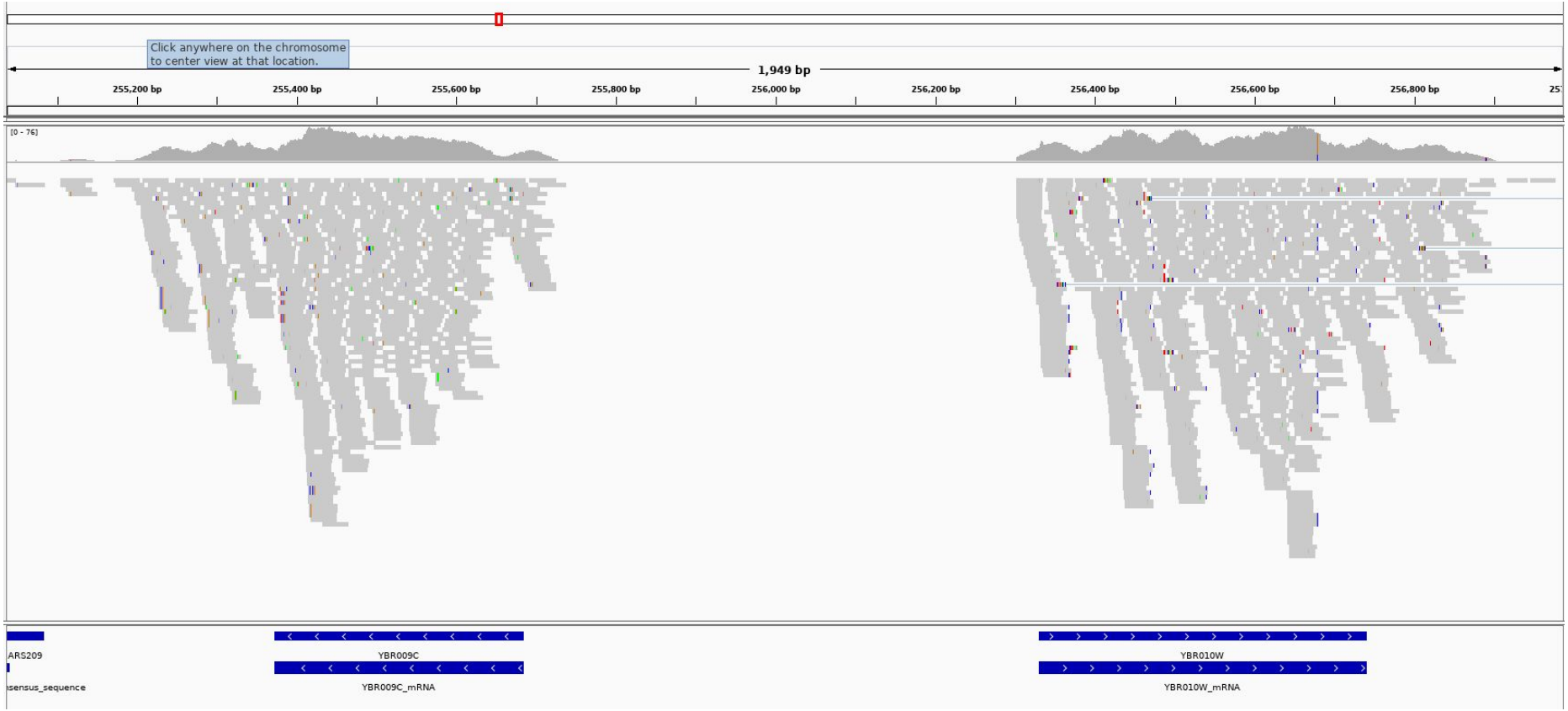
Partly Known

Novel

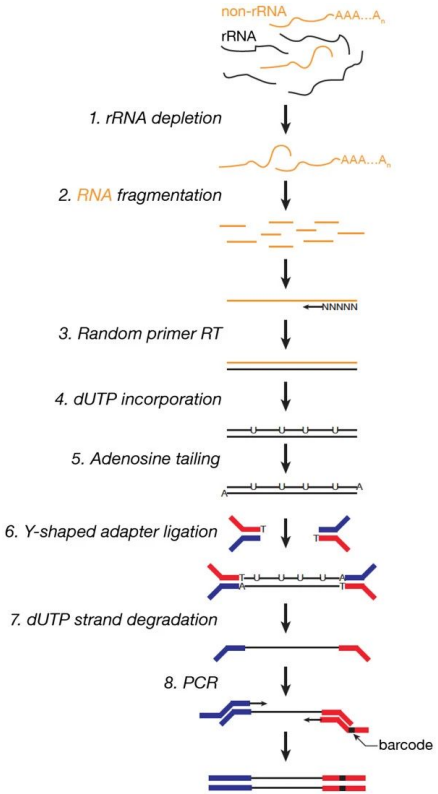


Coverage profile along genes



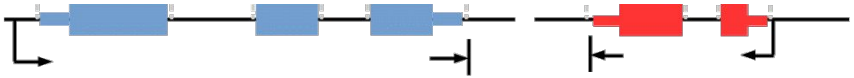

Click anywhere on the chromosome
to center view at that location.
1,949 bp
255,200 bp
255,400 bp
255,600 bp
255,800 bp
256,000 bp
256,200 bp
256,400 bp
256,600 bp
256,800 bp
[0 - 76]
ARS209
YBR009C
YBR010W
sensus_sequence
YBR009C_mRNA
YBR010W_mRNA

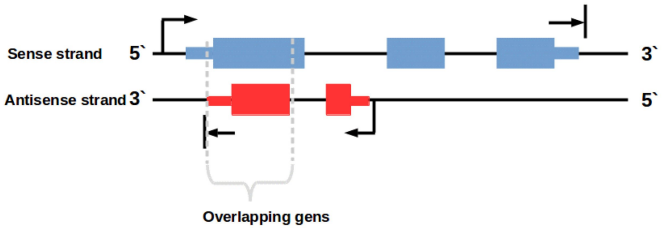
Strand-specific (stranded) protocols

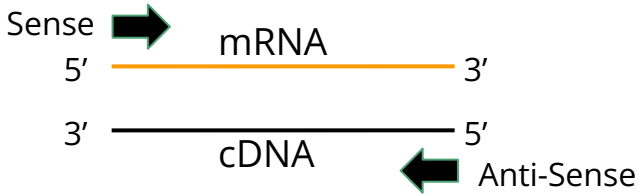


Zhang, Zhao, et al. "Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly (A) selection." *Silence* 3.1 (2012): 9.

Why is strand specificity important?







https://www.ecseq.com/support/ngs/how-do-strand-specific-sequencing-protocols-work

K-mer abundance analysis

- An alternative to RNA-seq reads mapping for expression quantification
- Much faster!
- Using **quasi-alignment** - no spliced alignments required
- Results are close to optimal
- Not an option when the actual mappings are required
- Several available tools:
  - Kallisto
  - Salmon
  - iMOKA

How does it work?

Step 1: generate the k-mer table

Transcript 1:

TTCTCTGTTACCATCTAAAATTCTTGAACTCT
TTTTCCCATTATTTTTCCACCACCTTTGC...

Transcript 2:

CTCCCATCACCCGACCTCAACGCCCGCCTCAT
CCTCCTCTACCGCCTCTTCGCCGCCCG...

Transcript 3:

TCTTGTTGGAAGAGATGCAGTGTAAGGGGGTT
GATTTGAATGTTGTAATATTTAACACG...

TTCTCTGTT : Transcript 1
TCTCTGTTA : Transcript 1
CTCTGTTAC : Transcript 1

CACCTTTGC : Transcript 1

CTCCCATCA : Transcript 2

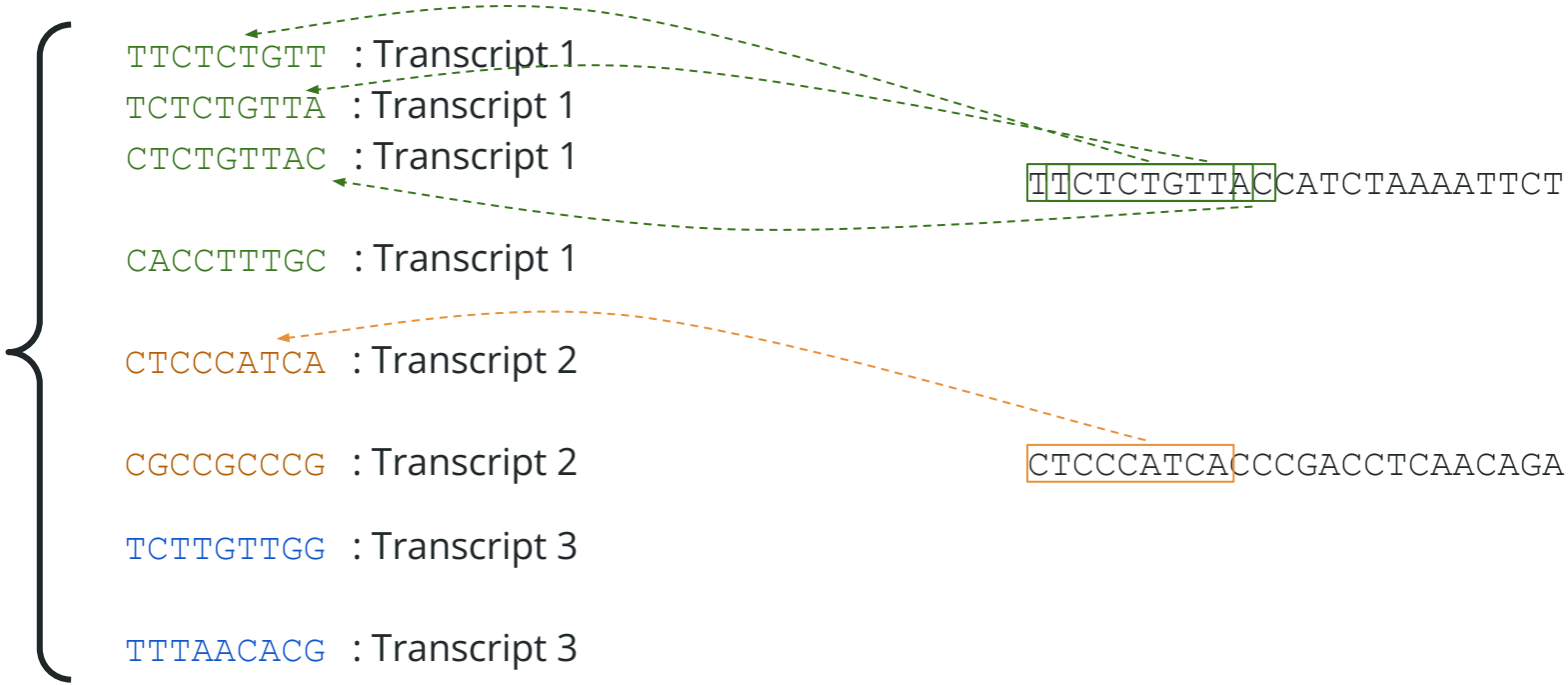
CGCCGCCCG : Transcript 2

TCTTGTTGG : Transcript 3

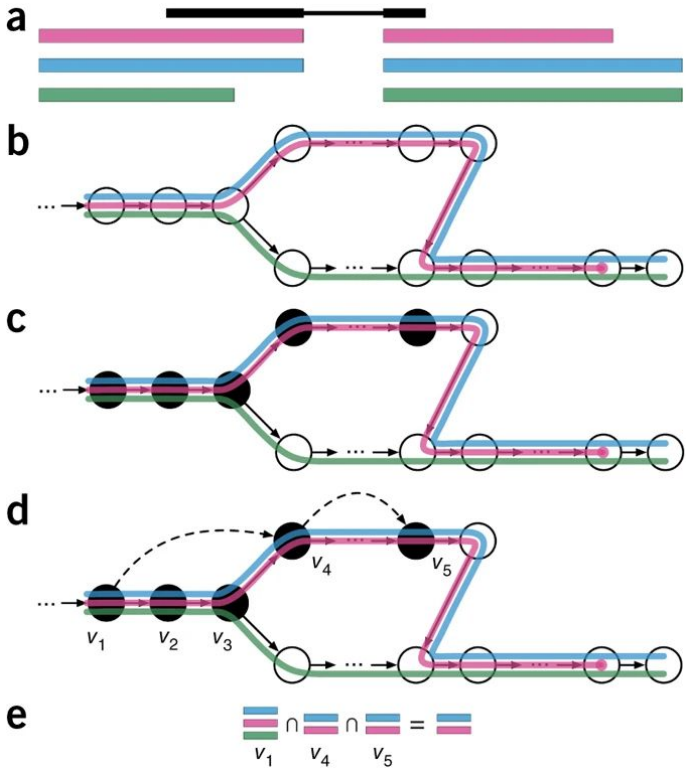
TTTAACACG : Transcript 3

How does it work?

Step 2: quasi-alignment of reads



How does it work in practice? (Kallisto)



Bray, Nicolas L., et al. "Near-optimal probabilistic RNA-seq quantification." *Nature biotechnology* 34.5 (2016): 525-527.

Detecting expression variants between yeast strains

- So far, we explored genomic variation between the reference strain S288C and the wine strain RM11
- Now we want to find which genes are differentially expressed between the strains