Introduction to Transcriptomics

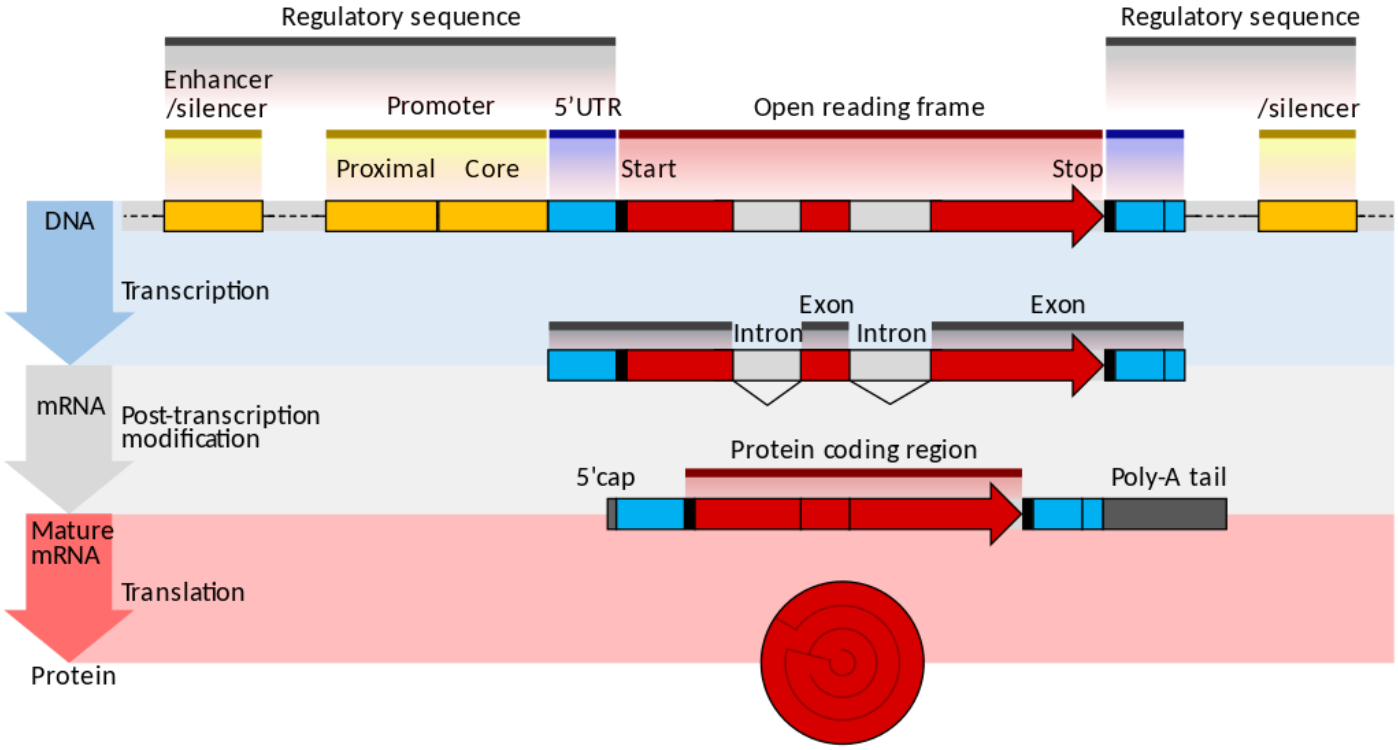
Requirements

Before diving into this slide deck, we recommend you to have a look at:

- Galaxy introduction
- Quality control

What is RNA sequencing?

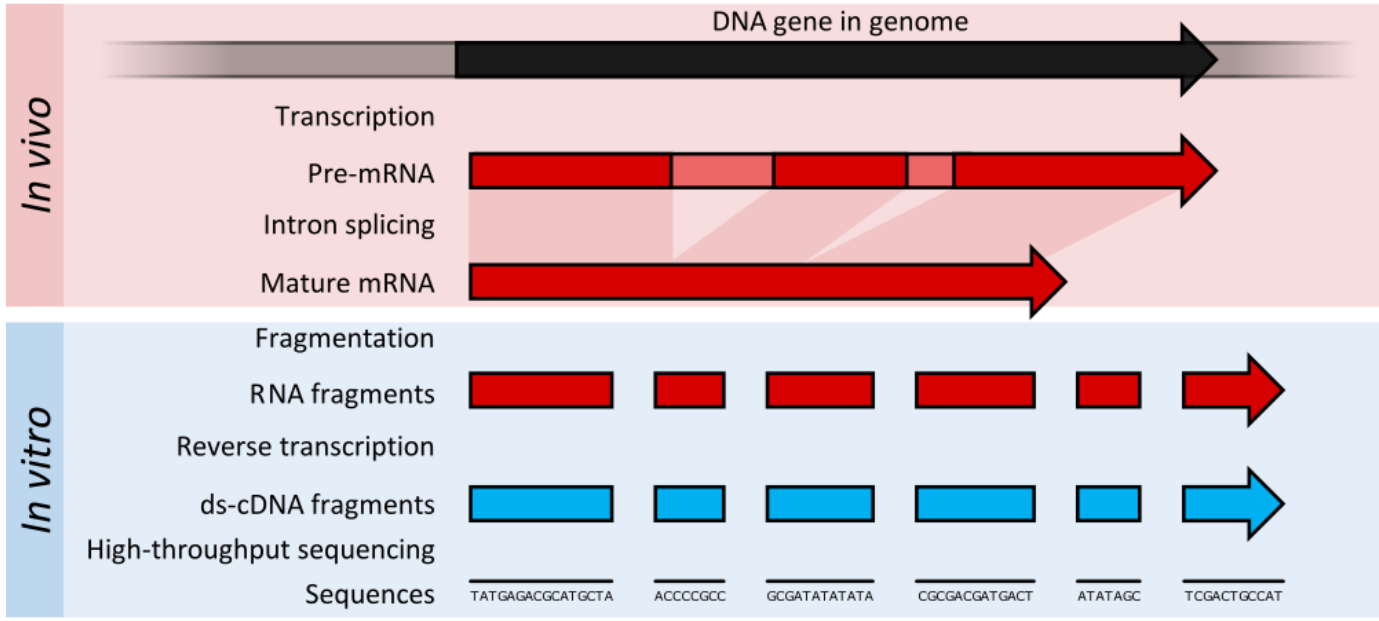
RNA



- Transcribed form of the DNA
- Active state of the DNA

Credit: Thomas Shafee

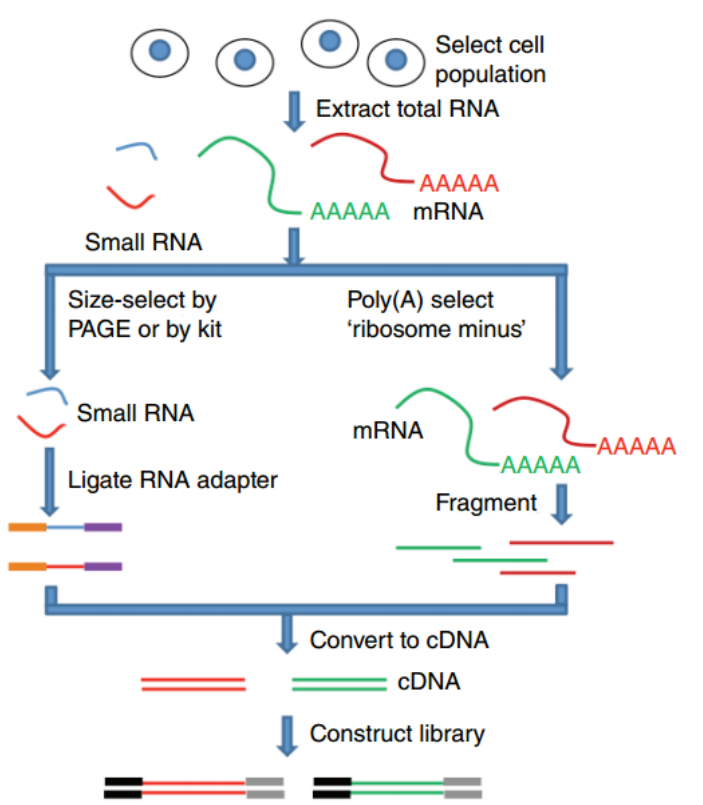
RNA sequencing



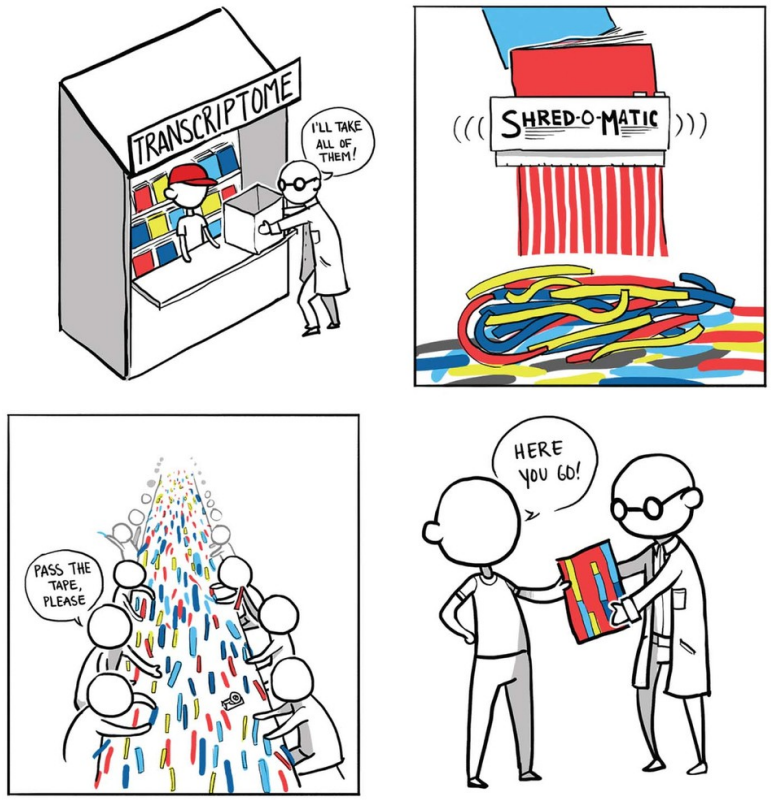
- RNA quantification at single base resolution
- Cost efficient analysis of the whole transcriptome in a high-throughput manner

Credit: Thomas Shafee (adapted)

Where does my data come from?



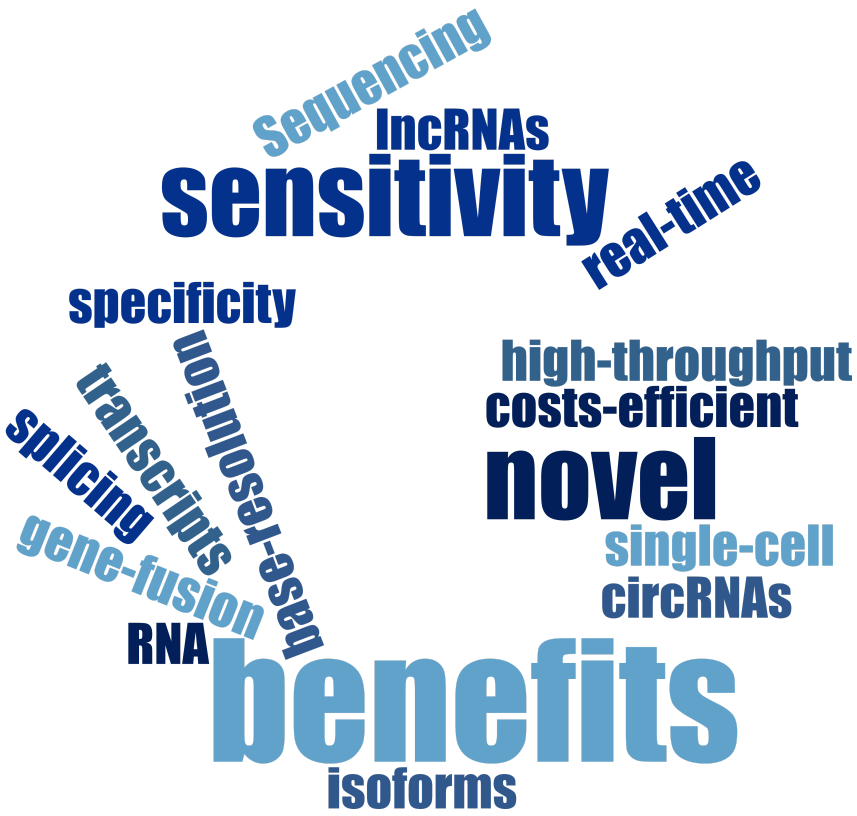
Principle of RNA sequencing



Challenges of RNA sequencing

- Different origin for the sample RNA and the reference genome
- Presence of incompletely processed RNAs or transcriptional background noise
- Sequencing biases (PCR library preparation)

Benefits of RNA sequencing



2 main research applications for RNA-Seq

- Transcript discovery

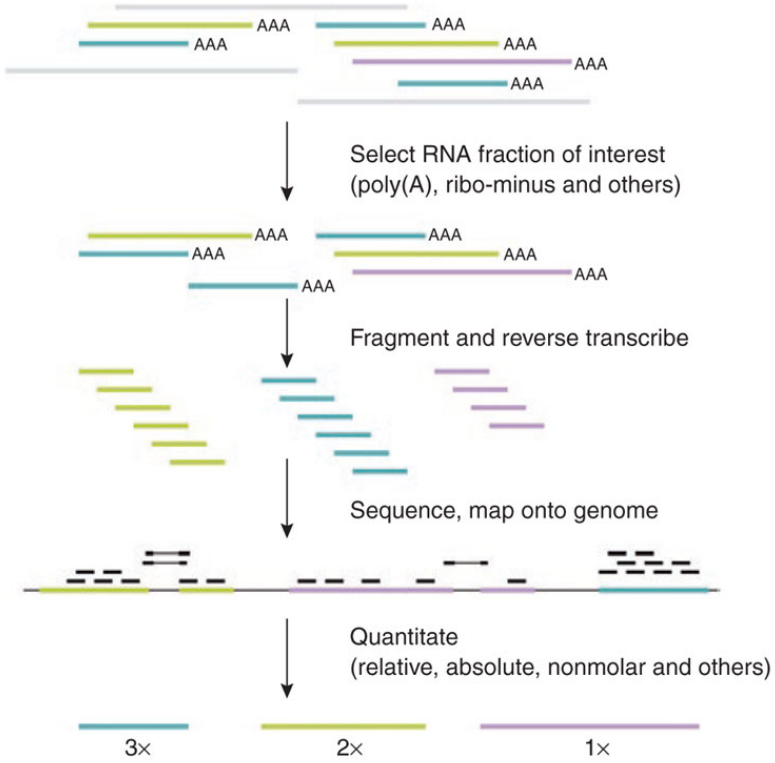
  Novel isoforms and alternative splicing, Non-coding RNAs, Single nucleotide variations, Fusion genes

- RNA quantification

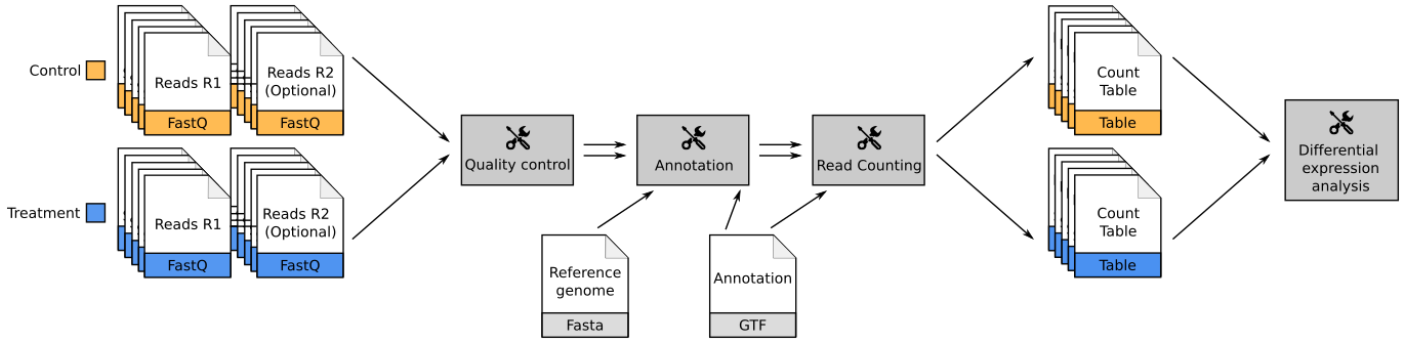
  Absolute gene expression (within sample), Differential expression (between biological samples)

How to analyze RNA seq data for RNA quantification?

RNA quantification



Overview of the Data Processing



- No available standardized workflow
- Multiple possible best practices for every dataset

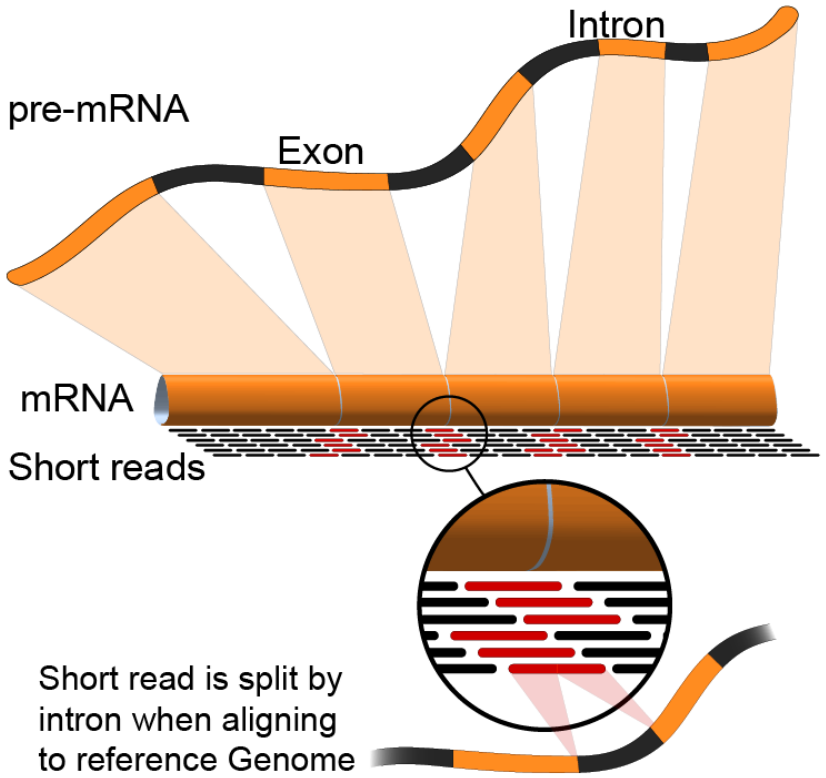
Data Pre-processing

1. Adapter clipping to trim the sequencing adapters
2. Quality trimming to remove wrongly called and low quality bases

See NGS Quality control

Annotation of RNA-Seq reads

Simple mapping on a reference genome? More challenging



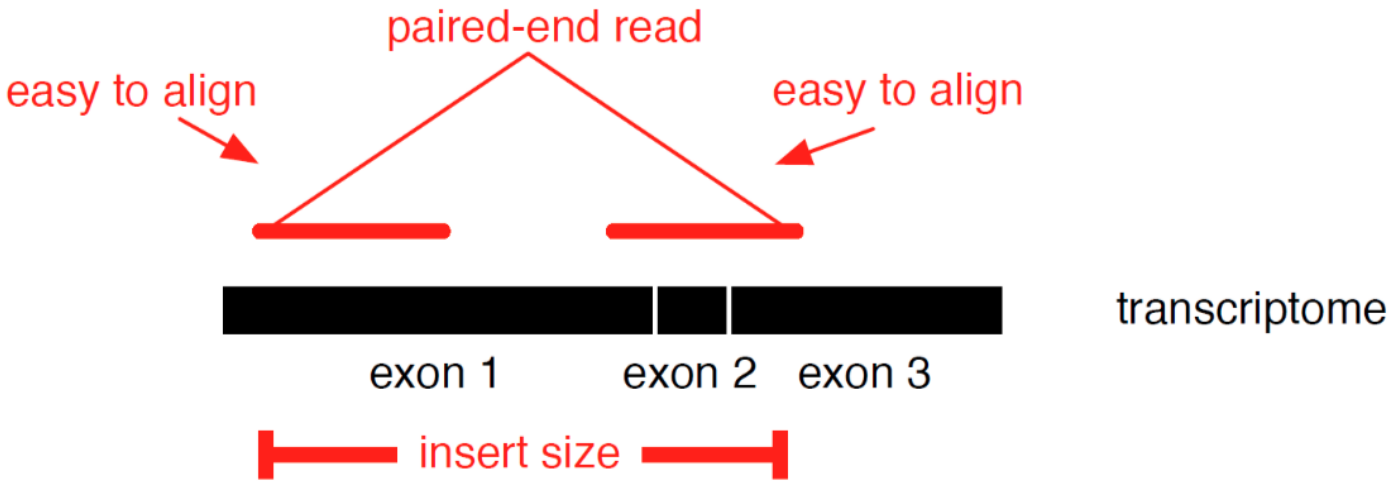
Credit: Rgocs

Annotation of RNA-Seq reads

3 main strategies for annotations

- Transcriptome mapping
- Genome mapping
- transcriptome assembly and annotation

Transcriptome mapping

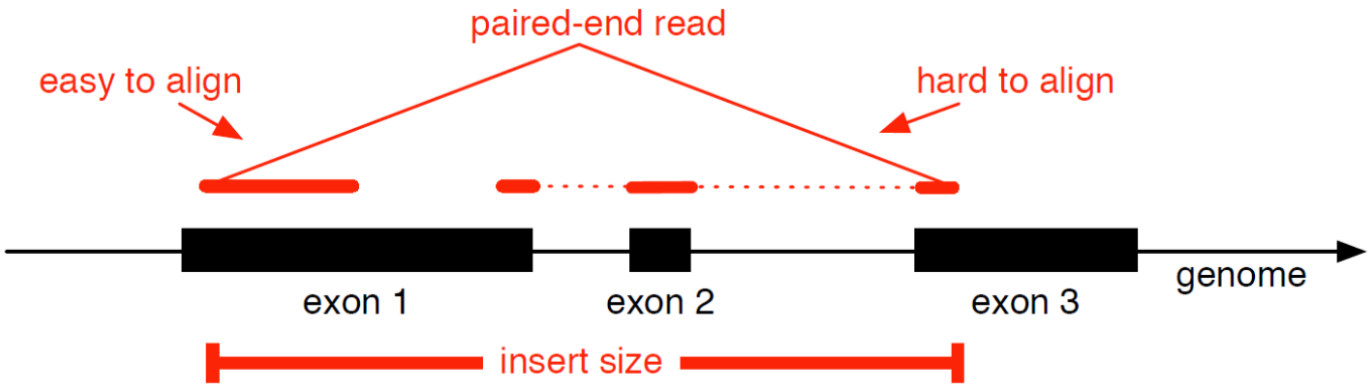


- Need reliable gene models
- No detection of novel genes

Figures by Ernest Turro, EMBO Practical Course on Analysis of HTS Data, 2012

Genome mapping

Splice-aware read alignment



Detection of novel genes and isoforms

Figures by Ernest Turro, EMBO Practical Course on Analysis of HTS Data, 2012

Transcriptome and Genome mapping

Needed

- Reference genome/transcriptome in FASTA
- Annotations of known genes, … in GTF

Where to find?

- Joint projects to produce and maintain annotations on selected organisms: EMBL-EBI, UCSC, RefSeq, Ensembl, …

De novo transcriptome assembly

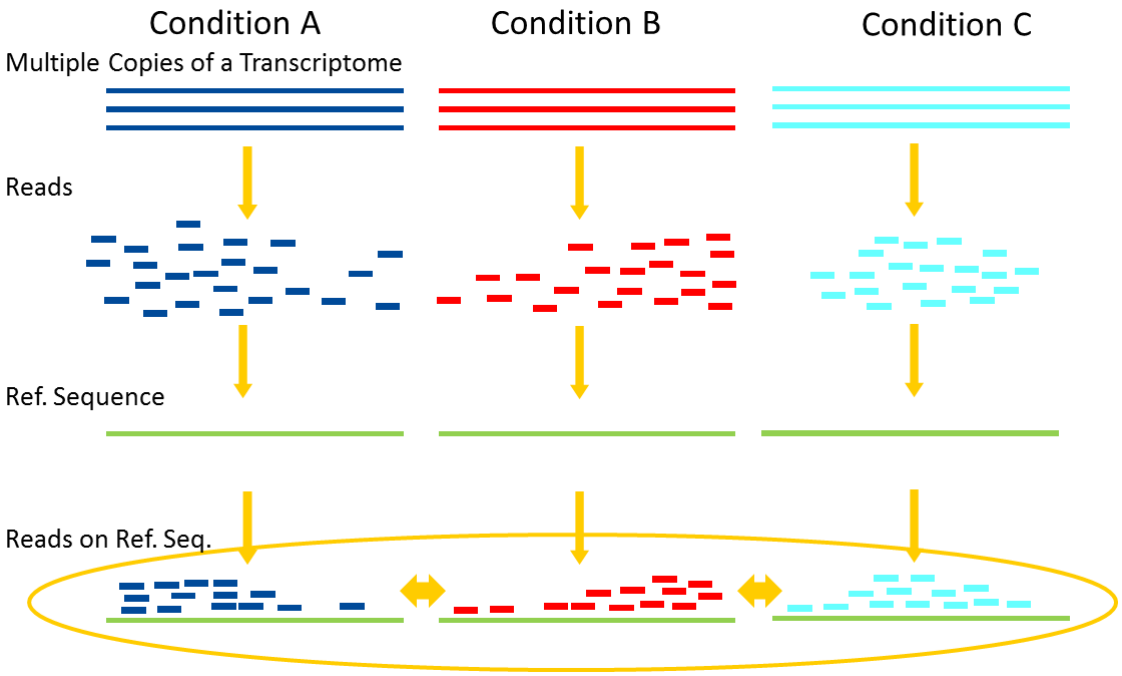
No need for a reference genome ...

1. Assembly into transcripts
2. Map reads back

Quantification

- Counting the number of reads per features: Easy!!
- Challenges
  - How to handle multi-mapped reads (reads with multiple alignments)?
  - How to distinguish between different isoforms?
    - At gene level?
    - At transcript level?
    - At exon level?

Differential Expression Analysis



Account for variability of expression across biological replicates with the help of counts

Differential Expression Analysis: Normalization

- By Features: genes, isoforms
- By Samples

- Methods
  - (Cufflinks/Cuffdiff)
  - (edgeR)
  - (DESeq2)

Normalize counts for gene in library by size factor

- Dillies et al., Brief Bioinf, 2013

Impact of sequencing depth and number of replicates

Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Example of calculations for the probability of detecting differential expression in a single test at a significance level of 5 %, for a two-group comparison using a Negative Binomial model, as computed by the RNASeqPower package

Visualization

- Integrative Genomics Viewer ( ) or Trackster

  Visualization of the aligned BAM files

-

  Quantitative visualization of read coverage along exons and splice junctions

-

  Visualization package for Cufflinks high-throughput sequencing data

Related tutorials

Thank you!

This material is the result of a collaborative work. Thanks the Galaxy Training Network and all the contributors (Bérénice Batut, Anika Erxleben, Markus Wolfien) !

Found a typo? Something is wrong in this tutorial?
Edit it on GitHub