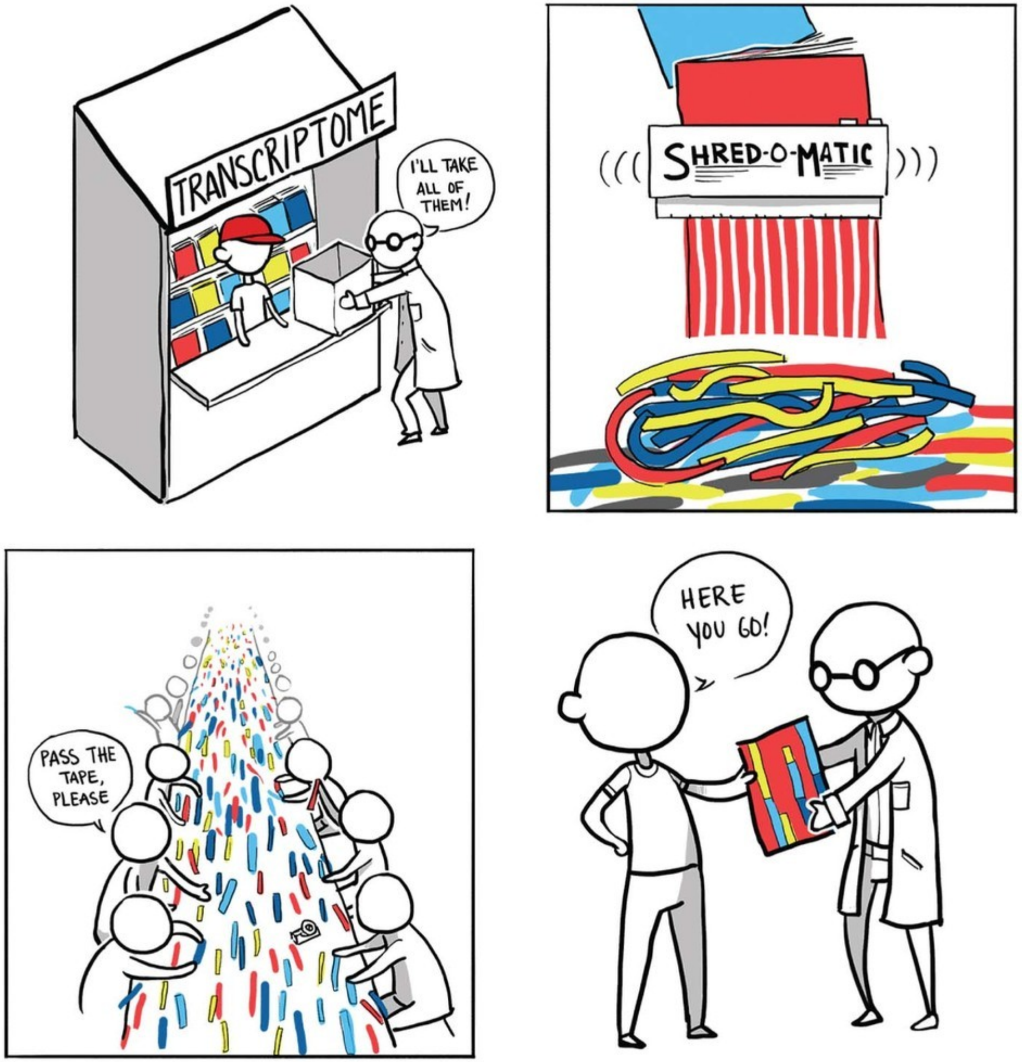
RNA-seq analysis in Galaxy

Fabian Kilpert[1] & Pavan Videm[2]

[1]MPI-IE Freiburg
[2]University of Freiburg

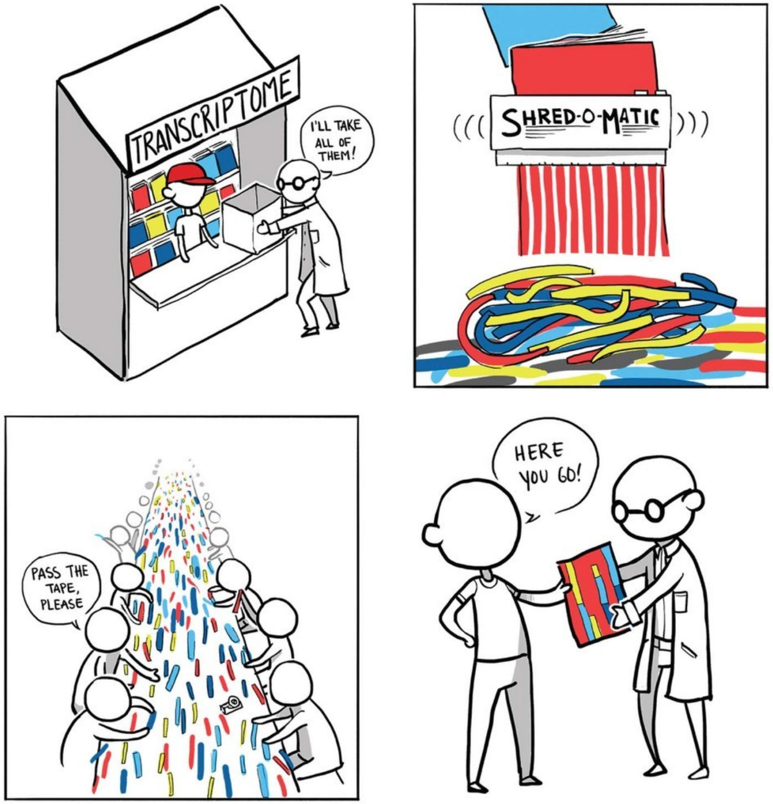
The RNA-seq approach



cheap & fast

Transcriptome reconstruction—akin to reassembling magazine articles after they have been through a paper shredder. [Korf, Nat Meth, 2013]

...and the quality?

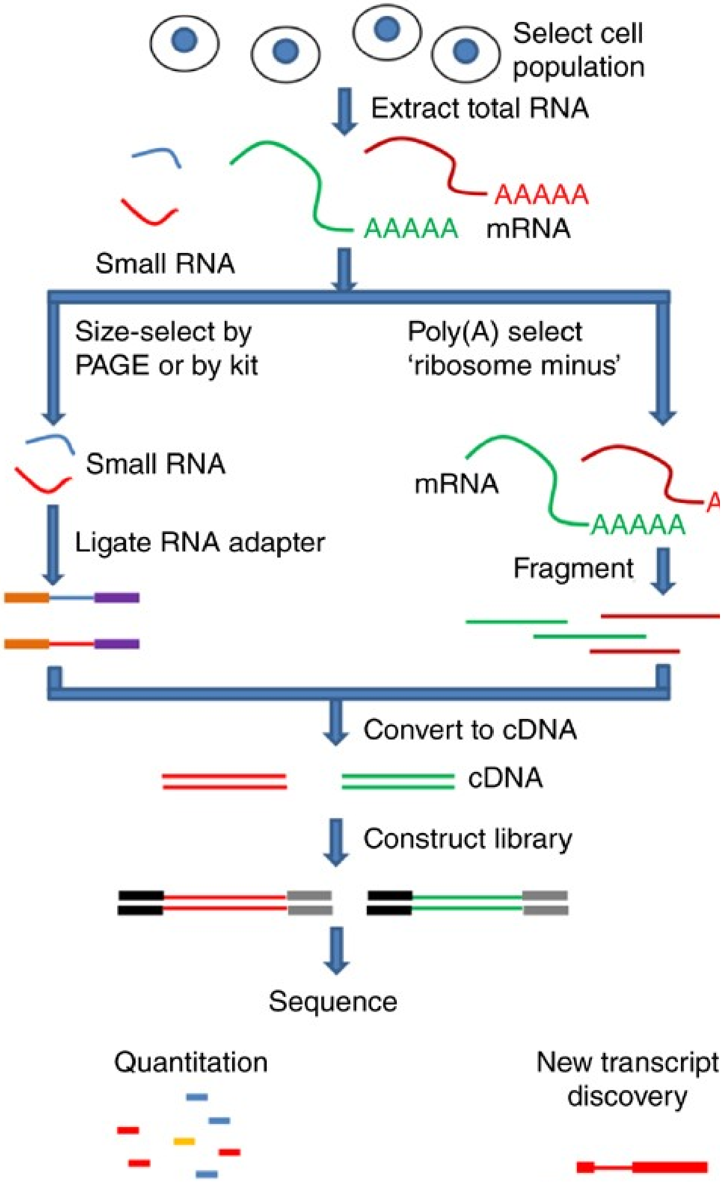


[Korf, Nat Meth, 2013]

Challenges:

- Sample RNA is from a different source than the reference genome
- Incompletely processed RNAs or transcriptional noise
- Biases in sequencing (e.g. PCR in library preparation)

RNA-seq library construction



[Zeng & Mortazavi, Nat Immun, 2012]

RNA-seq applications

Allows for:
high-throughput
whole genome
gene expression analysis

Two main research aims:

I. **Transcript discovery**

Which RNA molecules are in my sample?

- novel isoforms and alternative splicing
- non-coding RNAs
- single nucleotide variations
- fusion genes

II. **RNA quantification**

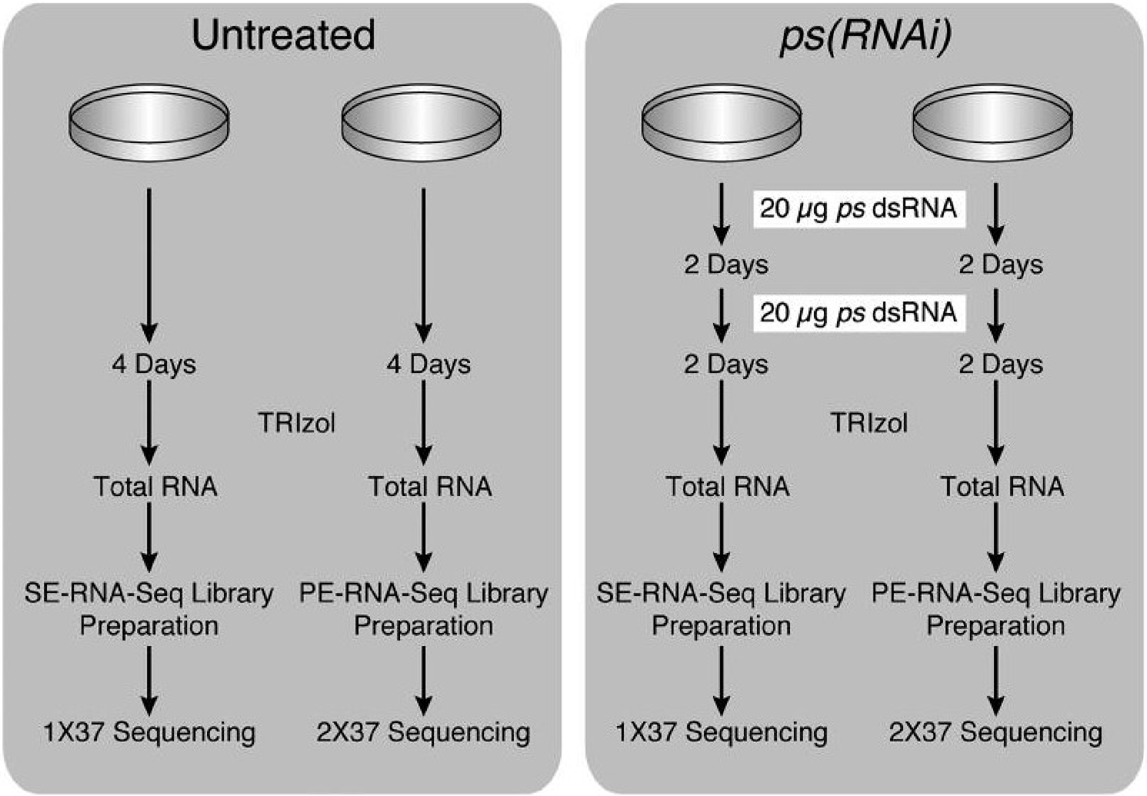
What is the concentration of RNAs?

- absolute gene expression (within sample)
- differential gene expression (between biological samples)
- isoform expression / differential exon usage / alternative splicing

Hands-on example dataset

RNA-seq data from *Drosophila melanogaster*

wt vs ps depleted (treated)



[Brooks et al., Genome Res, 2011]

Step 1: Preprocessing of raw sequencing reads

- Sequencer → FASTQ files
- inspect quality of raw sequencing reads with FastQC
- optional (depending on QC results):
  - trim low quality bases from 3’ end of the reads
    (use e.g. `cutadapt`, `Trim Galore!`)
  - clip adapter sequences
    (use e.g. `cutadapt`, `Trim Galore!`)
  - trim poly(A) tails
    (use e.g. `PRINSEQ`)

Galaxy exercise

http://galaxy.uni-freiburg.de/u/tutor/p/rna-seq

Step 1: Inspecting the FASTQ files

Hands-on!

Step 2: Annotating RNA-seq reads

How do I identify my reads?

If there is reference data (model organism):

Reference-based mapping
of RNA-seq reads to a genome and/or transcriptome using a sequence aligner
(Bowtie, BWA, SHRiMP, Stampy, etc.)

Reference-based mapping (splice-aware)
of RNA-seq reads to a genome using a splice-aware sequence aligner
(TopHat, MapSplice, SpliceMap, etc.)

If there is no reference data (non-model organism):

***De novo* assembly (+ additional annotation step)**
Does not use a reference genome
(Trinity, Trans-AbySS, Velvet-Oases, etc.)

Combined (*reference-based + de novo*)

Sources of reference annotations

Where do I get reliable reference annotations?

There are joint projects to produce and maintain annotations on selected organisms:

e!Ensembl EMBL-EBI wellcome trust sanger institute

UCSC genes UNIVERSITY OF CALIFORNIA SANTA CRUZ

RefSeq: NCBI Reference Sequence Database NCBI rather conservative

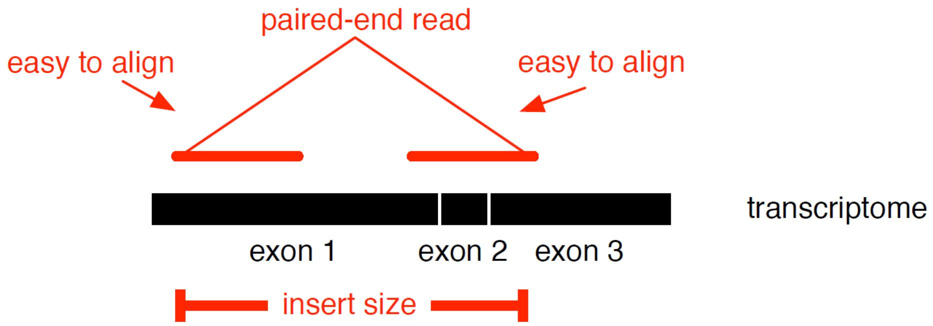
Annotations of known genes, repeats, etc. are provided in **GTF** (Gene transfer format) file format. Tab separated text file, e.g.:

```
1→transcribed_unprocessed_pseudogene→gene→11869→14409→.→+→.→gene_id "ENSG00000223972";
X→Ensembl→Repeat→2419108→2419128→42→.→.→hid=trf; hstart=1; hend=21;
```

You can use Galaxy to retrieve GTF files!

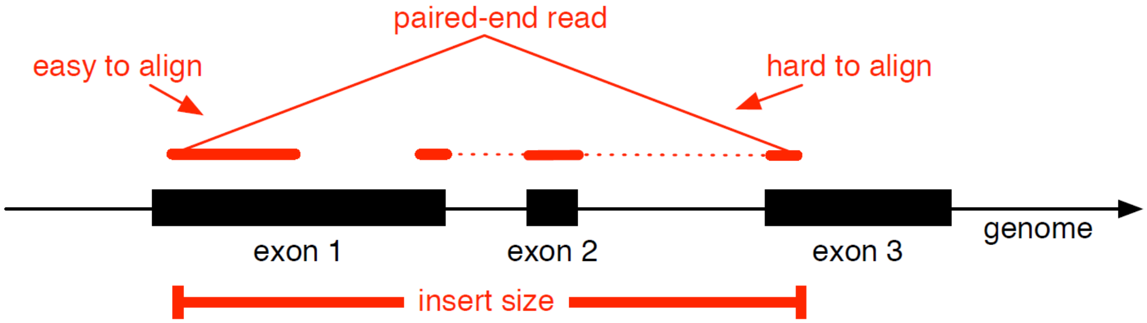
RNA-seq alignment strategies

Transcriptome alignment



- reliable gene models required
- no detection of novel genes

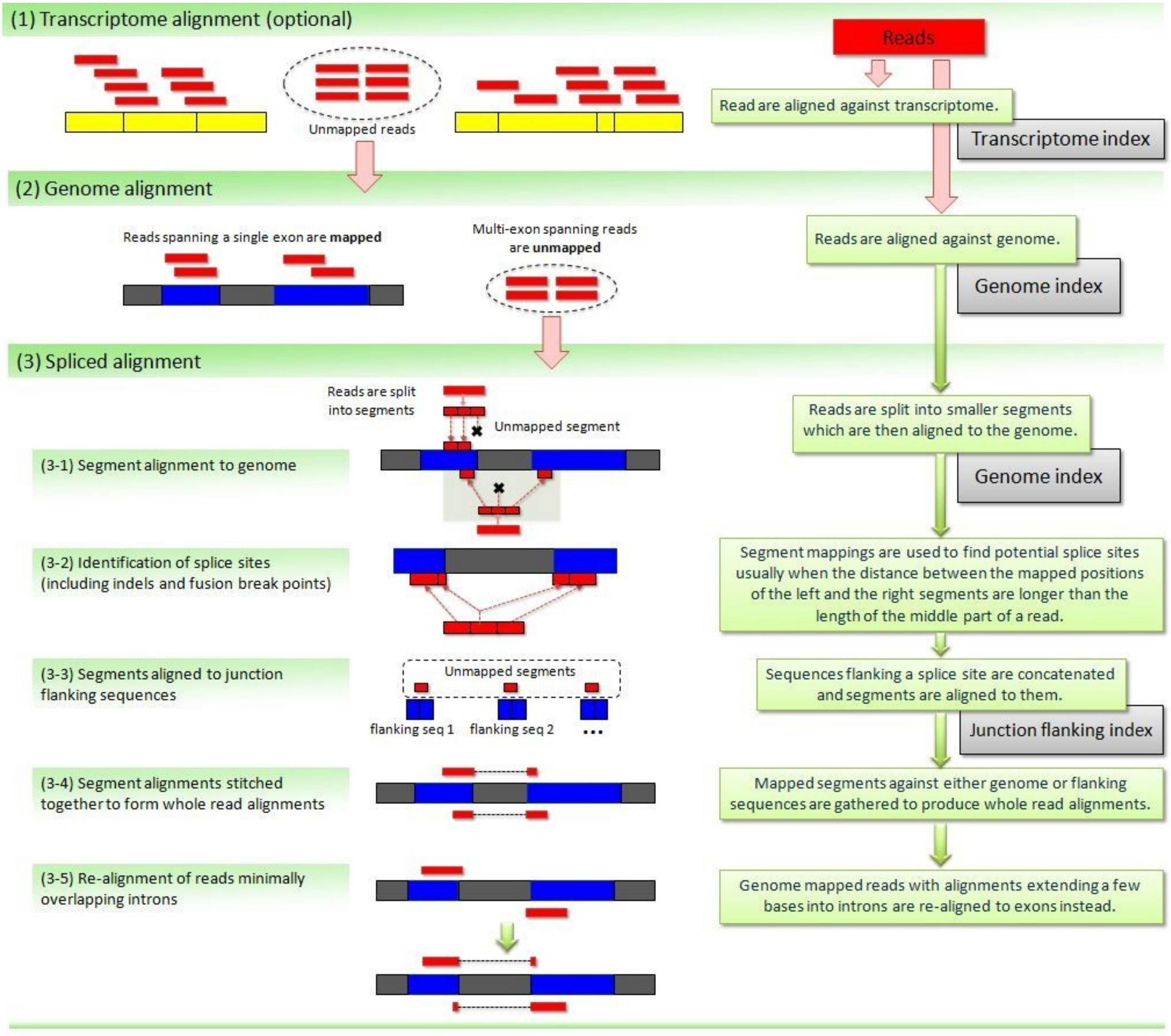
Genome alignment (splice-aware read alignment)



+ detection of novel genes and isoforms

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

TopHat2 – A popular splice-aware aligner



(1) Transcriptome alignment (optional)

Unmapped reads

(2) Genome alignment

Reads spanning a single exon are **mapped**

Multi-exon spanning reads are **unmapped**

(3) Spliced alignment

Reads are split into segments

Unmapped segment

(3-1) Segment alignment to genome

(3-2) Identification of splice sites (including indels and fusion break points)

(3-3) Segments aligned to junction flanking sequences

Unmapped segments

flanking seq 1 flanking seq 2 …

(3-4) Segment alignments stitched together to form whole read alignments

(3-5) Re-alignment of reads minimally overlapping introns

Reads

Read are aligned against transcriptome.

Transcriptome index

Reads are aligned against genome.

Genome index

Reads are split into smaller segments which are then aligned to the genome.

Genome index

Segment mappings are used to find potential splice sites usually when the distance between the mapped positions of the left and the right segments are longer than the length of the middle part of a read.

Sequences flanking a splice site are concatenated and segments are aligned to them.

Junction flanking index

Mapped segments against either genome or flanking sequences are gathered to produce whole read alignments.

Genome mapped reads with alignments extending a few bases into introns are re-aligned to exons instead.

🟥 Read
🟨 Exons from annotated transcripts
🟦 Unannotated exons (novel transcripts)
⬛ Intron or intergenic region

[Kim et al., Genome Res, 2013]

Galaxy exercise

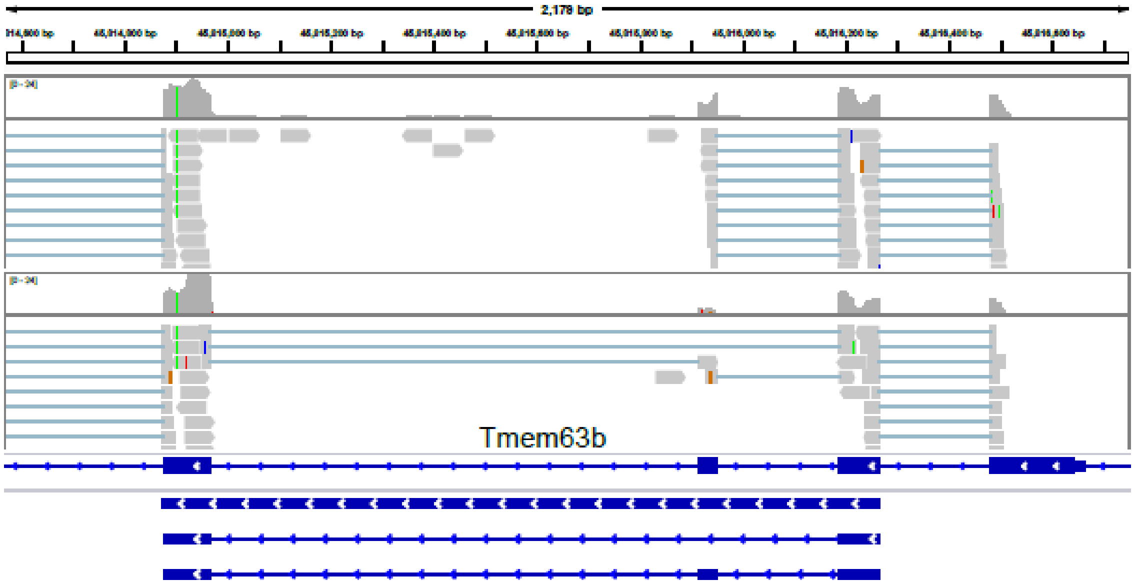
http://galaxy.uni-freiburg.de/u/tutor/p/rna-seq

Step 2: Mapping of the reads with TopHat

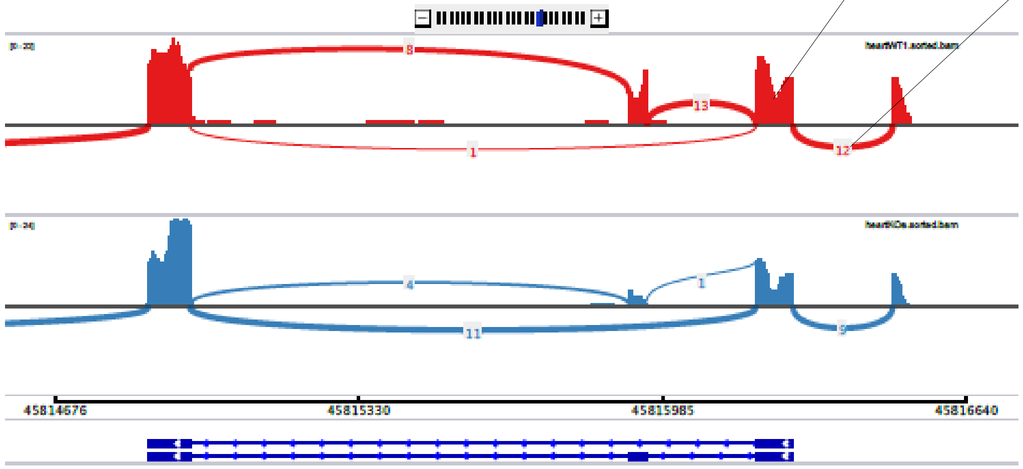
Hands-on!

Visualization of alternative splicing

Region of interest (IGV)



Interactive Sashimi plot



exon

splice junction

Sashimi plots:
Quantitative visualization of read coverage along exons and splice junctions

[Katz et al., bioRxiv, 2014]

Galaxy exercise

http://galaxy.uni-freiburg.de/u/tutor/p/rna-seq

Step 2: Mapping of the reads with TopHat

Hands-on!

Step 3: Inspecting the TopHat results

Hands-on!

Step 2 (continued): Annotating RNA-seq reads without reference

How do I identify my reads?

If there is reference data (model organism):

Reference-based mapping
of RNA-seq reads to a genome and/or transcriptome using a sequence aligner
(Bowtie, BWA, SHRiMP, Stampy, etc.)

Reference-based mapping (splice-aware)
of RNA-seq reads to a genome using a splice-aware sequence aligner
(TopHat, MapSplice, SpliceMap, etc.)

If there is no reference data (non-model organism):

***De novo* transcriptome assembly (+ additional annotation step)**
Does not use a reference genome
(Trinity, Trans-AbySS, Velvet-Oases, etc.)
Combined (*reference-based* + *de novo*)

Step 2 (continued): Annotating RNA-seq reads without reference

How do I identify my reads?

If there is reference data (model organism):

Reference-based mapping
of RNA-seq reads to a genome and/or transcriptome using a sequence aligner
(Bowtie, BWA, SHRiMP, Stampy, etc.)

Reference-based mapping (splice-aware)
of RNA-seq reads to a genome using a splice-aware sequence aligner
(TopHat, MapSplice, SpliceMap, etc.)

If there is no reference data (non-model organism):

***De novo* transcriptome assembly (+ additional annotation step)**
Does not use a reference genome
(Trinity, Trans-AbySS, Velvet-Oases, etc.)
Combined (*reference-based* + *de novo*)

NO Hands-on!

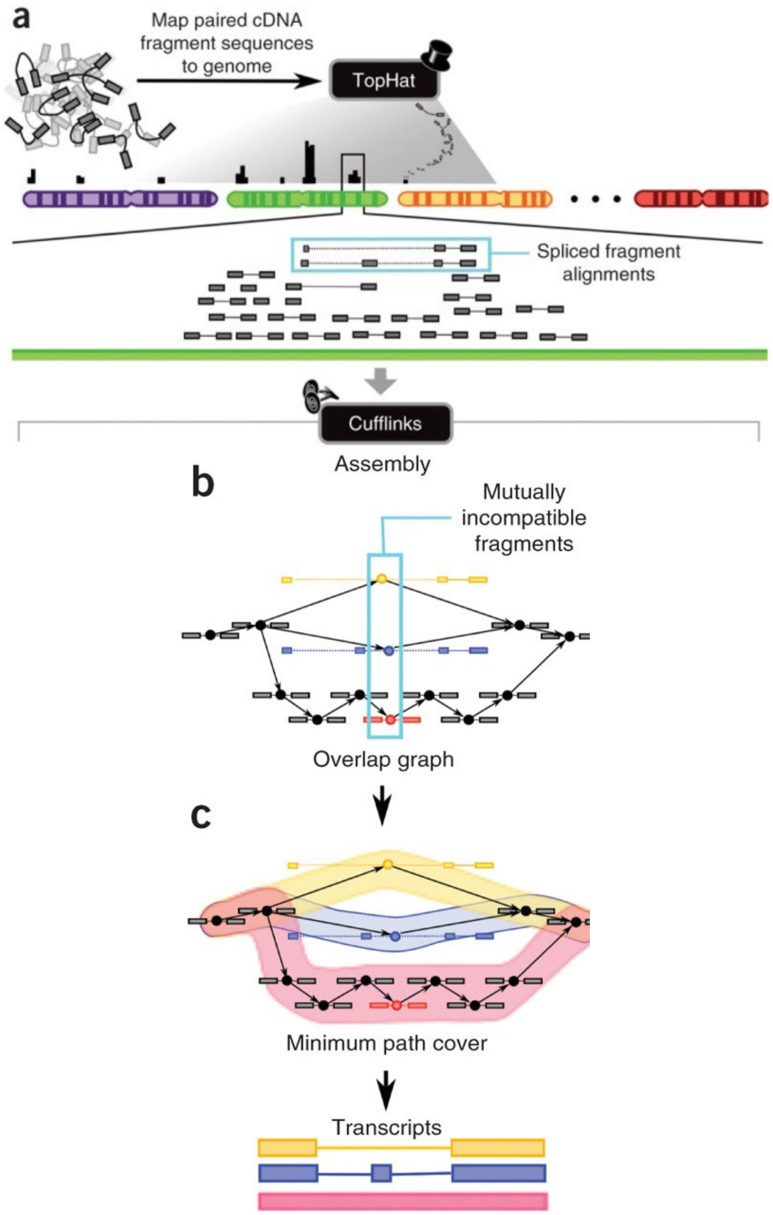
Step 4: Reconstruction of the transcriptome (transcriptome assembly)

How do I get a full catalog of transcripts and their variations from short reads?

Transcriptome assembly became possible due to new technology (RNA-seq) producing millions of short reads (with >100x coverage per base pair of a transcript)

→ Assemble near complete snapshot of the transcriptome
(including isoformes and rare transcripts)

Isoform detection using `cufflinks` (transcript assembly from mapped reads)


a
Map paired cDNA fragment sequences to genome
TopHat
Spliced fragment alignments
Cufflinks
Assembly
b
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts

a Paired-end reads (fragments) are mapped with a splice-aware aligner (e.g. TopHat2)

b `cufflinks` connects overlapping (compatible) fragments in an overlap graph
- node: fragment
- edge: connects compatible fragments
Overlap implies that fragments originate from the same isoform

c Paths through graph correspond to sets of mutually compatible fragments that could be merged into complete isoforms

`cufflinks` tries to find the minimum number of paths, each representing a different isoform of a transcript

[Trapnell et al., Nat Biotech, 2010]

Galaxy exercise

http://galaxy.uni-freiburg.de/u/tutor/p/rna-seq

Step 4: Predict novel transcripts with Cufflinks

Hands-on!

Step 5: Quantification of transcript level

What is the expression level of the genomic features (genes, isoforms, ...)?

→ count the reads per feature

- relatively easy: count the number of reads per gene, exon, …
- How to handle multi-mapping reads (i.e. reads with multiple alignments)?
  - discard multi-mapping reads: ok at gene and exon level
  - probabilistic selection: recommended for repetitive elements
- How to distinguish between different isoforms?
  - gene level (i.e. do not distinguish between isoforms)
  - transcript level (requires to estimate abundance of isoforms)
  - exon level

Galaxy exercise

http://galaxy.uni-freiburg.de/u/tutor/p/rna-seq

Step 5: Count the number of reads per annotated gene with htseq-count

Hands-on!

Normalization

- **Normalization** aims to make expression levels **comparable** across
  - **features** (genes, isoforms, ...)
  - **libraries** (samples)
- Normalization methods:
  - RPKM / FPKM (Cufflinks /Cuffdiff) [Mortazavi et al., Nat Meth, 2008]
  - TMM (edgeR) [Robinson & Oshlack, Genome Biol, 2010]
  - **DESeq2** (DESeq2) [Love et al., Genome Biol, 2014]

Normalization across samples

"Only the **DESeq and TMM normalization methods are robust to** the presence of **different library sizes** and widely **different library compositions**…"

Dillies et al., Brief Bioinf, 2013

- `DESeq`:
  normalise counts k_{ij} for gene i in library j by size factor s_j

$$s_j / s_{j'} = \operatorname*{median}_i \{k_{ij} / k_{ij'}\}$$

Analysis of differential gene expression

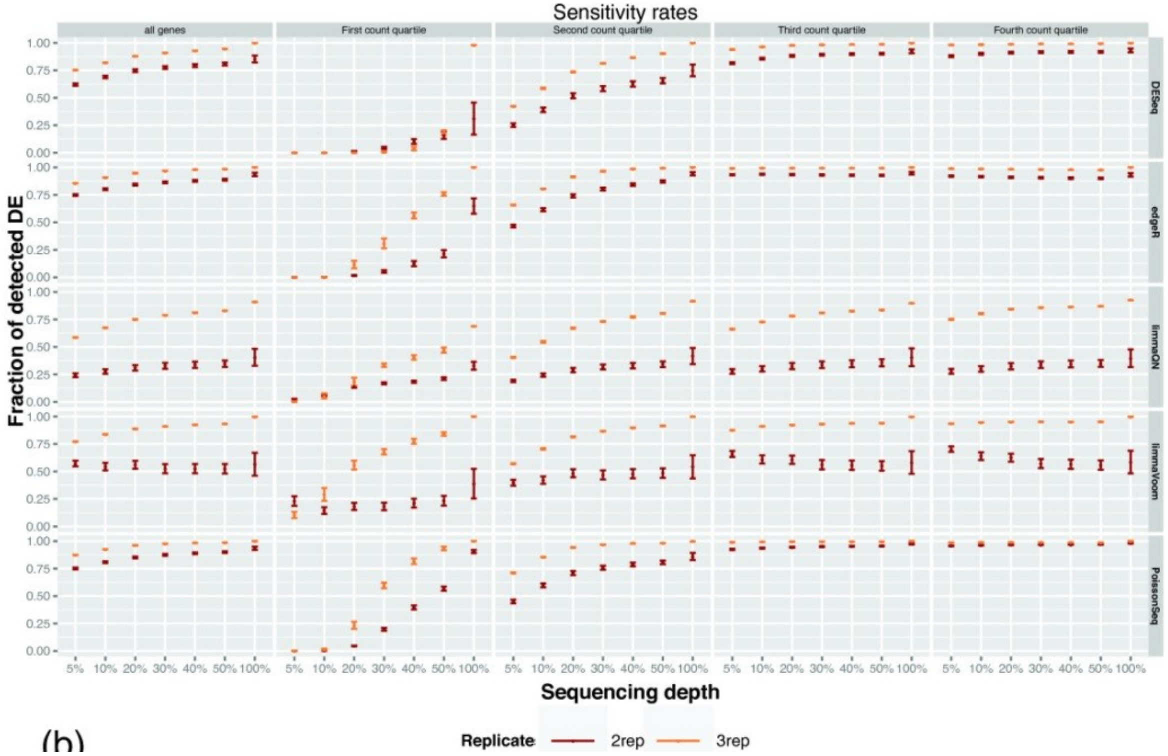
Selected programs to analyse differential expression (DE)

- at gene level:
  - `DEseq2` [Love et al., Genome Biol., 2014]
  - `edgeR` [Robinson et al., Bioinformatics, 2010]
- at transcript level:
  - `Cuffdiff2` [Trapnell et al., Nat. Biotech., 2013]
- differential usage of exons:
  - `DEXseq` [Anders et al., Genome Res., 2012]

Core idea:

- model the gene counts by negative binomial distribution
- account for variability of gene expression across biological replicates

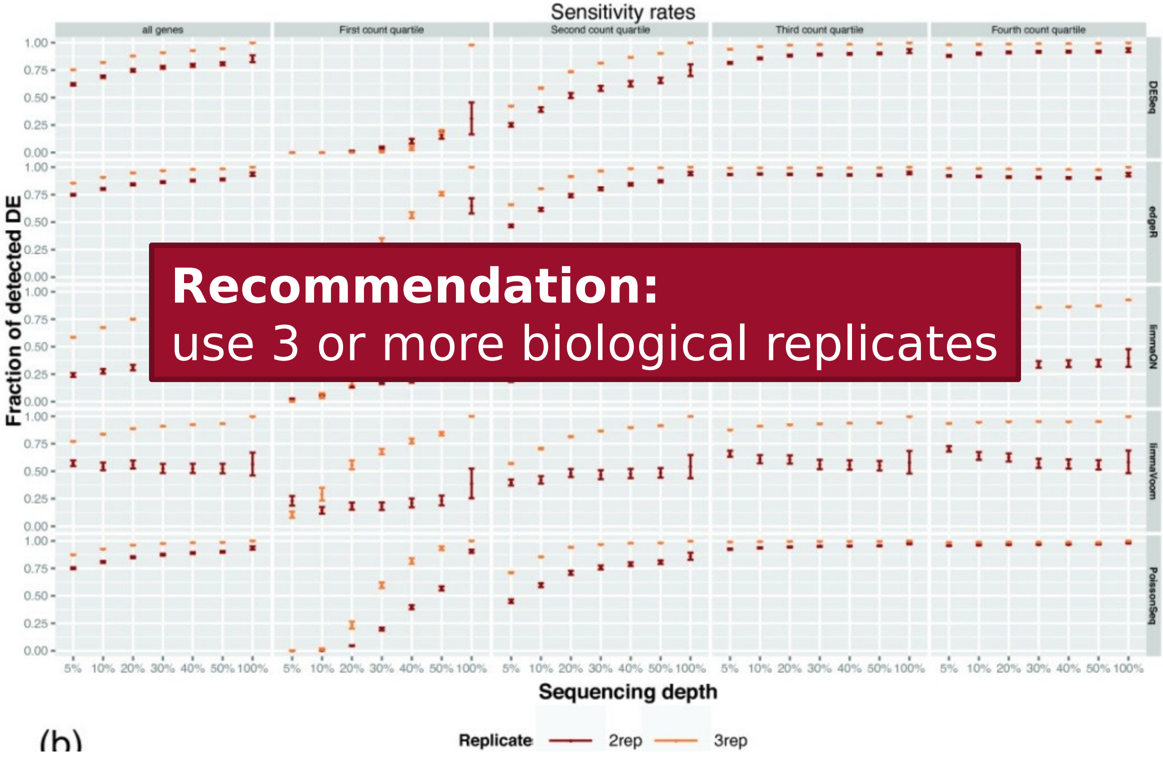
Impact of sequencing depth and number of replicates on differential expression analysis



- number of replicates has greater effect on DE detection accuracy than sequencing depth (more replicates = increased statistical power)
- DE detection of lowly expressed genes is very sensitive to number of reads and replication
- DE detection of highly expressed genes possible already at low sequencing depth

[Rapaport et al., Genome Biol., 2013]

Impact of sequencing depth and number of replicates on differential expression analysis



Recommendation:
use 3 or more biological replicates

- number of replicates has greater effect on DE detection accuracy than sequencing depth (more replicates = increased statistical power)
- DE detection of lowly expressed genes is very sensitive to number of reads and replication
- DE detection of highly expressed genes possible already at low sequencing depth

[Rapaport et al., Genome Biol., 2013]

Galaxy exercise

http://galaxy.uni-freiburg.de/u/tutor/p/rna-seq

Steps 5&6: Analysing differential gene expression with DESeq2

Hands-on!

Galaxy exercise

http://galaxy.uni-freiburg.de/u/tutor/p/rna-seq

Step 7: Functional enrichment among differentially expressed genes

Hands-on!

Differential splicing analysis

- selected programs to analyse differential splicing
  - at isoform level:
    - `Cuffdiff2` [Trapnell et al., Nat. Biotech., 2013]
    - `MISO` [Katz et al., Nat. Methods, 2010]
  - at exon level:
    - `DEXseq` [Anders et al., Genome Res., 2012]
  - at junction level:
    - `MATS` [Shen et al., Nucleic Acids Res., 2012]

Recent review:
Hooper JE. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. Human Genomics, 2014

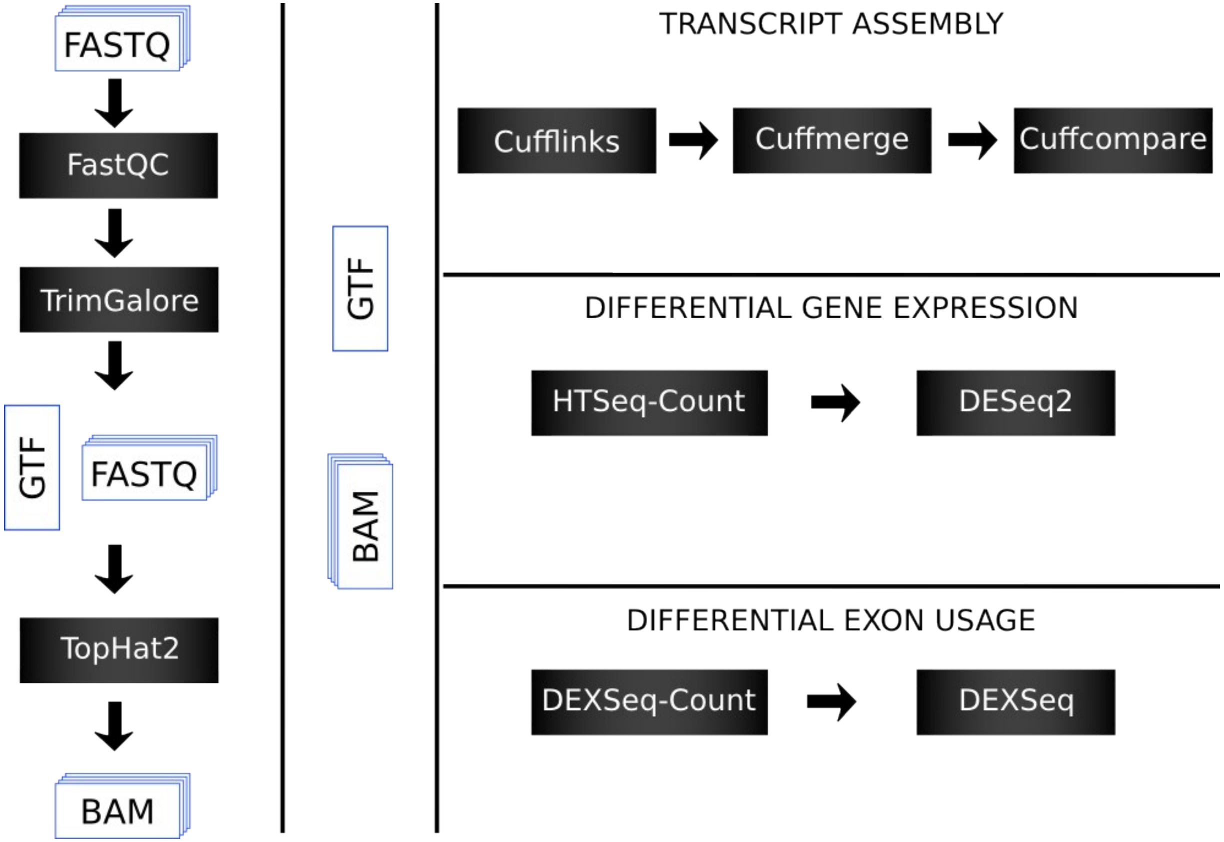
Galaxy exercise

http://galaxy.uni-freiburg.de/u/tutor/p/rna-seq

Steps 8&9: Inference of differential exon usage with DEXSeq

Hands-on!

Tutorial Overview



The End.

Thank you for your attention!

References

Brooks, A. N. et al. Conservation of an RNA regulatory map between Drosophila and mammals. Genome Res. 21, 193–202 (2011)

Dillies, M.-A. et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 14, 671–683 (2013)

Katz, Y. et al. Sashimi plots: Quantitative visualization of alternative isoform expression from RNA-seq data. bioRxiv (2014)

Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 14, R36 (2013)

Korf, I. Genomics: the state of the art in RNA-seq analysis. Nat Meth 10, 1165–1166 (2013)

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth 5, 621–628 (2008)

Rapaport, F. et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biology 14, R95 (2013)

Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515 (2010)

Zeng, W. & Mortazavi, A. Technical considerations for functional sequencing assays. Nat Immunol 13, 802–807 (2012)