Vall d'Hebron
Institut de Recerca

Introduction to RNA-Seq Data Analysis

Curs de Bioinformàtica per a la Recerca Biomèdica
21/11/2018

Ricardo Gonzalo Sanz
ricardo.gonzalo@vhir.org

Vall d'Hebron Hospital · Generalitat de Catalunya Departament de Salut · Institut Català de la Salut · UAB Universitat Autònoma de Barcelona · Generalitat de Catalunya Departament d'Economia i Coneixement · Institució CERCA Centres de Recerca de Catalunya · Instituto de Salud Carlos III

Vall d'Hebron
Institut de Recerca

TABLE OF CONTENTS

1. What is RNA-seq?

Vall d'Hebron
Institut de Recerca

- RNA-seq is the high throughput sequencing of cDNA using NGS technologies

- RNA-seq works by sequencing every RNA molecule and profiling the expression of a particular gene by counting the number of time its transcripts have been sequenced.

- The summarized RNA-seq data is widely known as *count table*

	Condition A			Condition B		
Gene1	4	0	2	12	14	13
Gene2	0	23	50	47	22	0
Gene3	0	2	6	13	11	15
...	...	...	...	...	...	...
GeneG	156	238	37	129	51	118

1. What is RNA-seq?

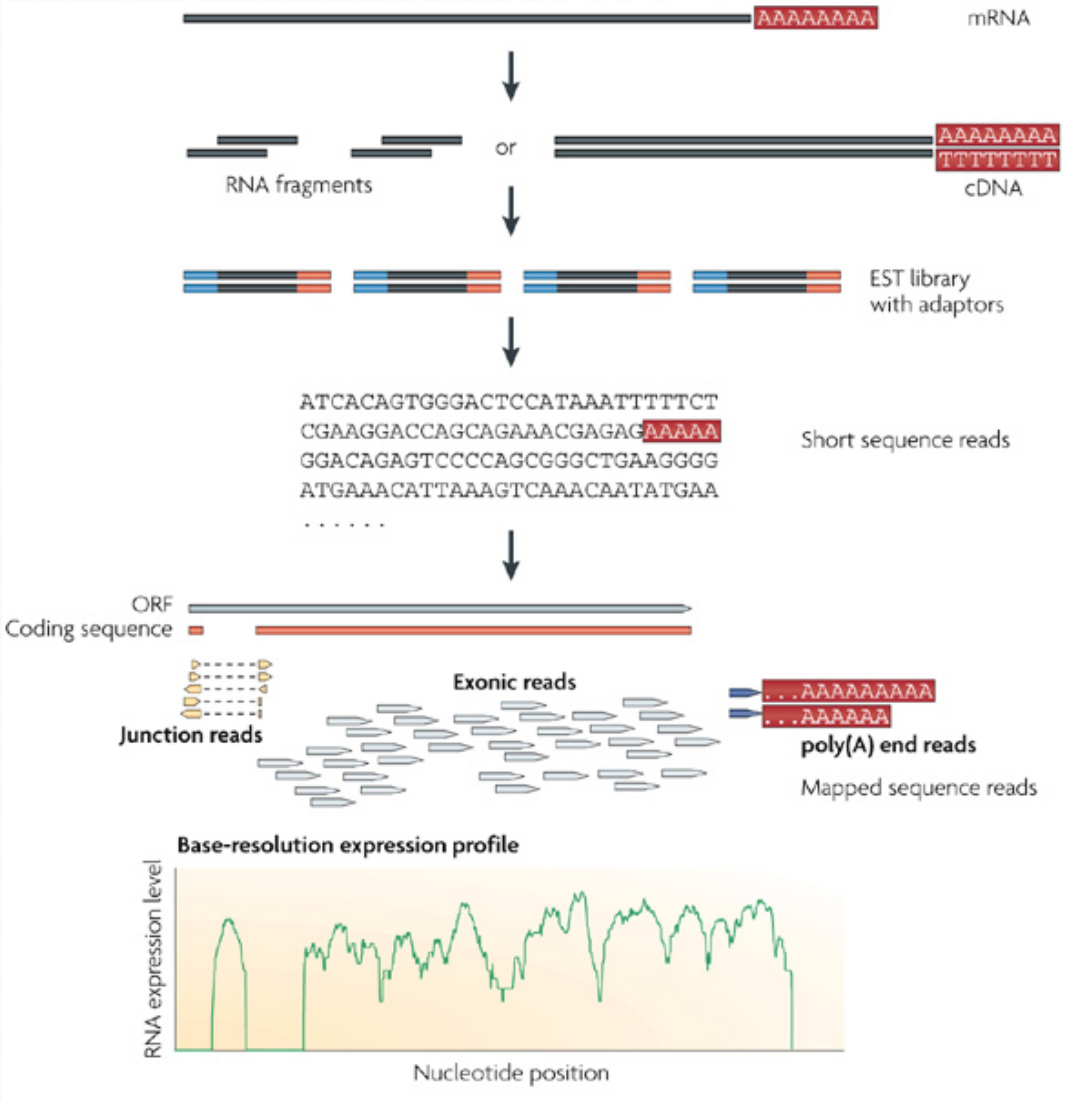
Vall d'Hebron
Institut de Recerca

Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation (see main text).

Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology.

The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom; a yeast ORF with one intron is shown.

Nature Reviews Genetics **10**, 57-63 (January 2009)



2. Why RNA-seq instead of DNA-seq?

Vall d'Hebron
Institut de Recerca

- Functional studies
  - Genome may be constant but an experimental condition has a pronounced effect on gene expression
    - e.g. Drug treated vs. untreated cell line
    - e.g. Wild type versus knock out mice

- Some molecular features can only be observed at the RNA level
  - Alternative isoforms, fusion transcripts, RNA editing

- Predicting transcript sequence from genome sequence is difficult
  - Alternative splicing, RNA editing, etc.

Vall d'Hebron
Institut de Recerca

TABLE OF CONTENTS

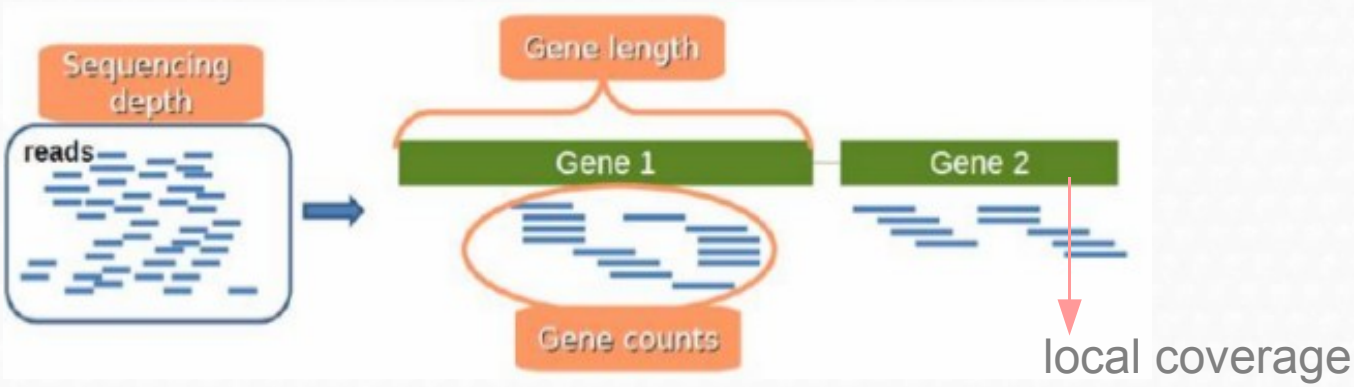
3. Basic key concepts

- Sequencing depth: Total number of reads mapped to the genome. (Library size) Could also be applied to samples.

- Coverage: Number of reads mapped to a specific region (average of them if we are talking about the whole genome...)

- Gene length: Number of bases that a gene has.

- Gene counts: Number of reads mapping to that gene (expression measurement).



4. Main challenges in RNA-seq

Vall d'Hebron
Institut de Recerca

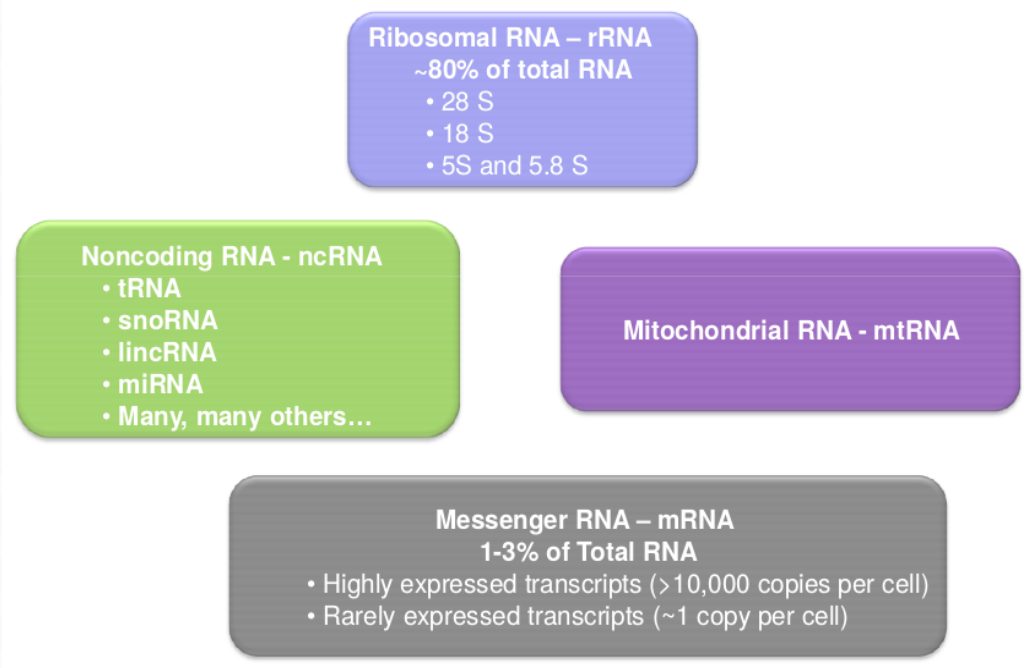
- **Sample**
  - Purity? Quantity? Quality?

- **RNAs consist of small exons that may be separated by large introns**
  - Mapping reads to genome is challenging
  - Non-uniformity coverage of the genome

- **The relative abundance of RNAs vary wildly**
  - 10^5 – 10^7 orders of magnitude
  - Since RNA sequencing works by random sampling, a small fraction of highly expressed genes may consume the majority of reads

- **RNAs come in a wide range of sizes**
  - Small RNAs must be captured separately
  - PolyA selection of large RNAs may result in 3' end bias

- **RNA is fragile compared to DNA** (easily degraded)



T. Stempfl, 2012

4. Main challenges in RNA-seq (and other NGS cases)

Vall d'Hebron
Institut de Recerca

- Independently of the software used,
  one needs to think about
  DATA STORAGE & MANAGEMENT!!

1 Illumina Flow Cell equals up to 1.5 Bn individual Clusters
= 3 Bn Reads
= 300 Gbases raw sequence
= 2.5 TByte of disk space (raw data)
> 100 GByte of disk space (fastq data)

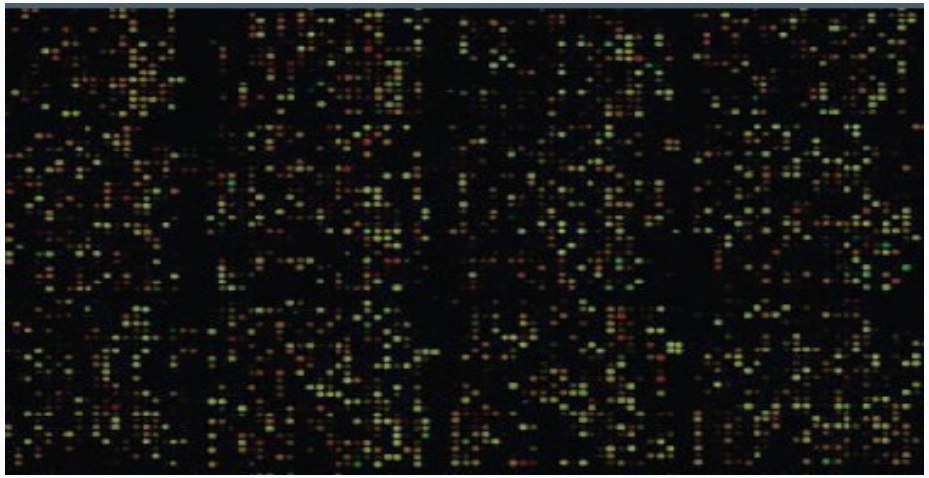
T. Stempfl, 2012

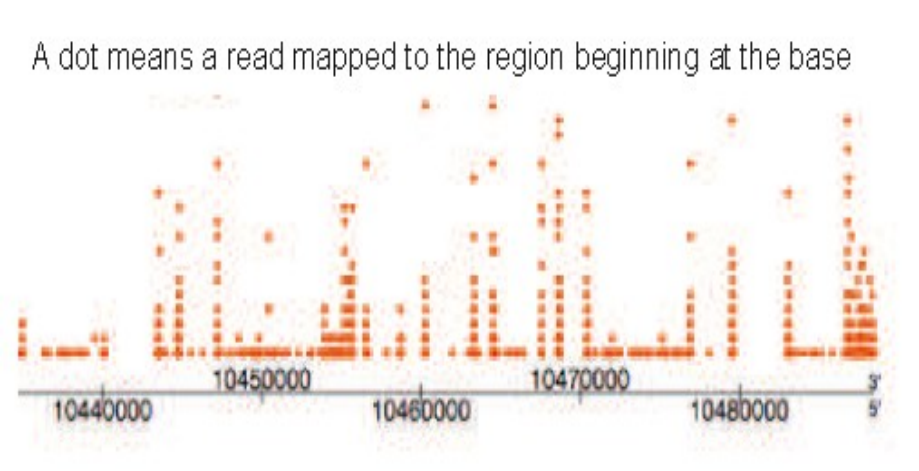
Vall d'Hebron
Institut de Recerca

TABLE OF CONTENTS

5. RNA-seq VS Microarrays

RNA-seq can be seen as the NGS-counterpart of microarrays

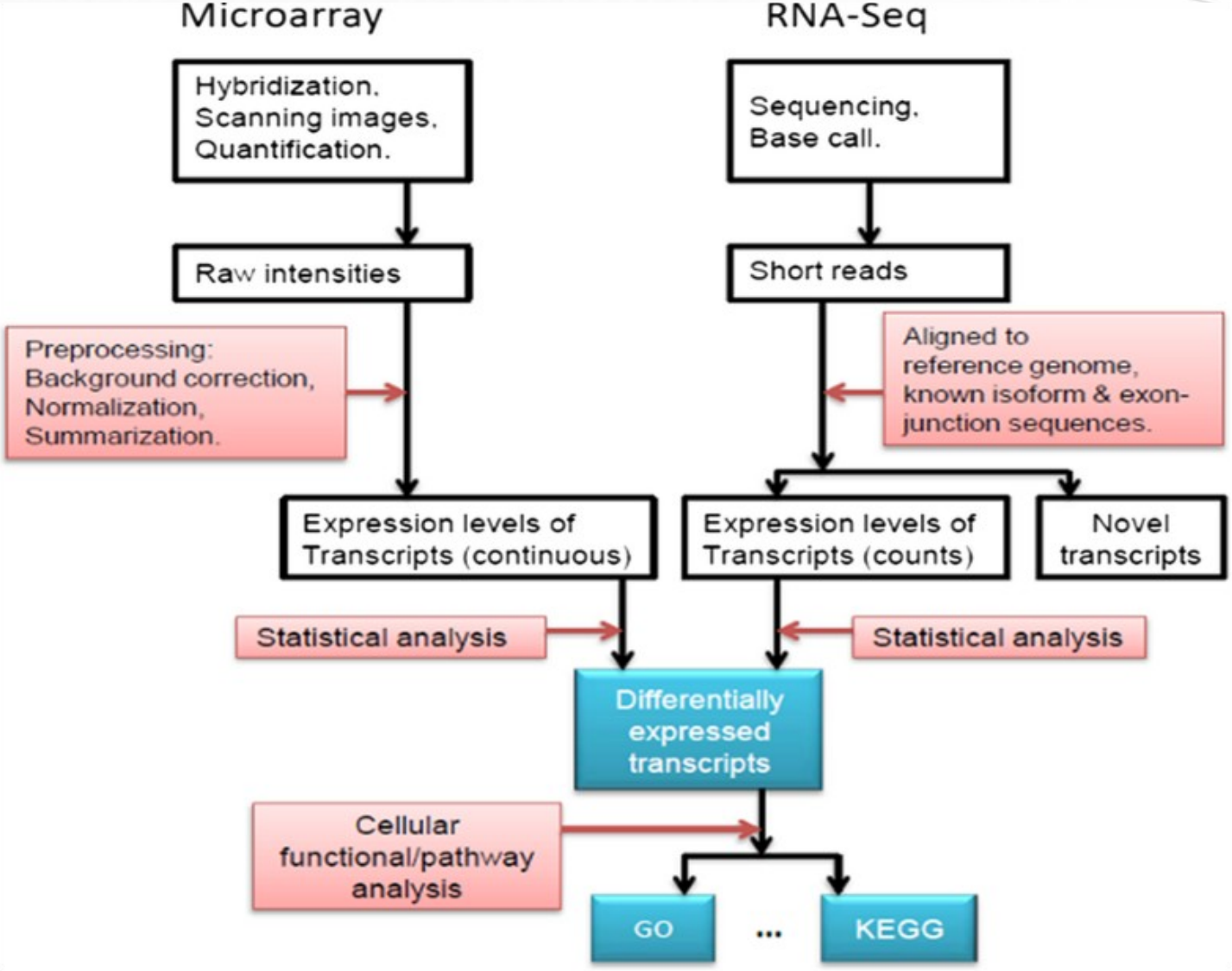




- **Analog Signal**
- Easy to convey the signal's information
- Continuous strength
- Signal loss and distortion

- **Digital Signal**
- Harder to achieve & interpret
- Reads counts: discrete values
- Weak background or no noise

5. RNA-seq VS Microarrays



5. RNA-seq VS Microarrays

Vall d'Hebron
Institut de Recerca

Pros and cons of both technologies

Microarrays

- Costs,
- well established methods, small data
- Hybridization bias,
- sequence must be known

RNA-seq

- High reproducibility,
- not limited to expression
- Costs,
- complexity of analysis

"**High correlation** between gene expression profiles generated by the two platforms."

"RNA-Seq sequencing technology is new to most researchers, more expensive than microarray, **data storage is more challenging** and **analysis is more complex**."

Zhao et al, 2014

6. Resources, Tools, Guidelines

Vall d'Hebron
Institut de Recerca

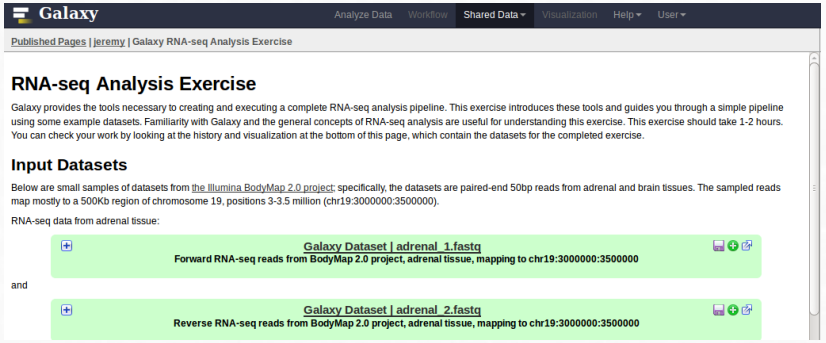
- Alignment
  - BWA (PMID: 20080505)
    - Align to genome + junction database
  - Tophat (PMID: 19289445), MapSplice (PMID: 20802226), hmmSplicer (PMID: 21079731)
    - Spliced alignment to genome
- Expression, differential expression alternative expression
  - Cufflinks/Cuffdiff (PMID: 20436464), ALEXA-seq (PMID: 20835245), RUM (PMID: 21775302)
- Fusion detection
  - ChimeraScan (PMID: 21840877), Defuse (PMID: 21625565), Comrad (PMID: 21478487)
- Transcript assembly
  - Trinity (PMID: 21572440), Oases (PMID: 22368243), Trans-ABySS (PMID: 20935650)
- Mutation calling
  - SNVMix (PMID: 20130035)
- **Visit the 'SeqAnswers', "RNA-Seq Blog" or 'BioStar' forums for more recommendations and discussion**
  - http://seqanswers.com/
  - http://www.rna-seqblog.com/
  - http://www.biostars.org/

ENCODE

RNA-Seq Guidelines (ENCODE Consortium)

RNA-Seq Blog
Transcriptome Sequencing Research & Industry News

http://www.rna-seqblog.com/

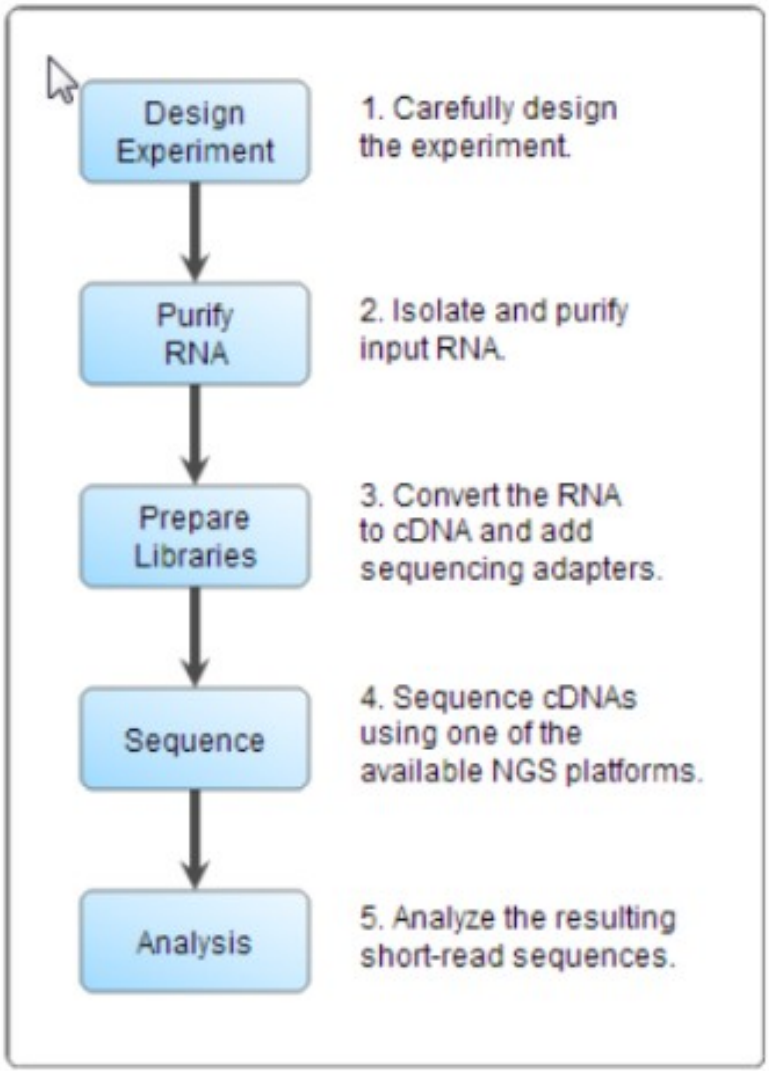


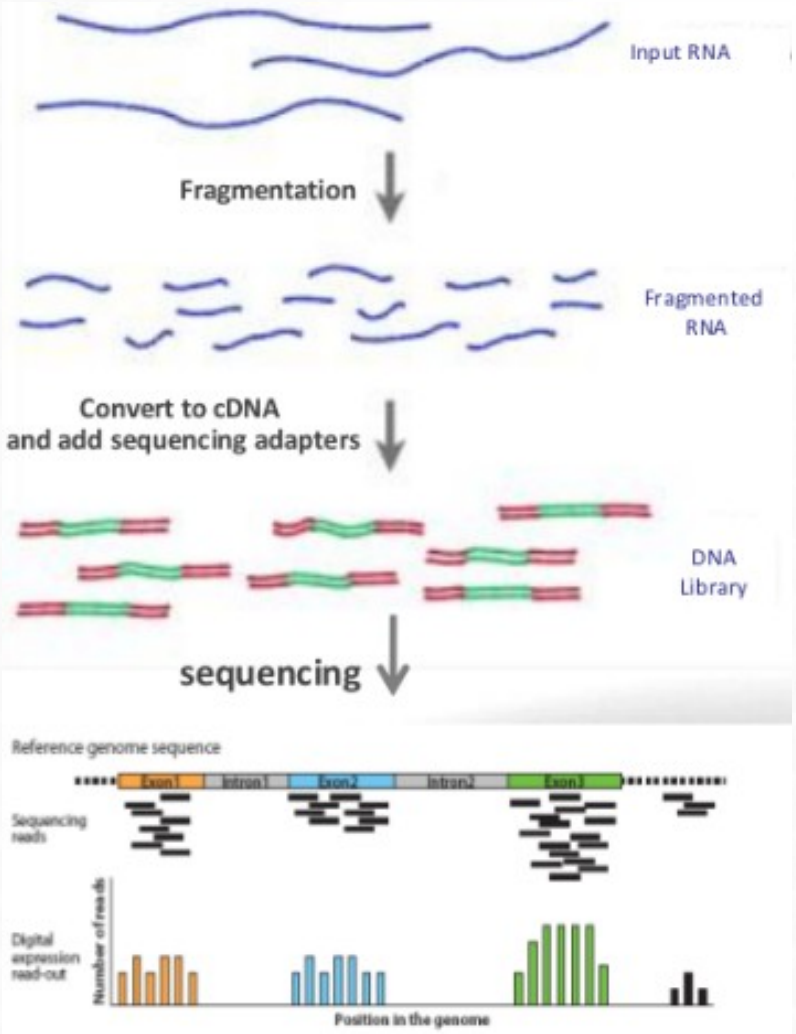
https://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise

Vall d'Hebron
Institut de Recerca

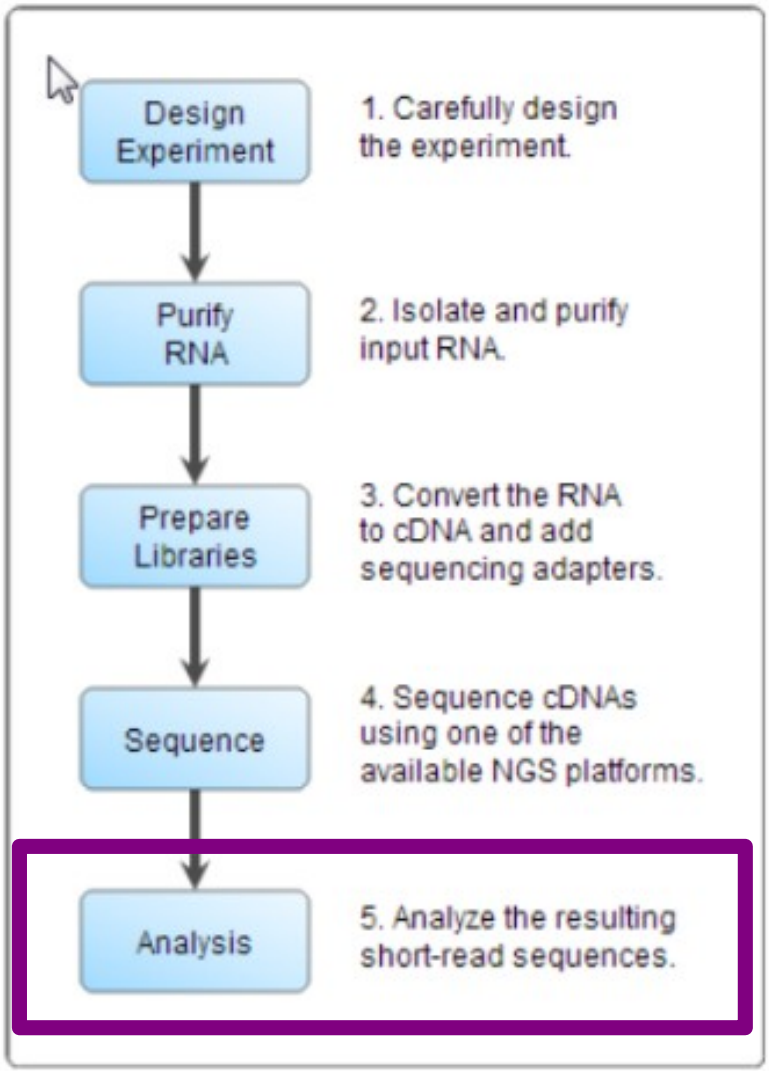
TABLE OF CONTENTS

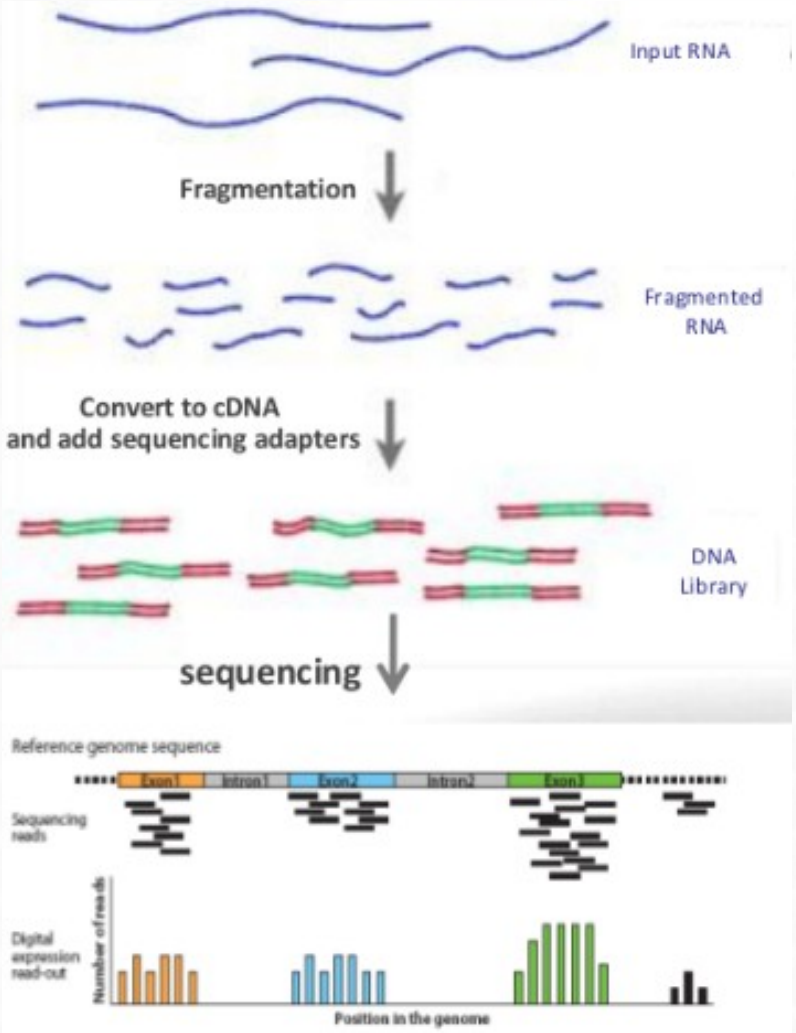
7. RNA-seq analysis pipeline(s)





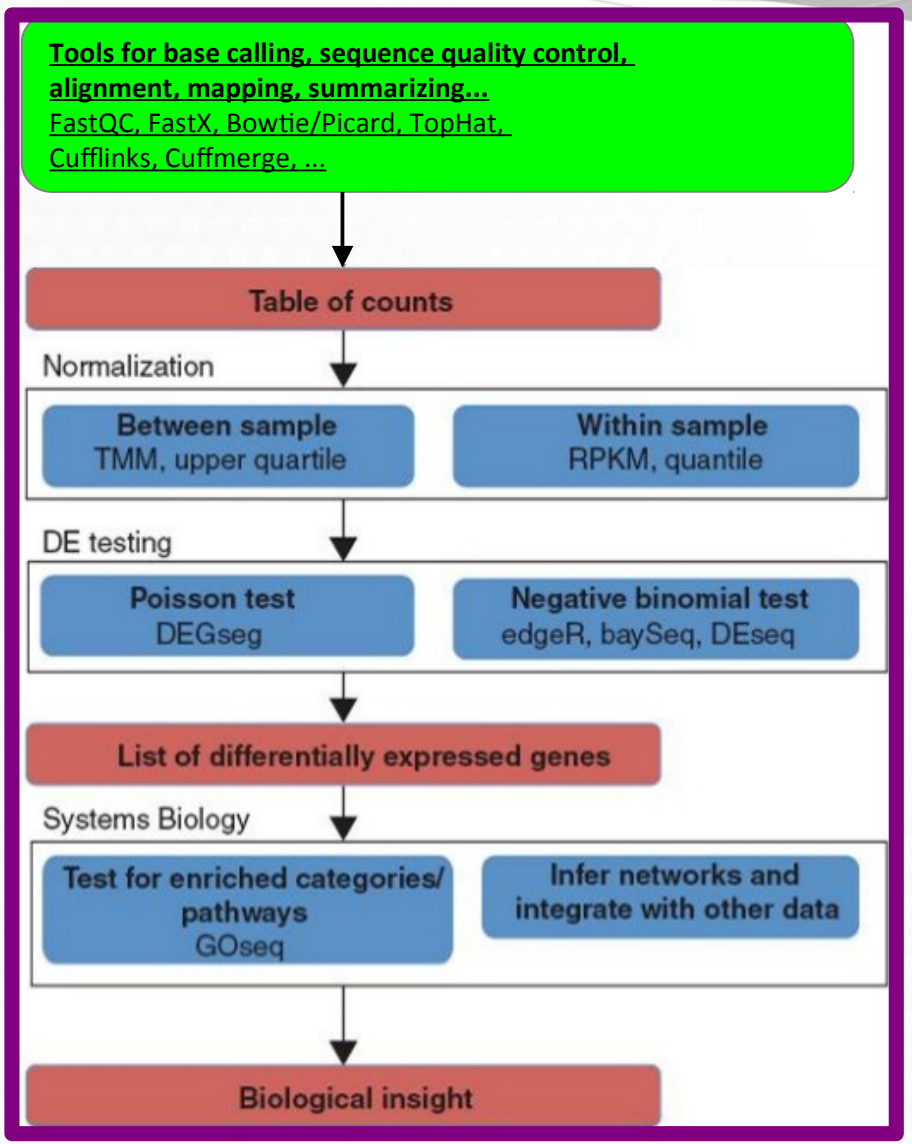
7. RNA-seq analysis pipeline(s)



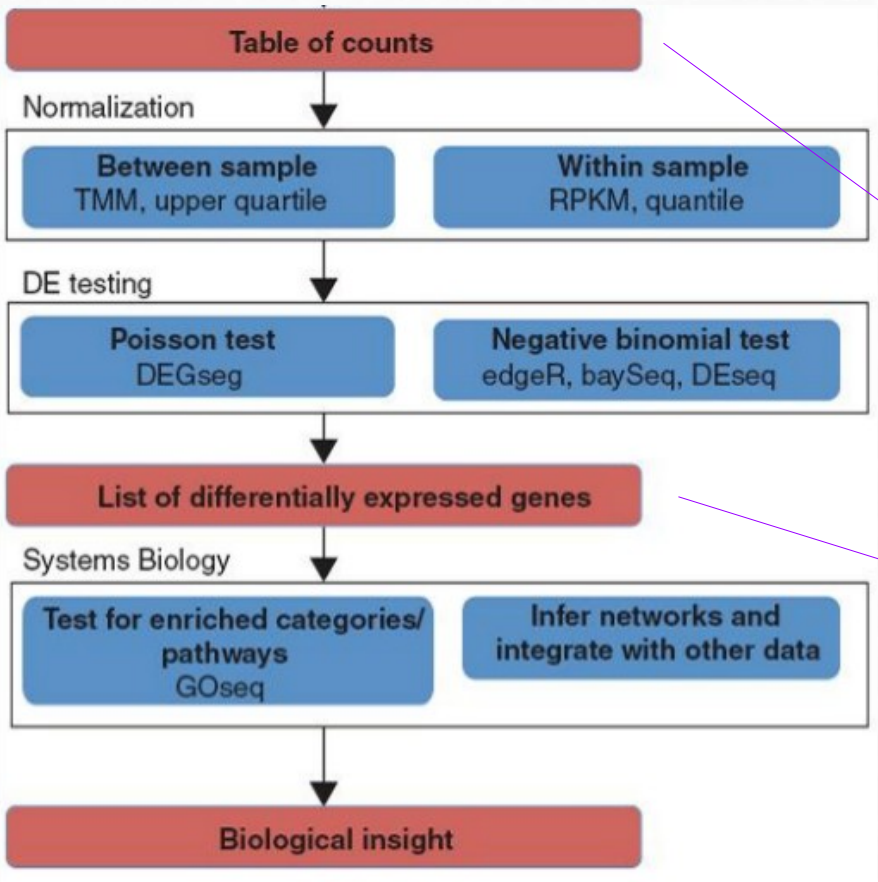


7. RNA-seq analysis pipeline(s)

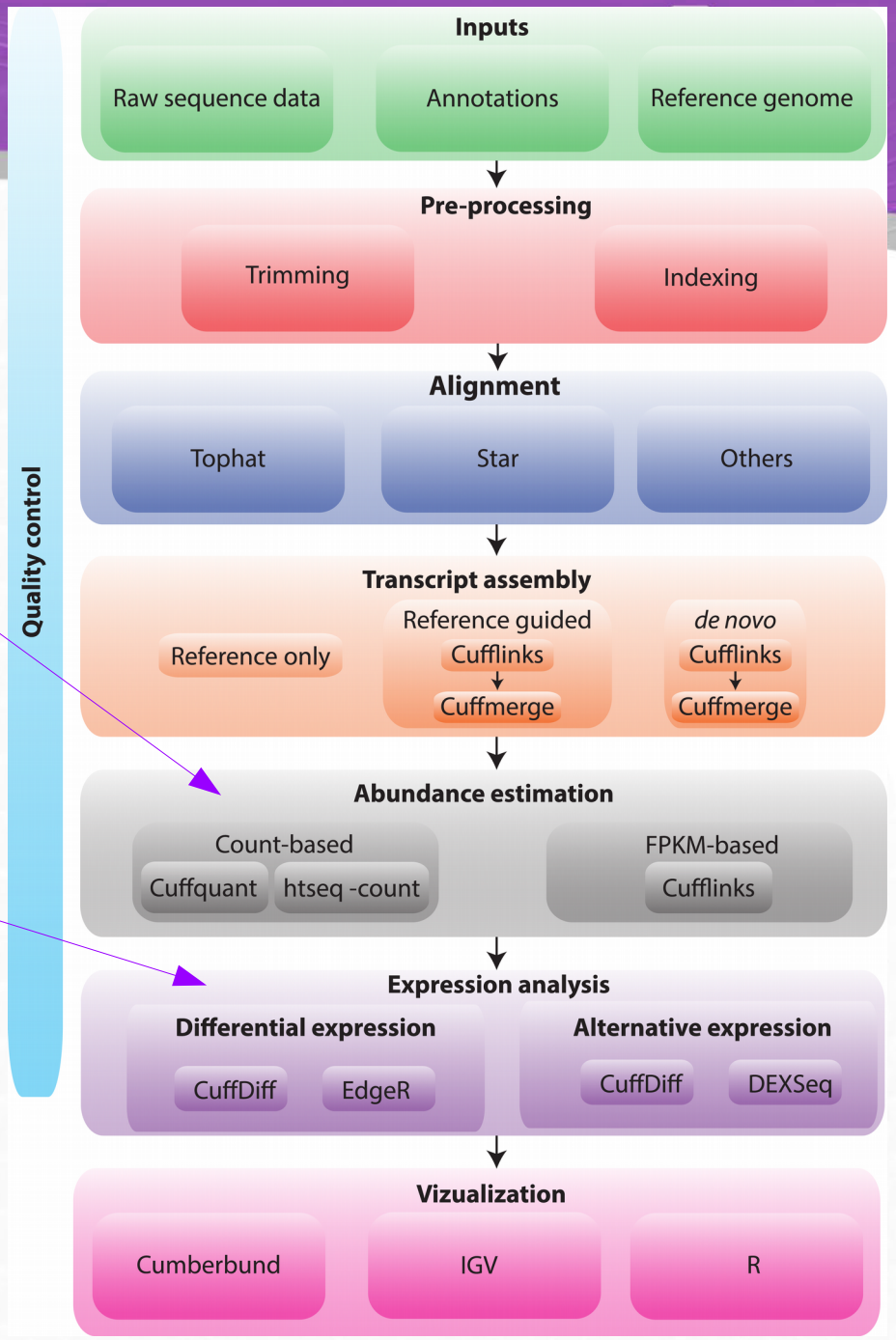
- Reads are mapped to the reference genome or transcriptome
- Mapped reads are assembled into expression summaries (tables of counts, showing how many reads are in coding region, exon, gene or junction)
- Data is normalized
- Statistical testing of differential expression (DE) is performed, producing a list of genes with p-values and fold changes
- Similar downstream analysis than microarray results (Functional Annotations, Gene Enrichment Analysis; Integration with other data...)



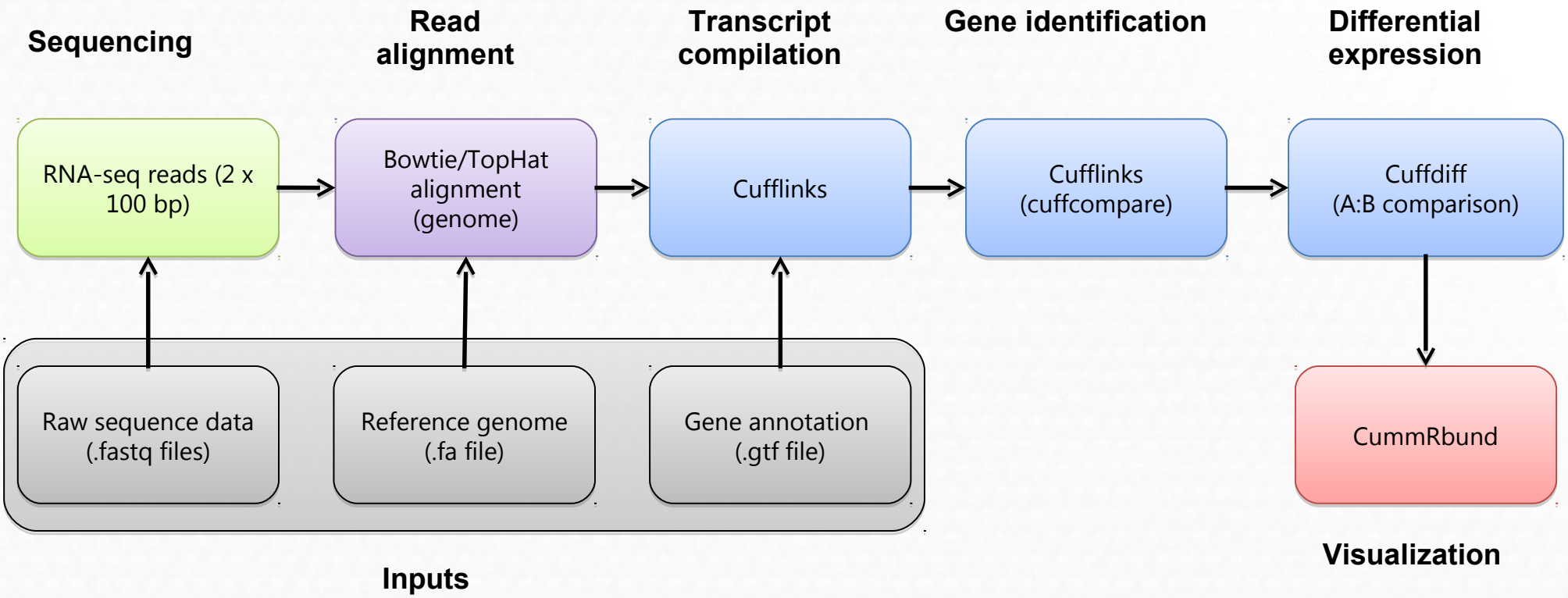
7. RNA-seq analysis pipeline(s)



Griffith, 2015 DOI: 10.1371/journal.pcbi.1004393



7. RNA-seq analysis pipeline(s)



7. RNA-seq analysis pipeline(s)

MAPPING

Main Issues:

- Number of allowed mismatches
- Number of multi-hits
- Mates expected distance
- Considering exon junctions

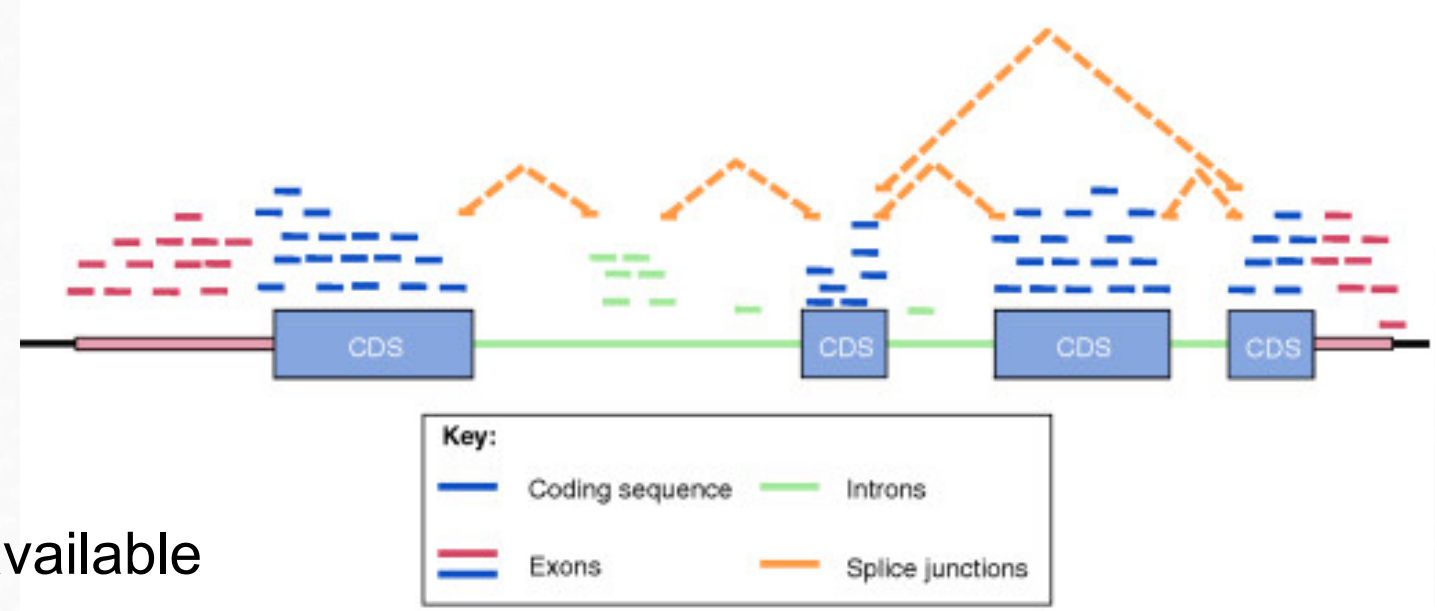
End up with a list of
# of reads (counts) per transcript

These will be our discrete)
response variable

7. RNA-seq analysis pipeline(s)

SUMMARISATION

- Summarise & aggregate reads over *some biologically meaningful unit*, such as exons, transcripts, genes, regions...



- Many methods available
  - Counts # of reads overlapping the exons in a gene,
  - Include reads along the whole length of the gene and thereby incorporate reads from 'introns'.
  - Include only reads that map to coding sequence…