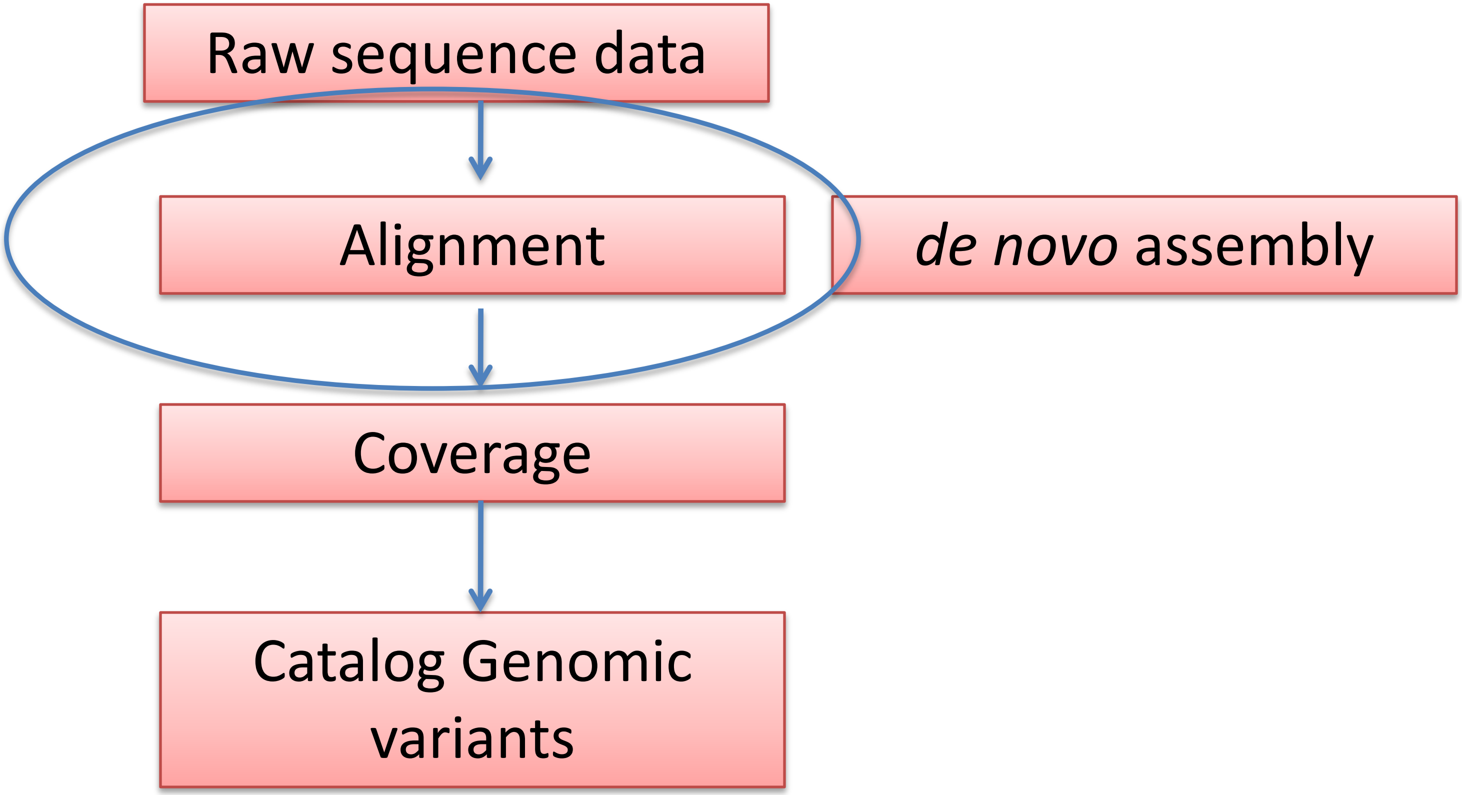
# RNA-Seq and differential expression

London School of Hygiene and Tropical Medicine

# Some aspects of the course


Raw sequence data

Alignment

*de novo* assembly

Coverage

Catalog Genomic variants

Population genetics

Whole genome Association studies

# Outline

- What is RNA-seq?
- Why?
- How?
- Differential expression
- Exercises

# RNA-seq

- Use of new sequencing technologies to capture and study the transcriptome
- Identify novel transcripts
- Exon/transcript boundaries
- Splice junctions/alternative splicing
- Measure transcript abundance
- Gene expression differences across multiple samples (i.e. differential expression)

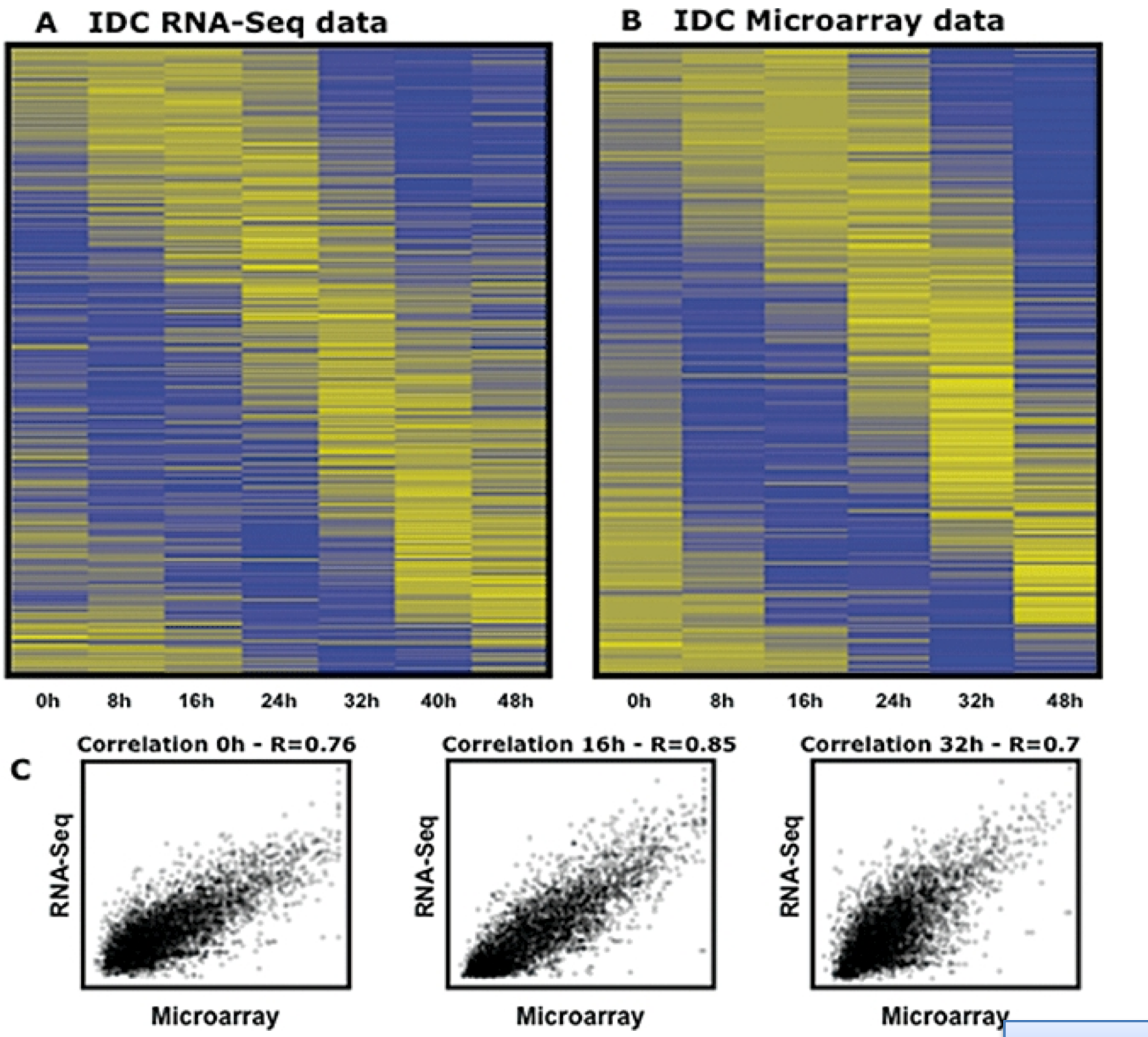
# Comparisons with previous approaches

| Technology | Tiling microarray | cDNA or EST sequencing | RNA-Seq |
|---|---|---|---|
| ***Technology specifications*** | | | |
| Principle | Hybridization | Sanger sequencing | High-throughput sequencing |
| Resolution | From several to 100 bp | Single base | Single base |
| Throughput | High | Low | High |
| Reliance on genomic sequence | Yes | No | In some cases |
| Background noise | High | Low | Low |
| ***Application*** | | | |
| Simultaneously map transcribed regions and gene expression | Yes | Limited for gene expression | Yes |
| Dynamic range to quantify gene expression level | Up to a few-hundredfold | Not practical | >8,000-fold |
| Ability to distinguish different isoforms | Limited | Yes | Yes |
| Ability to distinguish allelic expression | Limited | Yes | Yes |
| ***Practical issues*** | | | |
| Required amount of RNA | High | High | Low |
| Cost for mapping transcriptomes of large genomes | High | High | Relatively low |

Wang et al. 2009, *Nature Reviews Genetics* **10**, 57-63

# How does RNA-seq compare with microarray?

Otto et al. 2010

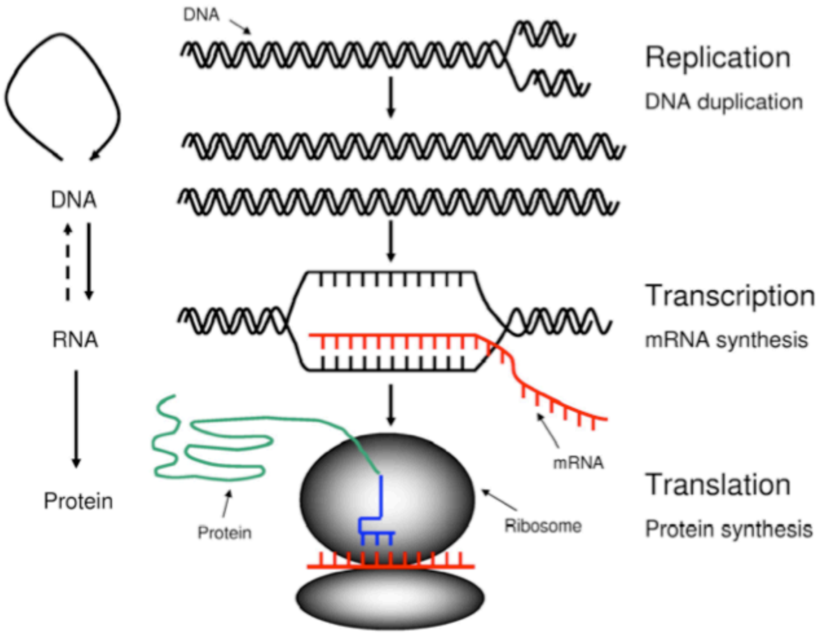


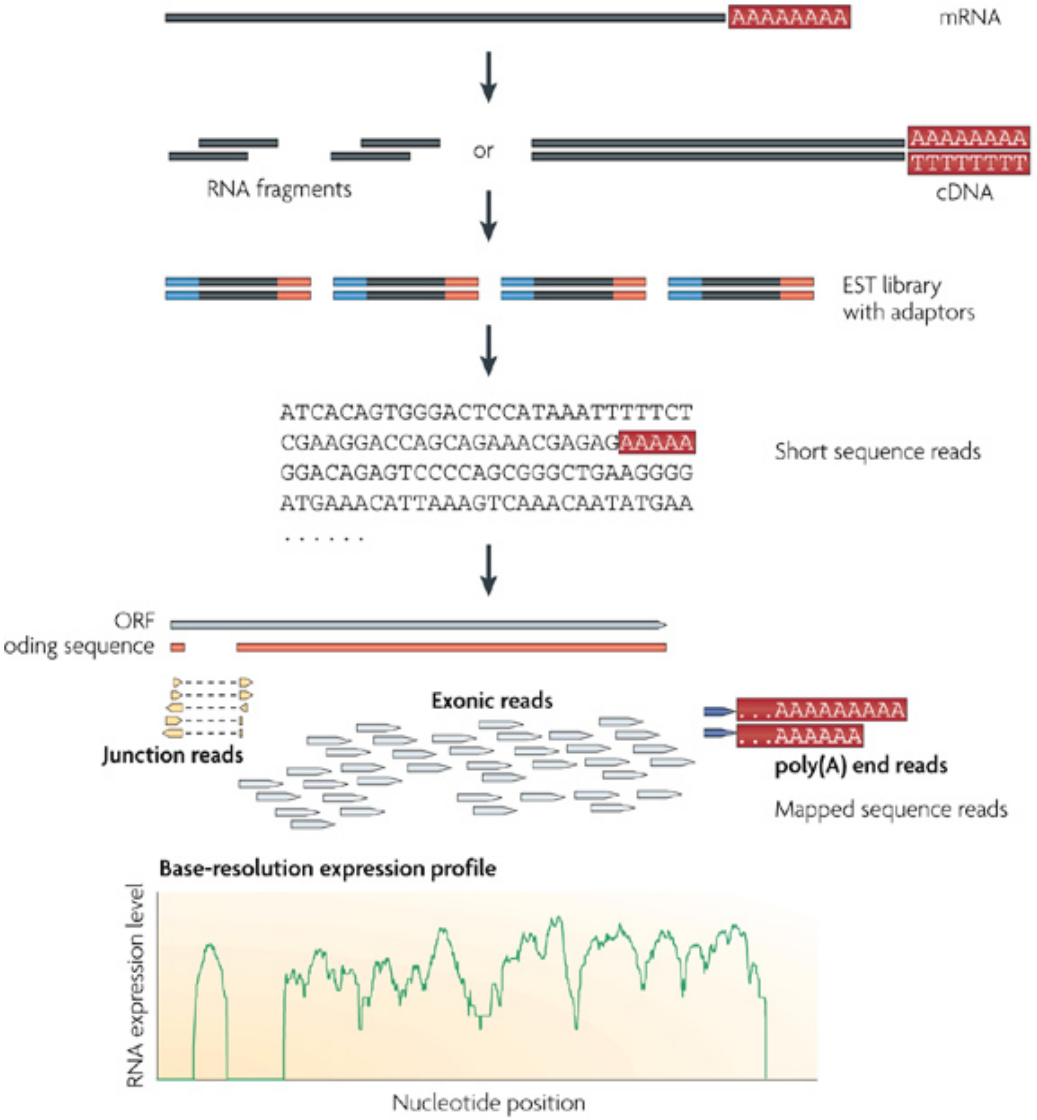
Bozdech et al. 2003

# Considerations for library preparation

- Total RNA?
- mRNA?
  - *Depletion of rRNA*
- Strand specific?
- Replicates?
  - Technical (multiple libraries from the same sample)
  - Biological (multiple samples from the same condition)
- Which platform?
- Multiple samples/multiplexing

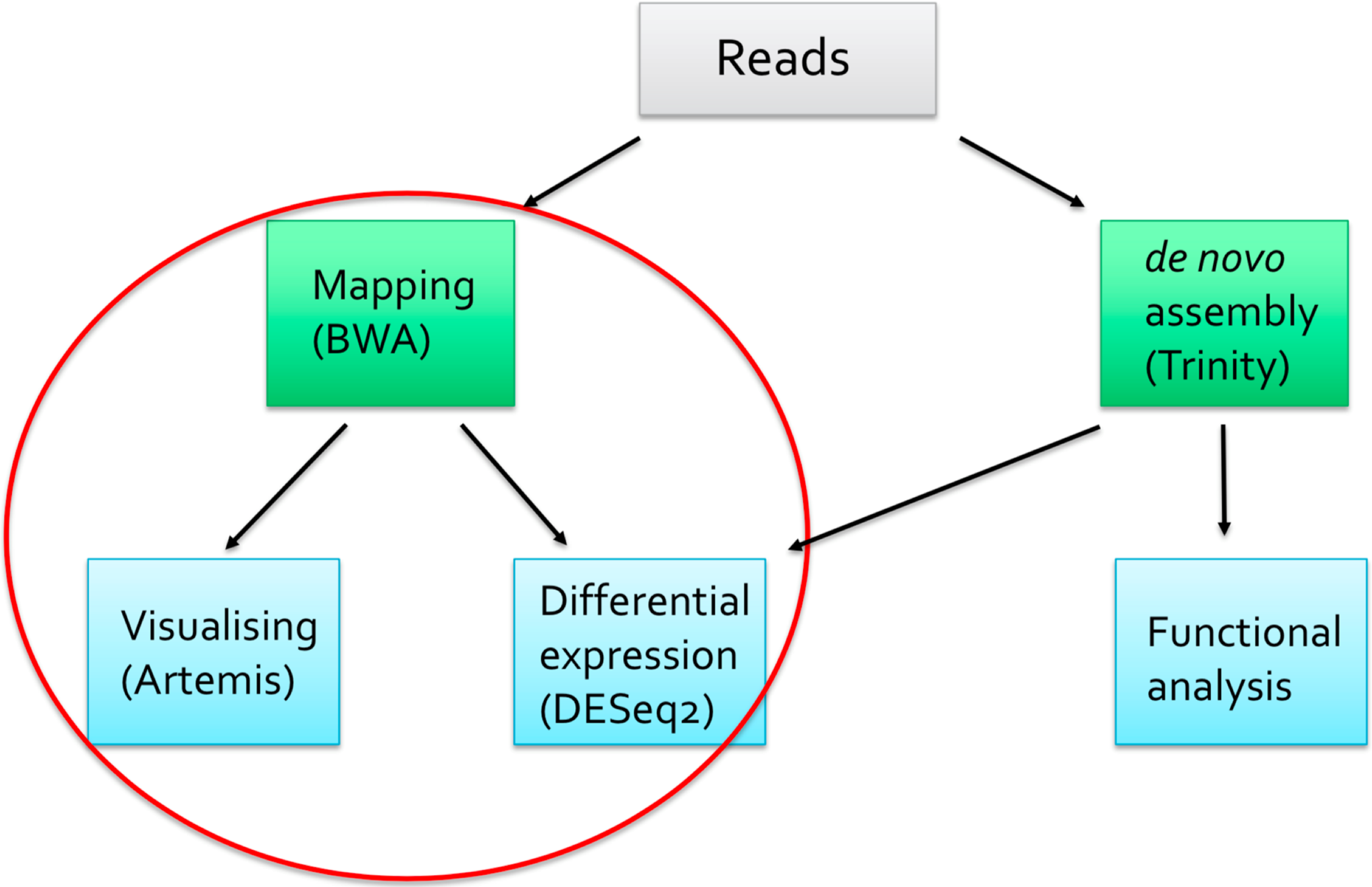
# Sequencing the transcriptome





Nature Reviews | Genetics

# RNA-Seq analysis pipeline



Reads

Mapping (BWA)

*de novo* assembly (Trinity)

Visualising (Artemis)

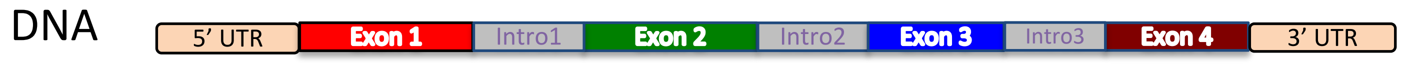
Differential expression (DESeq2)

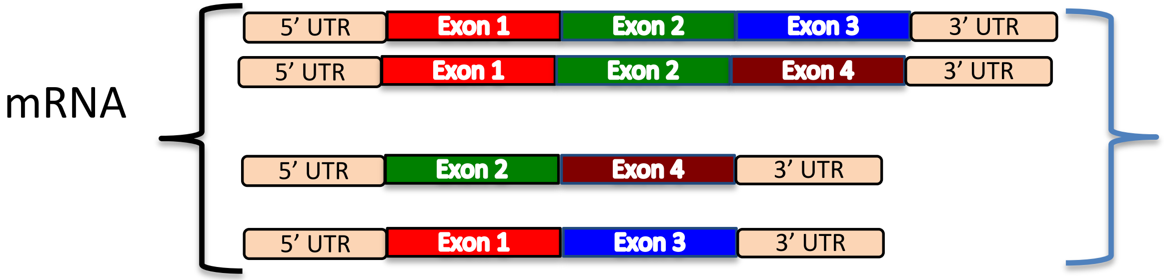
Functional analysis

# Analysis considerations

- Quality Control of sequence data
- Mapping to a reference genome
  - ***Should we align to the transcriptome?***
- Determine which genes are expressed and their abundance
  - Count reads over genes
    - ***Do you think this is enough?***
    - Discard poor quality reads
    - Discard non-uniquely aligned reads. ***Why?***
- ***Do we need a reference genome?***

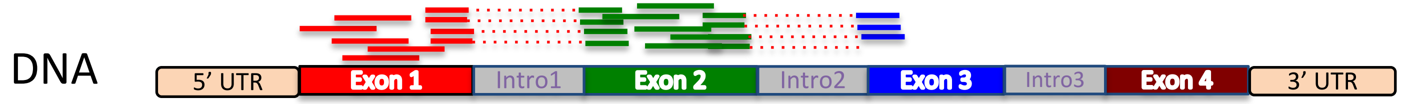
# Mapping





*Alternative splicing --> isoforms

* Exon skipping

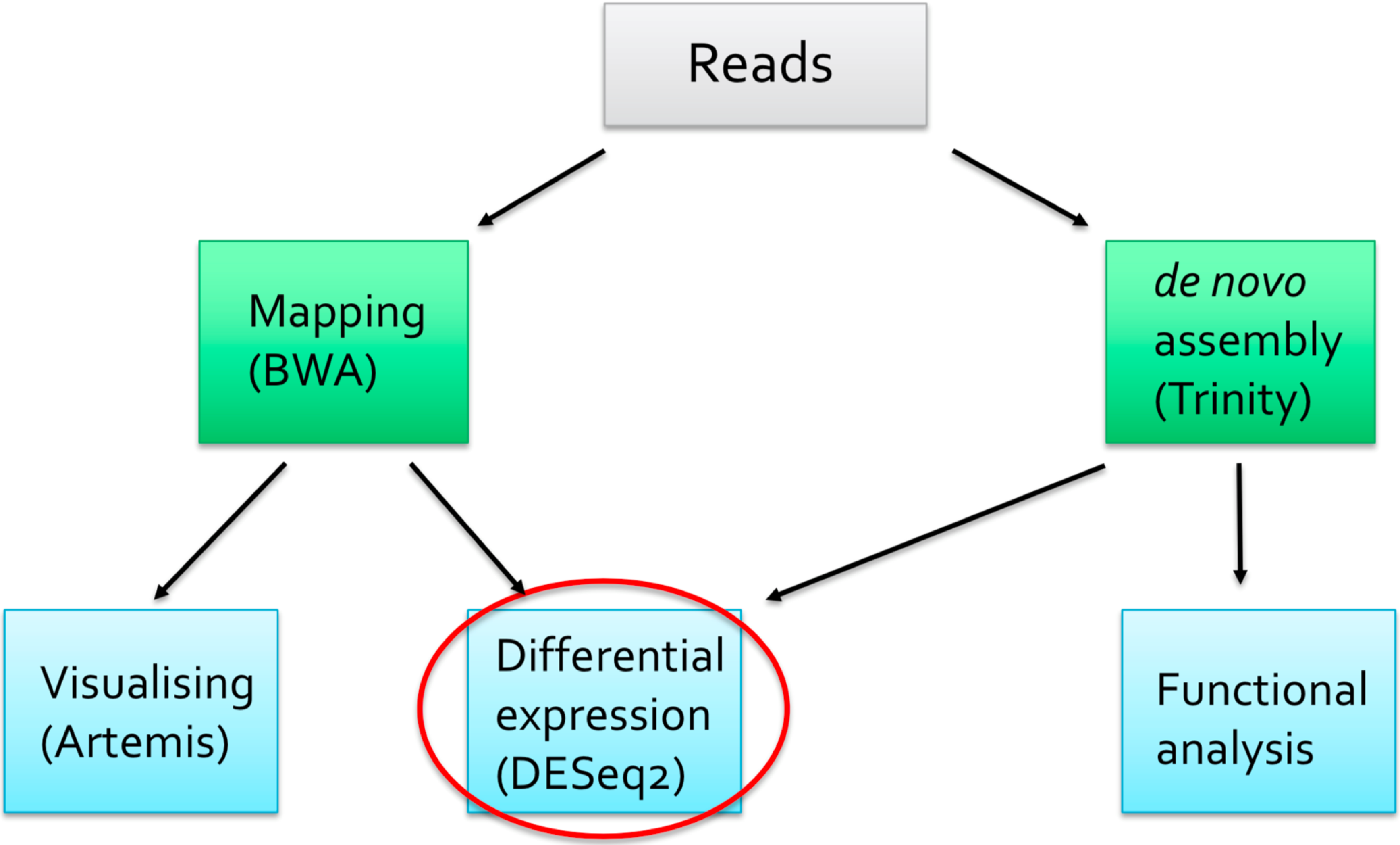


* Different aligners: BWA, HISAT2

# What are we looking for?

- Regions with high read-count
- Incorrectly annotated exons
- Alternative splicing/isoforms
- Differentially expressed genes
  - Normalisation required to account for differences due to library size (input cDNA) etc.

# RNA-Seq analysis pipeline



Reads

Mapping (BWA)

*de novo* assembly (Trinity)

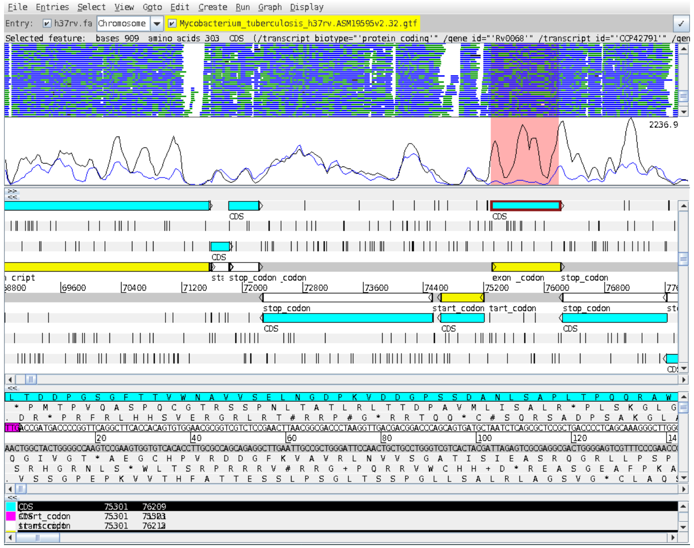
Visualising (Artemis)

Differential expression (DESeq2)

Functional analysis

# Differential Expression

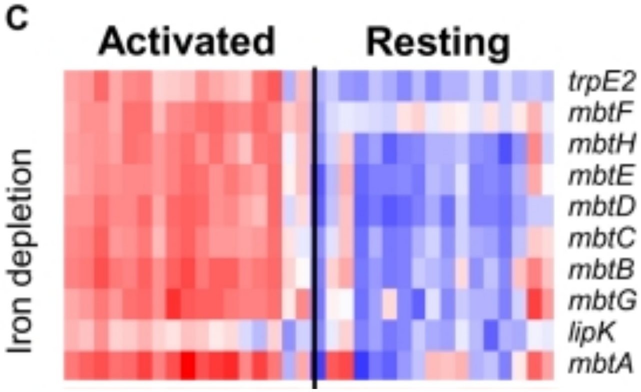
- DESeq/EdgeR
- Based on read counts assigned to transcripts
- Normalisation of gene expression values among samples (FPKM, TMM, DESeq…)
- Comparison of gene abundance under different conditions



Entry: ☑ h37rv.fa Chromosome ☑ Mycobacterium_tuberculosis_h37rv.ASM19595v2.32.gtf

Selected feature: bases 909 amino acids 303 CDS (/transcript biotype="'protein coding'" /gene id="'Rv0068'" /transcript id="'CCP42791'" /gen

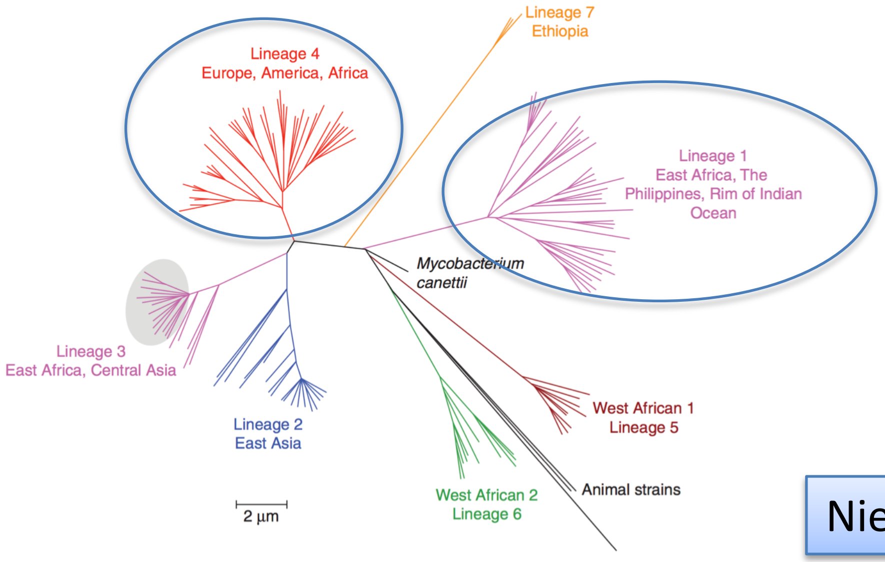
| | | |
|---|---|---|
| CDS | 75301 | 76209 |
| start_codon | 75301 | 75303 |



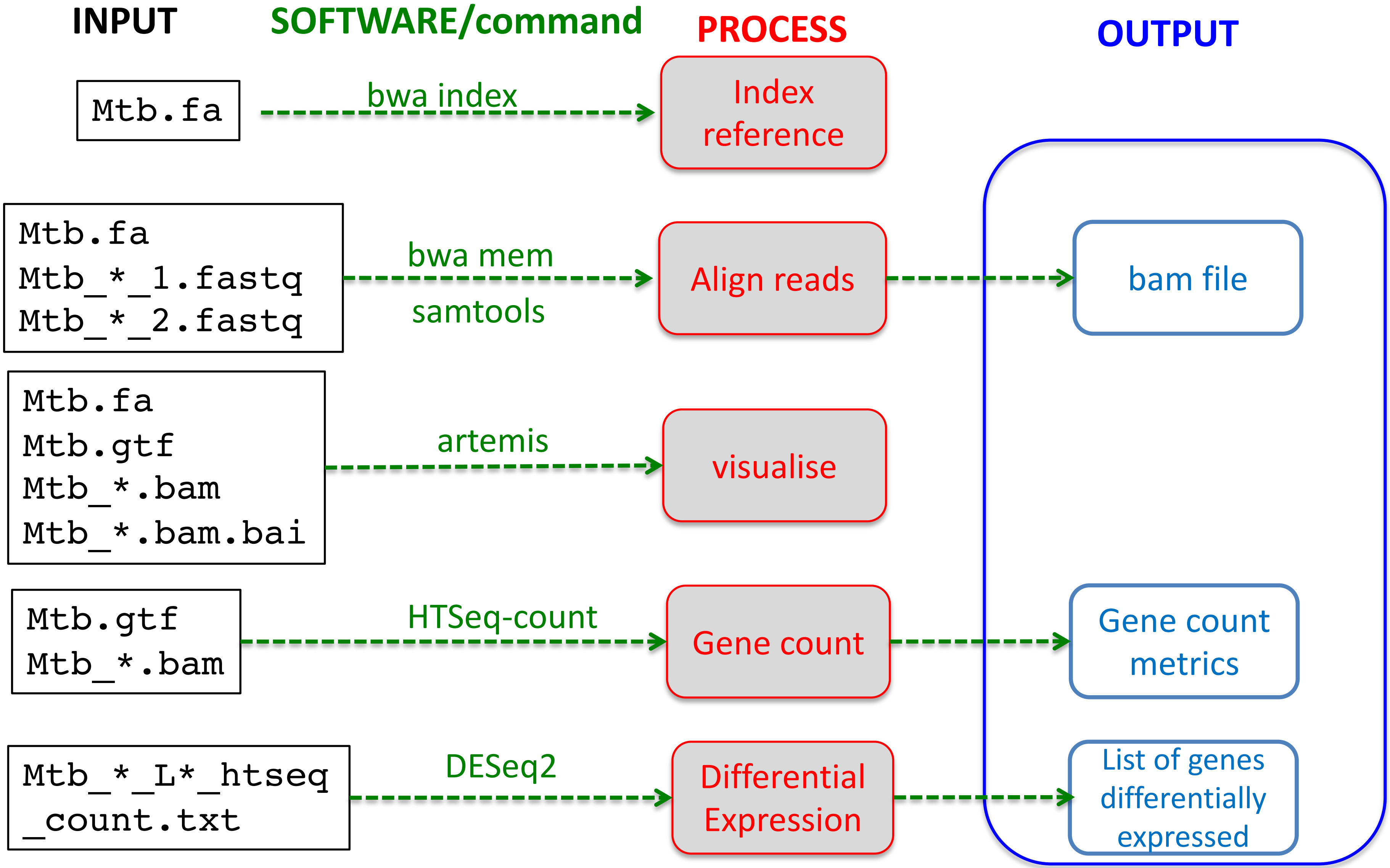
Homolka et al., 2010

# The practical

- Aligning RNA-seq reads using BWA
  - We will use data from two *Mycobacterium tuberculosis* lineages (lineage 1 and 4)
  - Align reads to H37Rv reference genome of *M. tuberculosis*
  - Visualise in Artemis
  - Count reads with HTSeq-count
  - Differential expression using DESeq2 R package



Niemann & Supply (2014)



INPUT

SOFTWARE/command

PROCESS

OUTPUT

Mtb.fa

bwa index

Index reference

Mtb.fa
Mtb_*_1.fastq
Mtb_*_2.fastq

bwa mem

samtools

Align reads

bam file

Mtb.fa
Mtb.gtf
Mtb_*.bam
Mtb_*.bam.bai

artemis

visualise

Mtb.gtf
Mtb_*.bam

HTSeq-count

Gene count

Gene count metrics

Mtb_*_L*_htseq_count.txt

DESeq2

Differential Expression

List of genes differentially expressed

* is used to represent the different lineages. Eg. Mtb_L1_1.fastq

Acknowledgements: Paula Josefina Gomez Gonzalez