Introduction to Transcriptomics

London School of Hygiene and Tropical Medicine

Matthew.Higgins@lshtm.ac.uk

Outline

This practical has 3 core objectives:

- Introduce transcriptomics
- Outline RNA-Seq
- Demonstrate utility of differential gene expression

What is Transcriptomics?

Transcriptomics is the investigation of all **RNA** molecules within a sample.

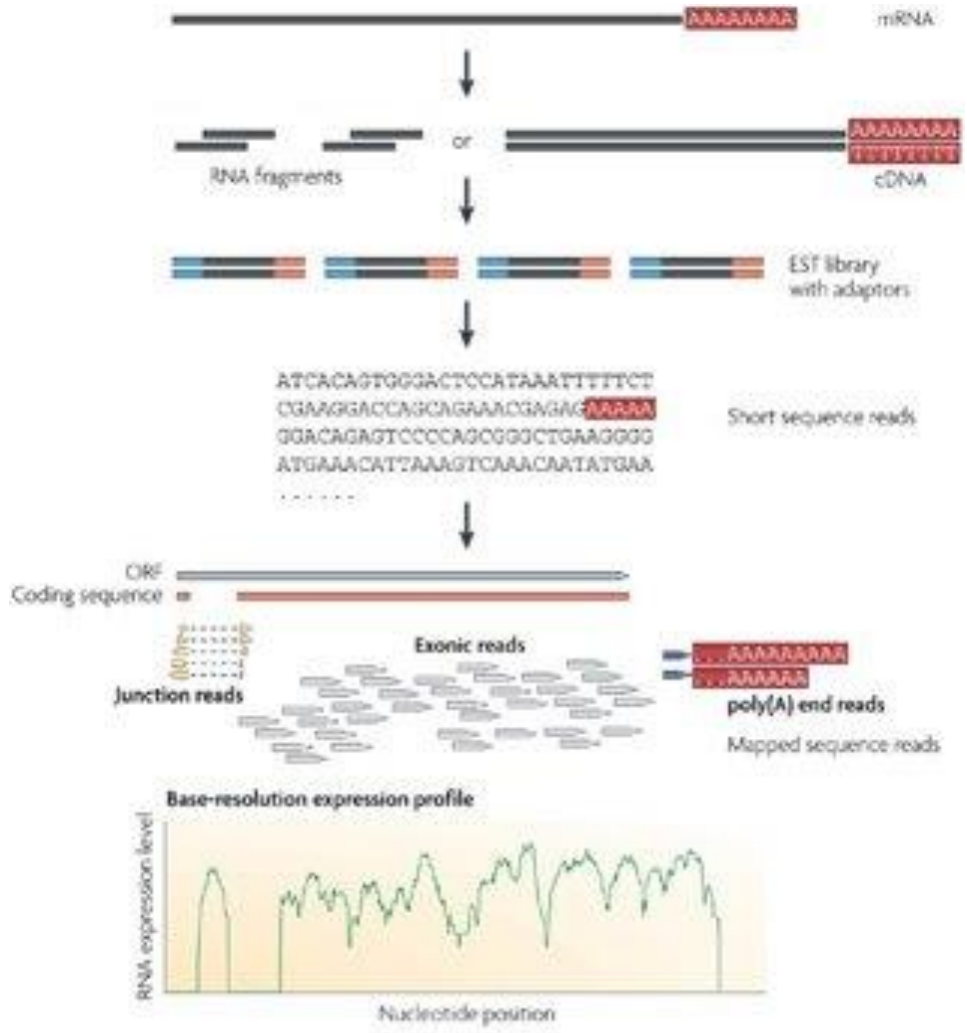
It can be used to:

- Measure transcript abundance.
- Investigate differences in gene expression.
- Identify novel transcript isoforms (splice variants).

What is RNA-Seq?

RNA-Seq is the most common technique used in transcriptomics. The stages are as follows:

1) RNA Isolation.
2) cDNA synthesis.
3) Library preparation (according to sequencing platform).
4) Sequencing & data analysis.



Comparison to Historic Approaches

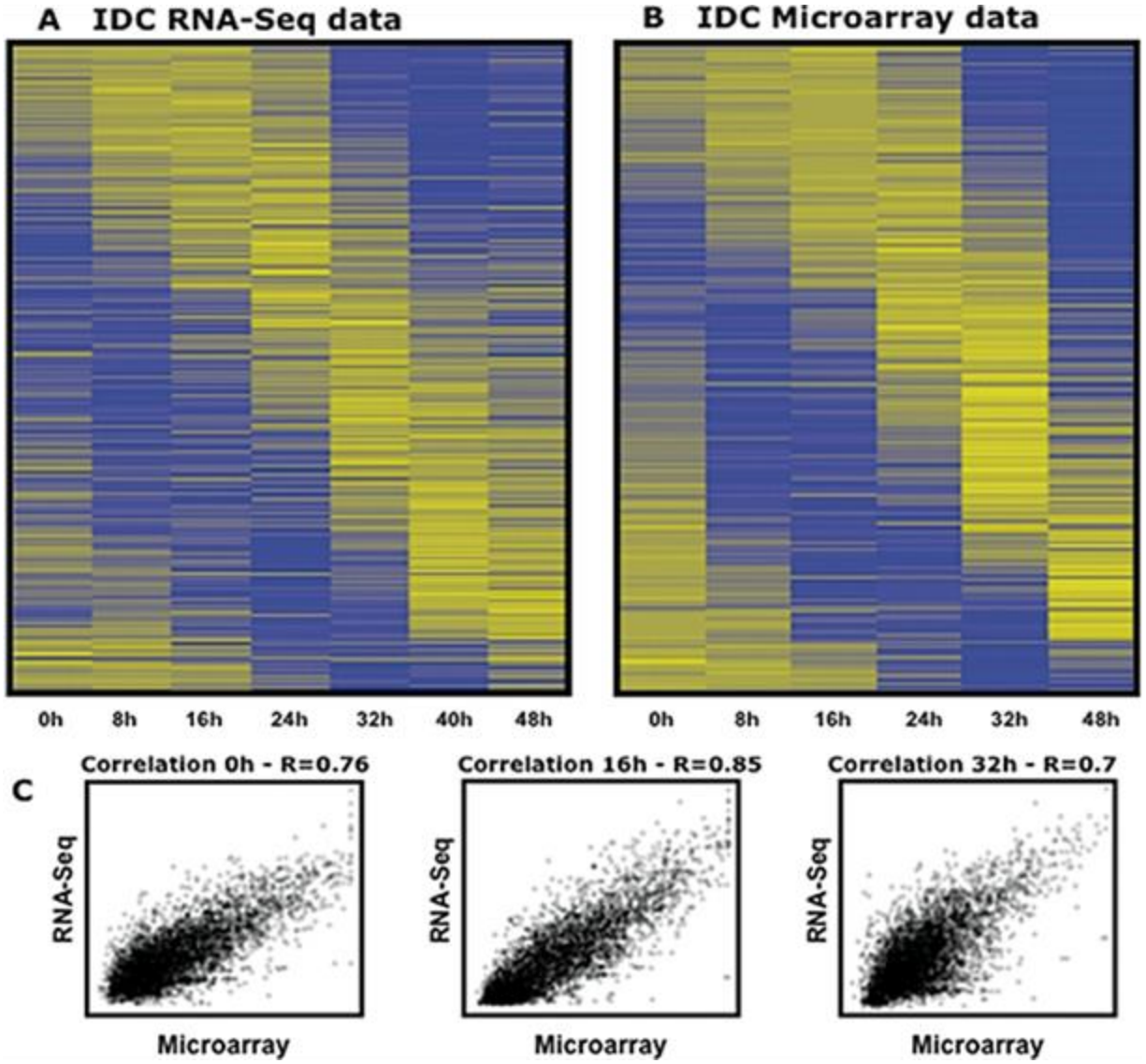
Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background 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. *Nature Reviews Genetics* 10, 57-63

RNA-Seq vs Microarray

Numerous benchmarking studies, including a 2010 study investigating ***P. falciparum*** development***:***

- 75% of previously predicted splice sites, confirmed.
- 202 novel spliced sites.
- 107 novel transcripts.



Experimental Design

Total vs mRNA Only

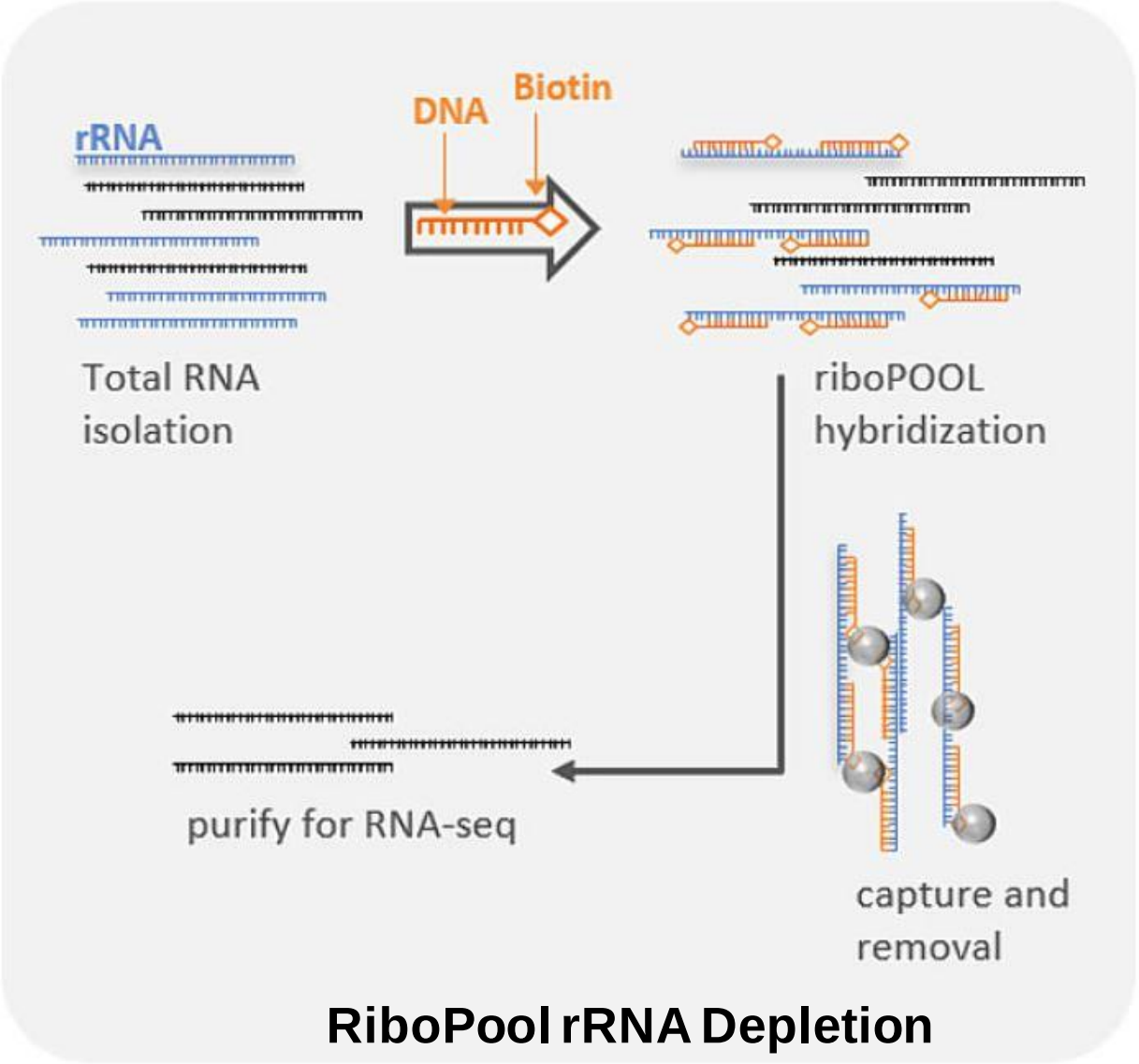
- Depletion of rRNA

Sequencing platform

- Illumina vs Nanopore

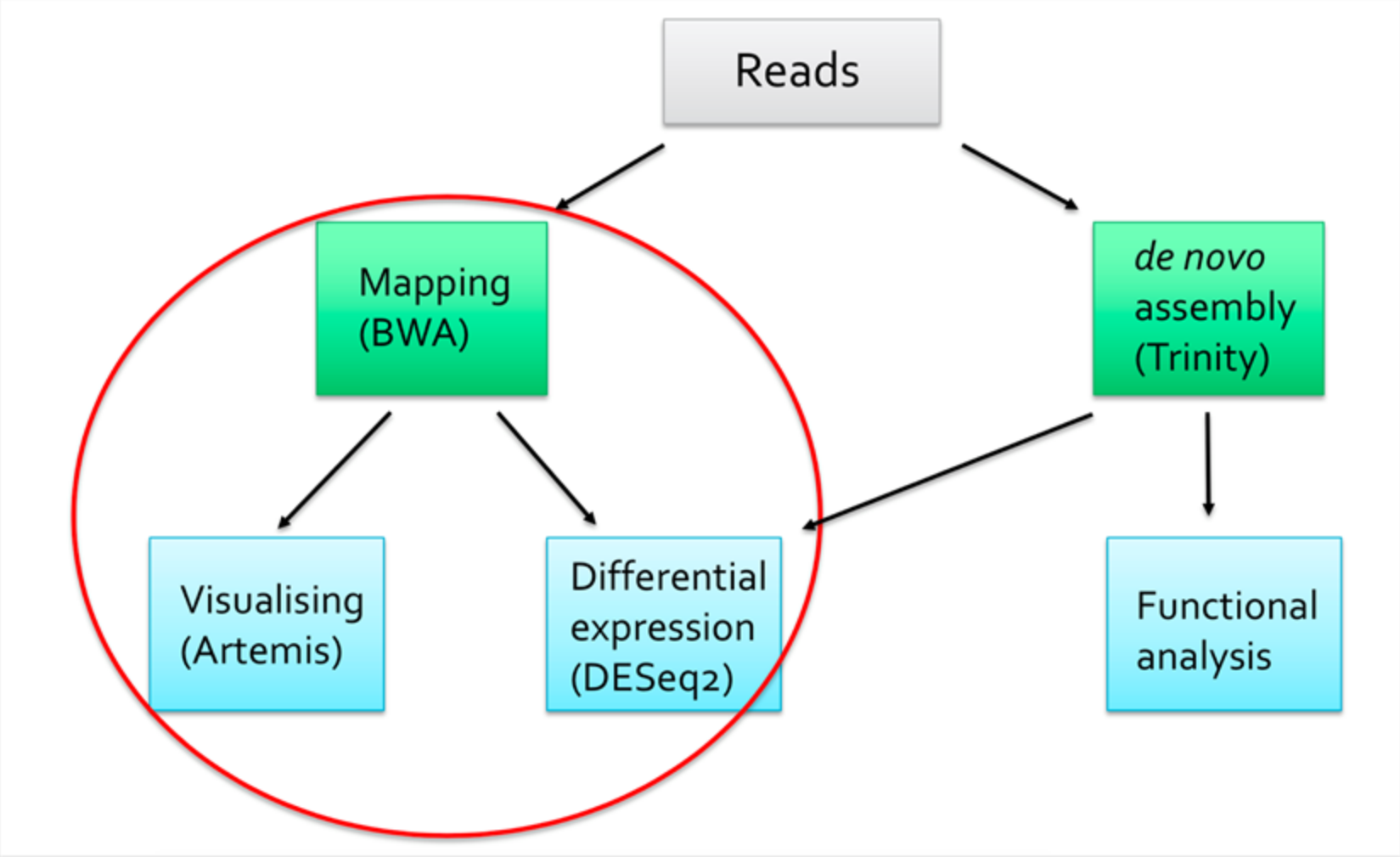
Multiplexing

- Size of transcriptome
- Library yield



RiboPool rRNA Depletion

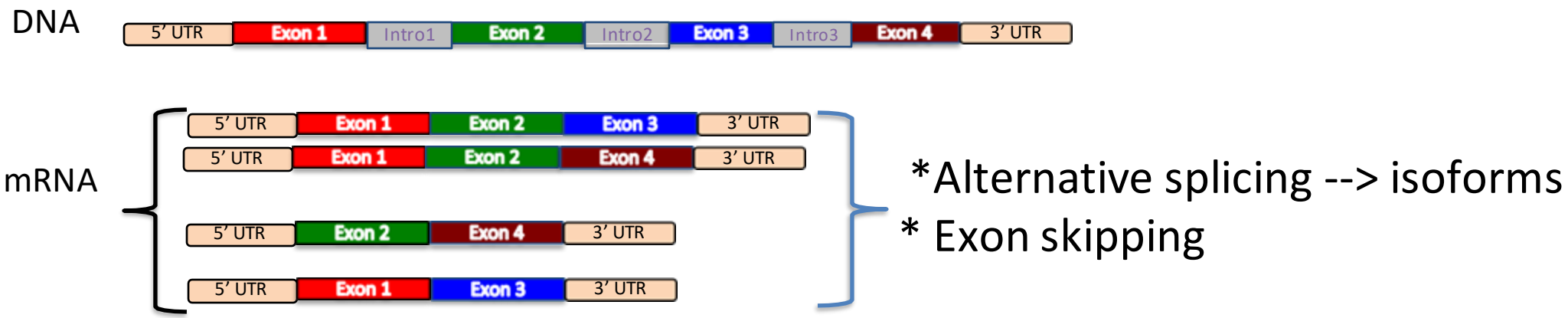
RNA-Seq Analysis Practical Pipeline



Analysis Workflow

- Data quality control (trimming & filtering).
- Map to reference genome or transcriptome.
- Filter mapped reads and quantify transcript abundance.
    - Discard poor quality reads.
    - Discard non-uniquely aligned reads.
- Investigate differential gene expression between sample cohorts.

Mapping Considerations



* Different aligners: BWA, HISAT2

Post Mapping Considerations

What are we looking for within a sample:

- Quantify gene expression based on number of reads which map uniquely to given transcript.
- Novel transcripts & exons.

What are we looking for between samples:

- Differential gene expression.

Important: Normalization required to account for differences due to library size etc.

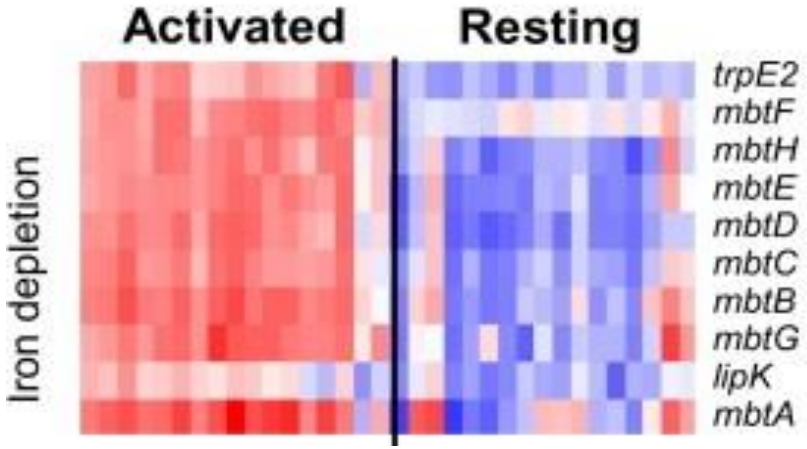
Differential Gene Expression

Why consider differential gene expression?

- Probe samples with different phenotypes (e.g. susceptible vs drug resistance)?

- Uncover role of genes (e.g. during development cycle)

Bioinformatic tools: DESeq2 & EdgeR



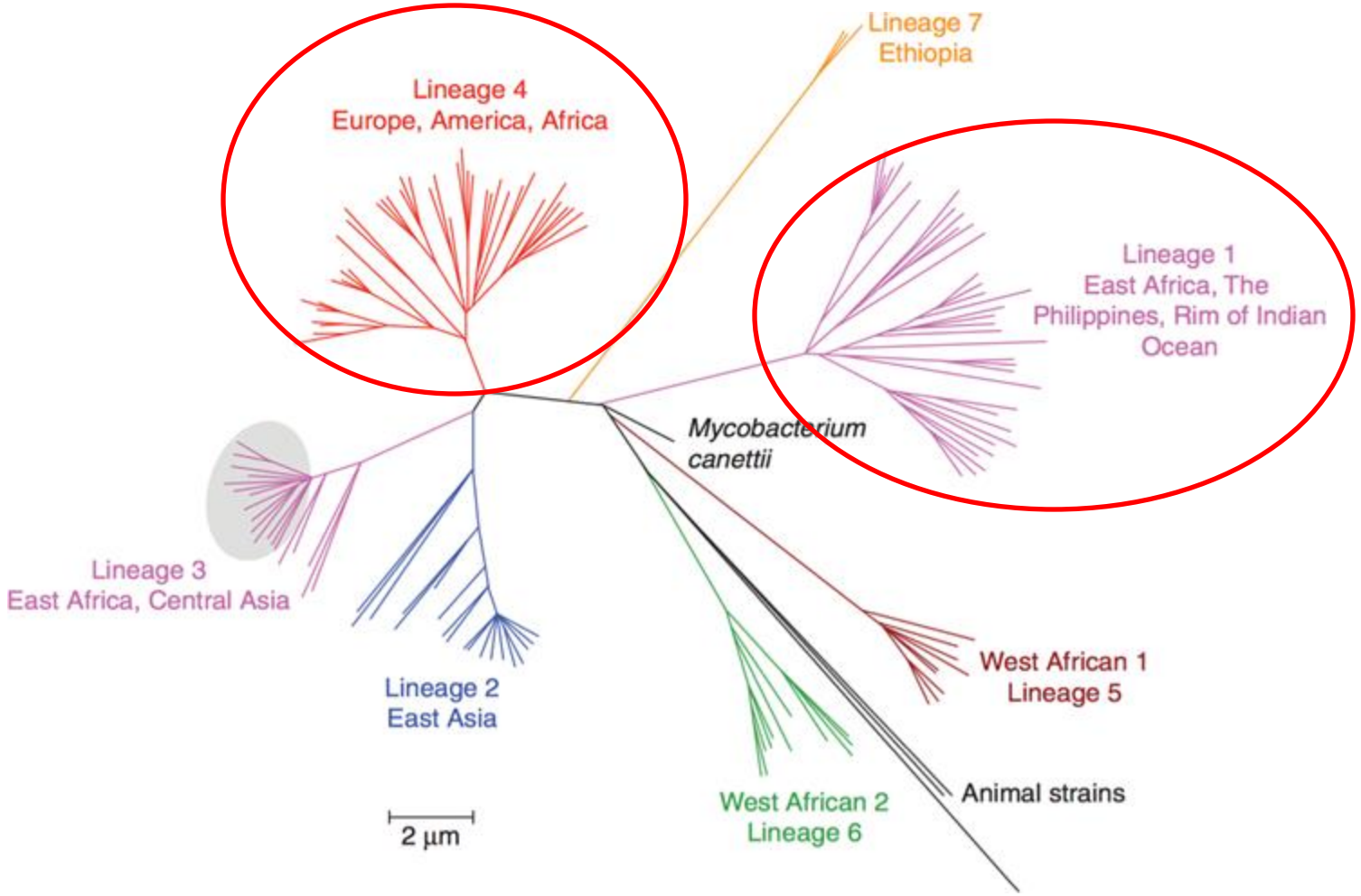
Heatmap demonstrating difference in gene expression between active vs resting TB

Homolka et al., 2010

Practical Background

- Investigating the transcriptome of *Mycobacterium tuberculosis* (TB).

- Comparing gene expression between lineage 1 vs 4.



Niemann et al 2014