Oxford Genomics CENTRE

UNIVERSITY OF OXFORD

RNA-Seq Data Analysis
27-28 November, 2017

Taught module for DPhil programme in
Genomic Medicine and Statistics

Organised by Bioinformatics Core at WHG:

Helen Lockstone M.Sc.

Ben Wright PhD

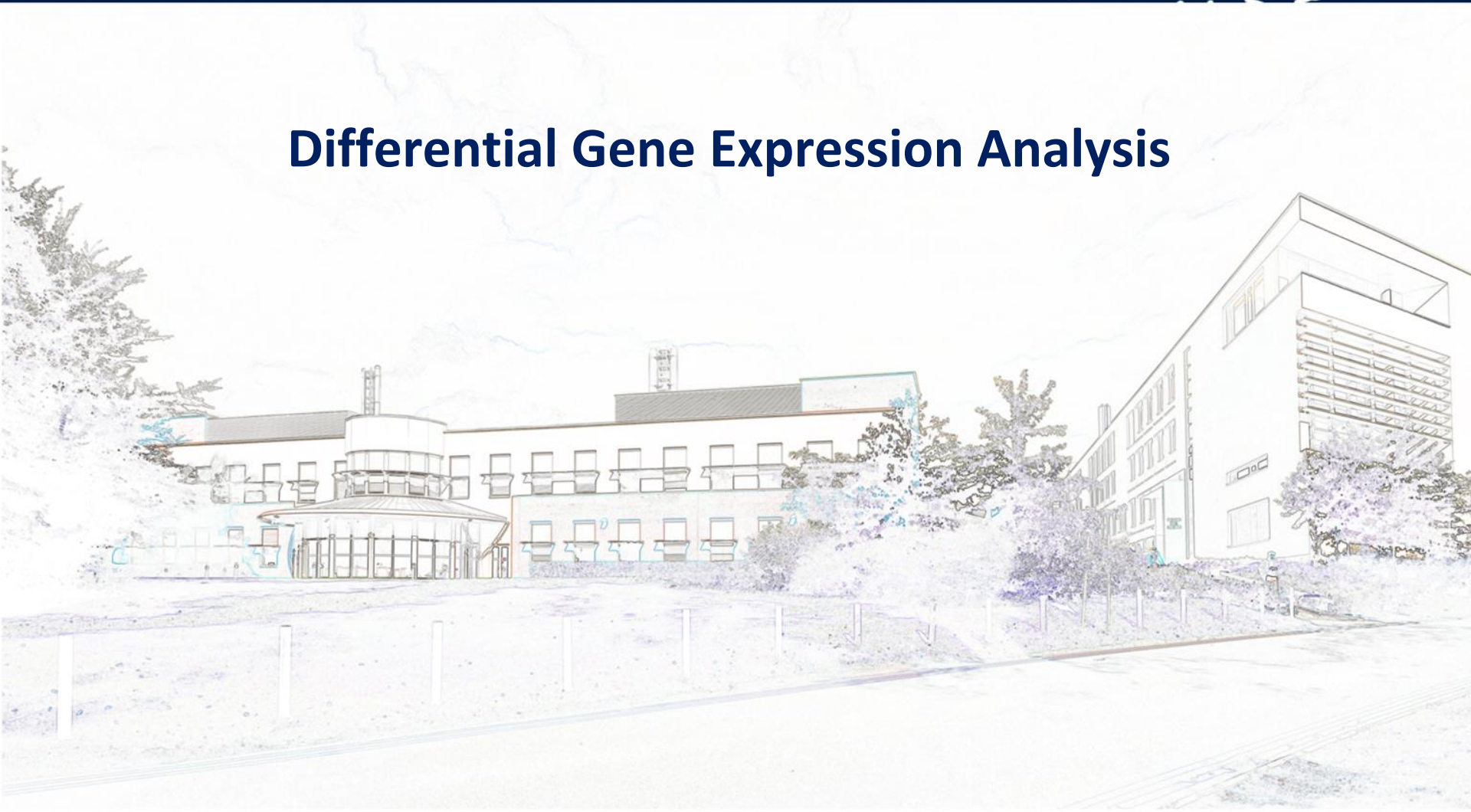
Santiago Revale, M.Sc.

wellcome centre human genetics

Oxford Genomics Centre

Oxford Genomics Centre
The Wellcome Trust Centre for Human Genetics

Differential Gene Expression Analysis



Helen Lockstone
Bioinformatics Core Group

whg

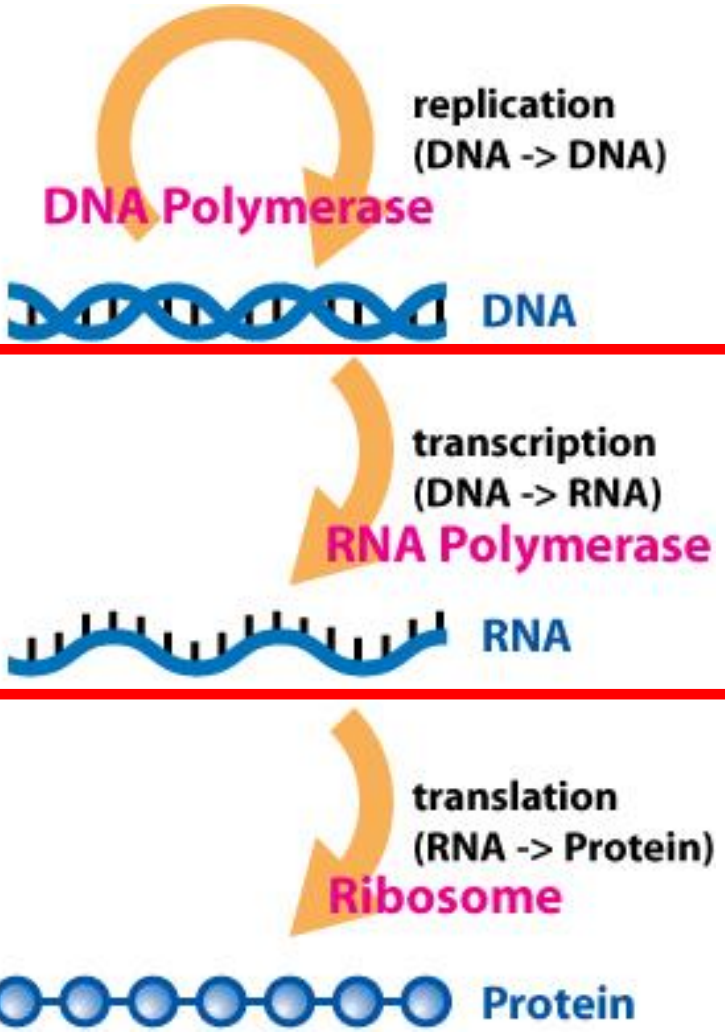
UNIVERSITY OF OXFORD

Overview

- General remarks on gene expression technology and data (RNA-Seq and microarray)
- Experimental design considerations
- Hypothesis testing
- Differential expression analysis using R/BioConductor
- Practical session working with a cancer dataset

A brief history of gene expression

Transcriptome Profiling



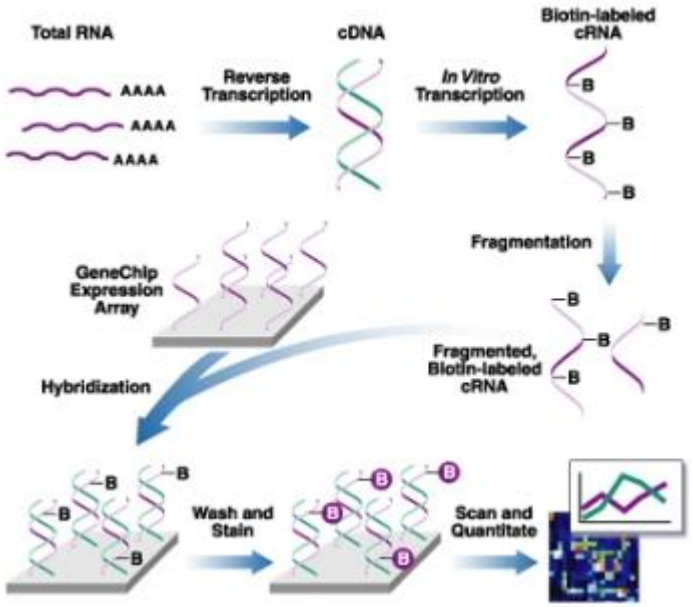
Transcriptome can be measured by microarrays or RNA-Seq

Widely-used techniques, provide insight into biological system, albeit a snapshot – highly dynamic and complex process (splicing, gene methylation, RNA stability/degradation, miRNA regulation etc)

Two key technologies

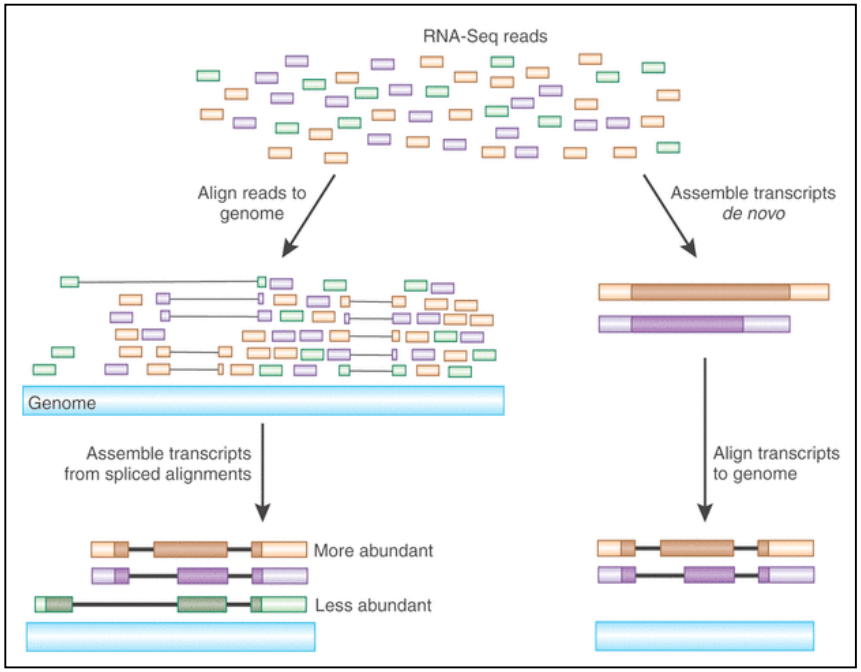
Oxford Genomics CENTRE

Microarrays



Complementary hybridisation
early 1990s onwards

RNA-Seq

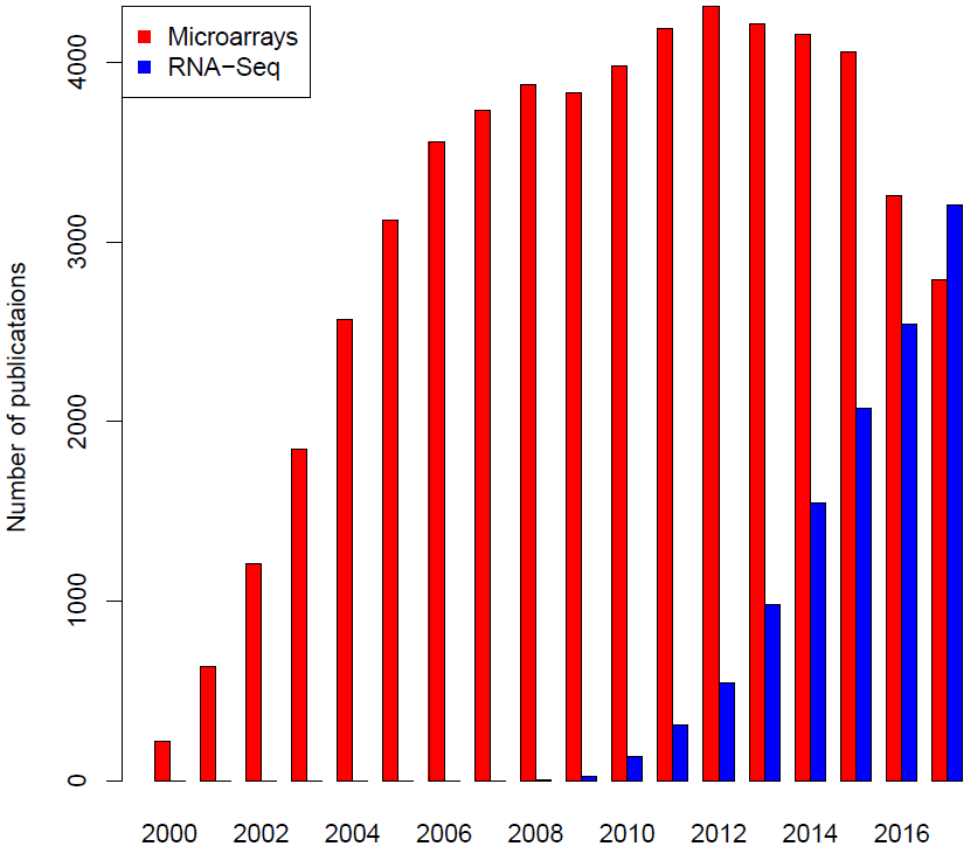


Next-generation sequencing
2007 onwards

wellcome centre human genetics

UNIVERSITY OF OXFORD

Publications by Technology



Which technology to use?

- Microarrays and RNA-Seq are complementary technologies (despite common perception that RNA-Seq superior)

- Choice usually depends how detailed a characterisation of the transcriptome is required
  - Gene level changes => microarrays sufficient, reliable and cheap
  - Isoform structure, splicing, novel transcripts => RNA-Seq
  - Note that exon arrays can also assess splicing

- Both report relative gene expression level estimates, influenced by a range of factors and biases inherent to each technology
- Fold-change concordance reasonably high between arrays and RNA-Seq

RNA-Seq Myths and Caveats

- Can detect low expressed genes better than arrays
  - Possibly but may need prohibitively expensive sequencing depth
  - In typical designs, up to half of all genes are too low expressed to be reliably detected (if at all)
  - Additional sequencing will still tend to be of highly expressed genes, so lower end hard to interrogate

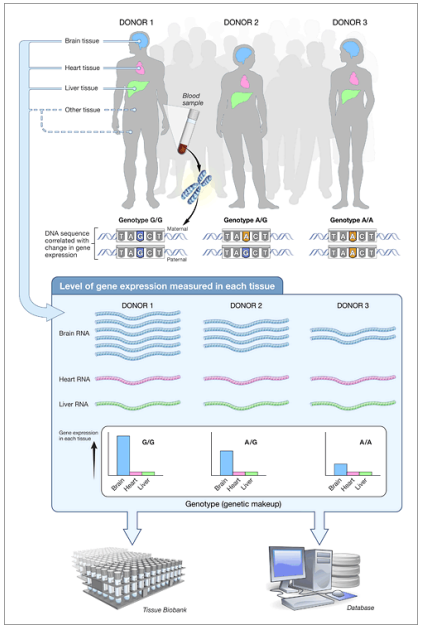
- The issue of low counts is even more problematic for splicing analysis where you may be comparing exons or junction-spanning reads

- What you sequence in an RNA-Seq library influences your data for all genes – very inter-dependent in a way that arrays are not

Large-scale gene expression projects

- ENCODE
- Allen brain atlas
- Genotype-Tissue Expression Project (GTEx)
- TGCA

- Public repositories
  - Gene Expression Omnibus (GEO)
    http://www.ncbi.nlm.nih.gov/geo/
  - Sequence Read Archive (SRA)
  - http://www.ncbi.nlm.nih.gov/sra



Typical experimental designs

- Disease vs control
- Gene knockdown/knockout vs wildtype
- Effect of treatment/stimulus/drug

- Clinical applications
  - Tumour-normal pairs
  - Good prognosis vs poor prognosis
  - Patient subgroups responding to different treatments
  - ‘Gene signature’ to predict who will respond well to a given treatment
- Time course
- Different tissues/stages of development

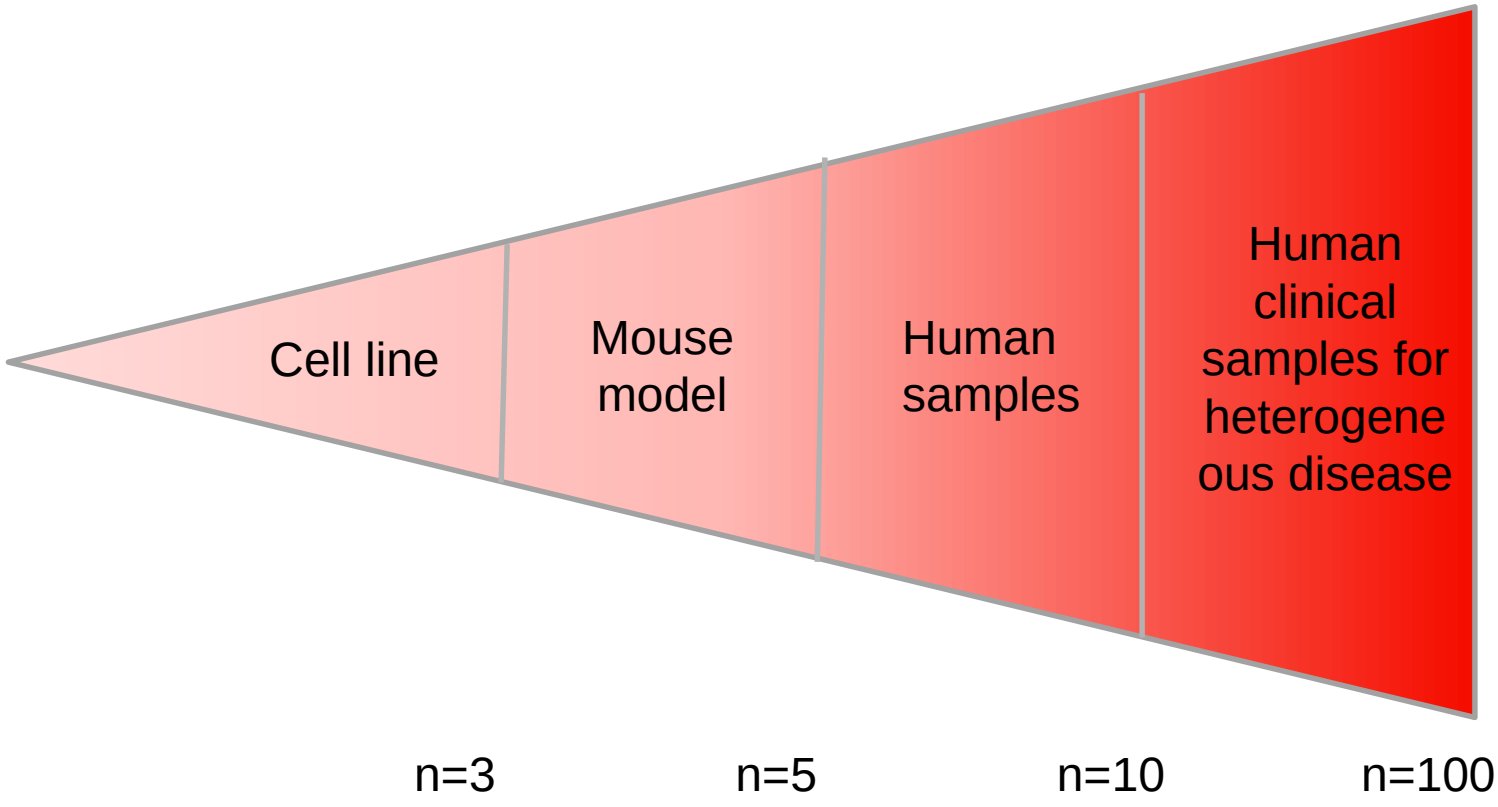
Limitations of transcriptomic profiling

- Comprehensive but inherently limited to descriptive results, no matter how well experiment performed or data analysed
- Produce large amounts of information; subjective interpretation, can be mined in different ways, always much left untouched (often publically available)
- Expensive and time-consuming so often published as a stand-alone experiment
- However best used as starting point for further work - following up hypotheses from gene expression data to uncover mechanistic/causal effects can produce elegant studies

Experimental design considerations

Replication

- Depends on context – type of sample, size of effect, heterogeneity within conditions



Sequencing Depth

- Number of reads required per sample depends on experimental question

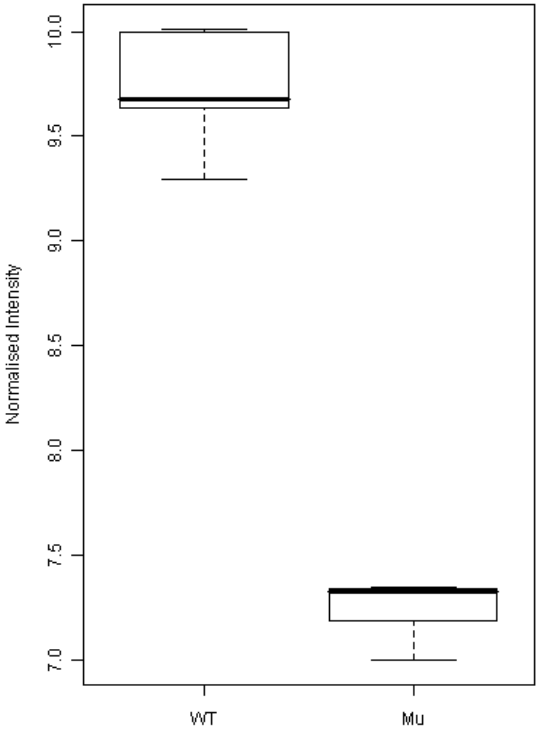
- HiSeq4000 – one lane = 250 million reads
- Multiplexing e.g. 10-plex human samples gives ~25m reads for each, plenty for quantifying gene expression (except for very low/unexpressed genes)

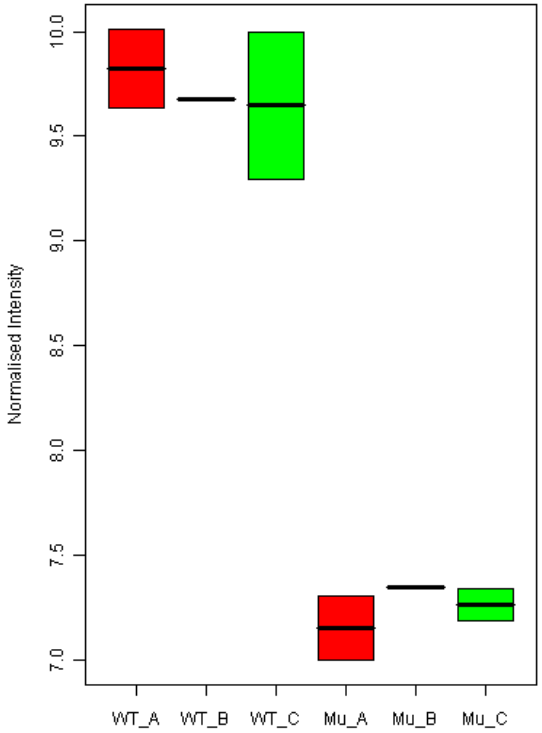
- Higher depth required in some situations e.g. for splicing analysis, certain library prep methods (ribo-depletion)

Potential confounds and covariates

- Gene expression data highly sensitive to many factors
  - Lab operator/conditions, day performed, sample collection methods, RNA extraction day and so on
  - Often influence the data to greater extent than any experimental effects
  - Any step where treated and control samples are handled differently could confound the experiment
  - If split into batches containing mix of treated/control samples, can account for potential effects in analysis
- Also be aware of potential effects from factors unrelated to the experiment on the data, which may need to be accounted for to optimise analysis

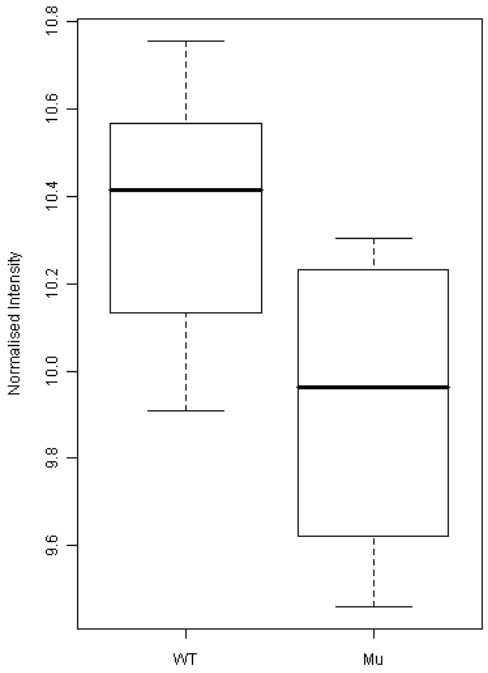
Gene not influenced by litter

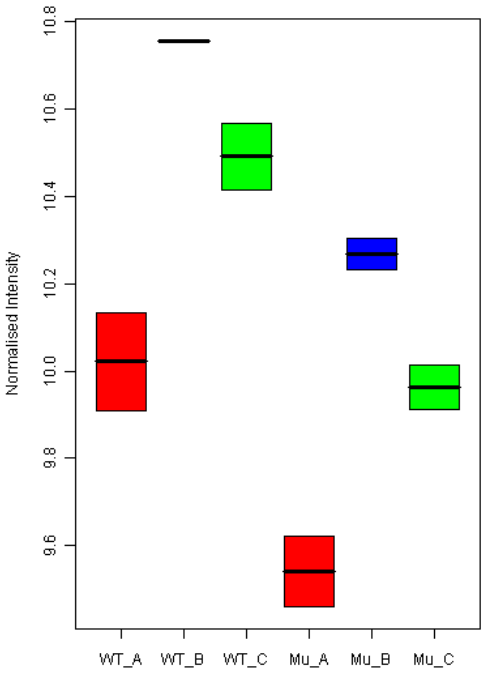




- Wt and Mut groups
- Three different litters
- Top gene ~ 5x higher expression in Wt compared to Mut
- Similarly expressed across litters in both genotypes

Gene with strong litter effect





- Within litters, consistent pattern of higher expression in WT vs Mut
- Within genotypes, B>C>A – expression depends on litter
- Accounting for this source of variability increases power to detect changes of interest