Concepts and Main Aspects of RNA-Seq

Wellcome Trust Centre for Human Genetics

25th, 28th and 29th April 2016

Organised by Helen Lockstone and Irina Pulyakhina

Overview of Gene Expression Profiling and Experimental Design

Helen Lockstone

25th April 2016

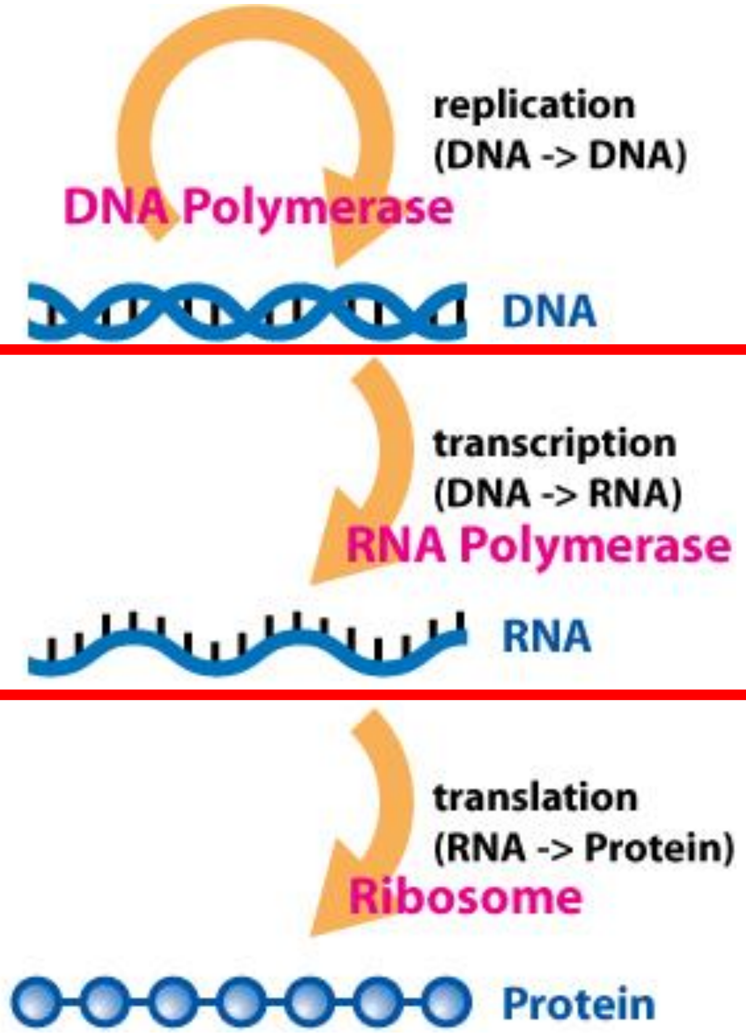
Day 2 Overview

- General remarks on gene expression data
- RNA-Seq and microarray technology
- Experimental design considerations
- Transcript quantification
- Data normalisation and quality
- Differential expression analysis
- Biological interpretation using pathway analysis

Schedule

Time	Topic
09.45-10.30	RNA-Seq experimental design
10.30-11.00	Transcript quantification, normalisation
11.00-11.15	Coffee break
11.15-12.00	Differential expression analysis
12.00-12.30	Pathway analysis
12.30-13.30	Lunch
13.30-16.00	Practical sessions (break at 15.00)

Transcriptome profiling



Entire 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)

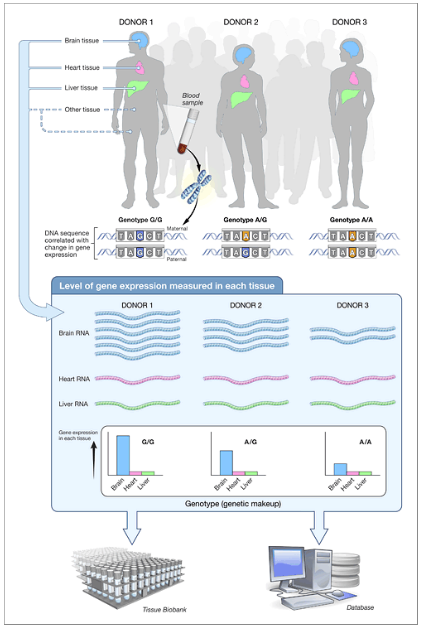
Examples of large gene expression projects

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

ALLEN BRAIN ATLAS
DATA PORTAL

THE CANCER GENOME ATLAS

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

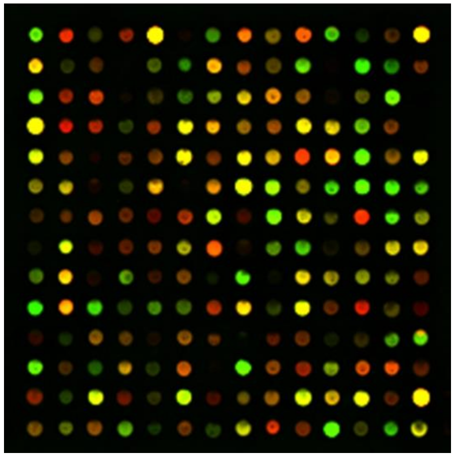


Typical expression profiling designs

- 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

Premise of gene expression profiling



- Compare gene expression in different conditions
- Differentially expressed genes may provide some biological insight
- But not magical solutions! Large amounts of descriptive data generated – what to do next?

Limitations of gene expression data

- Comprehensive but inherently limited to descriptive results, no matter how well experiment performed or data analysed

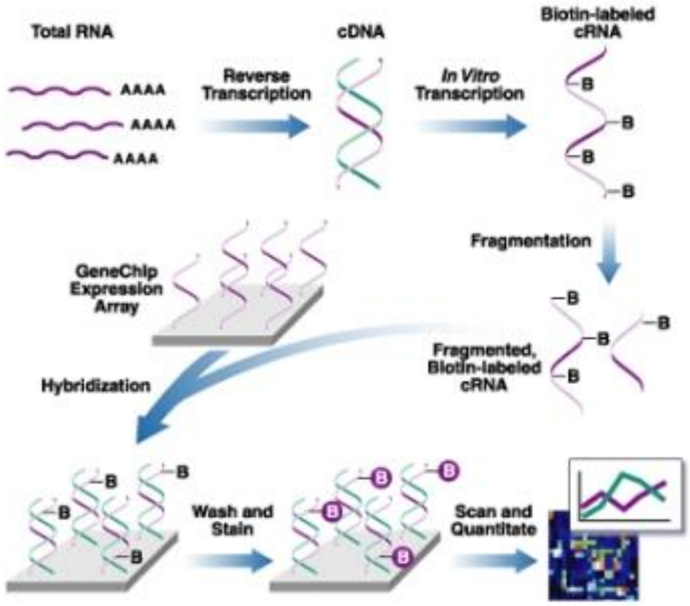
- Produce large amounts of information; subjective interpretation, can be mined in different ways, always much left untouched (often publically available)

- 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

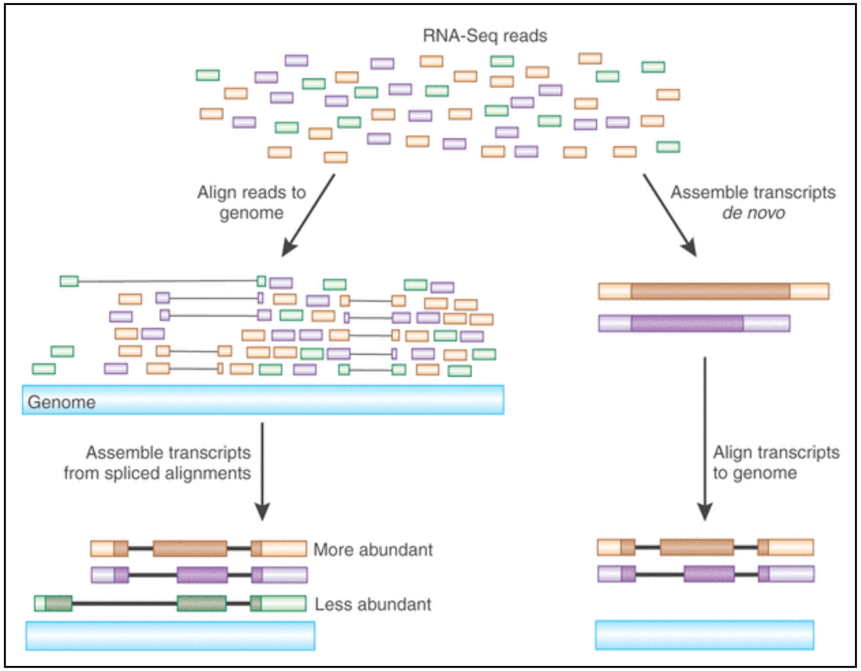
Two key technologies

Microarrays



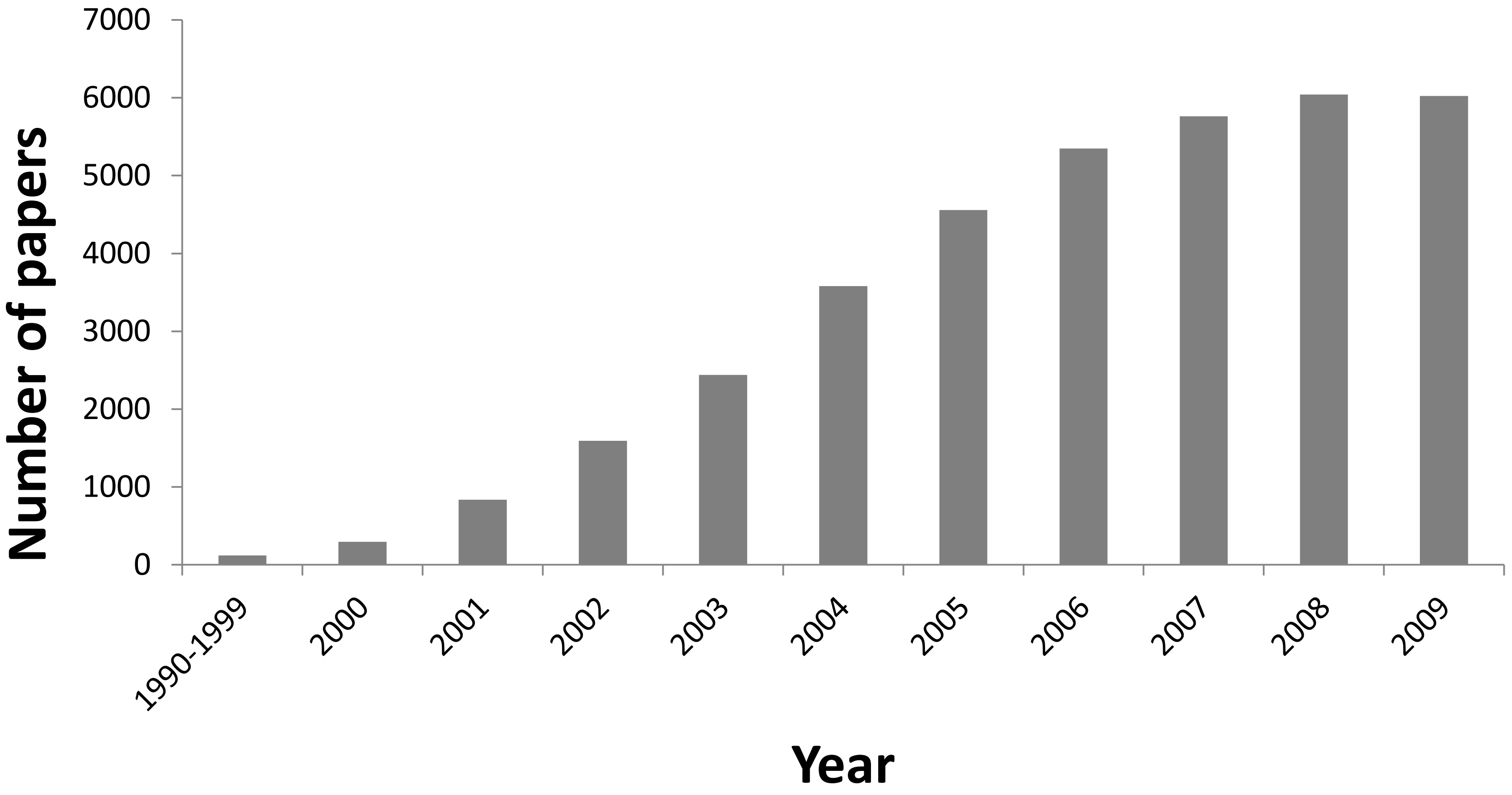
Complementary hybridisation
early 1990s onwards

RNA-Seq


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Next-generation sequencing
2007 onwards

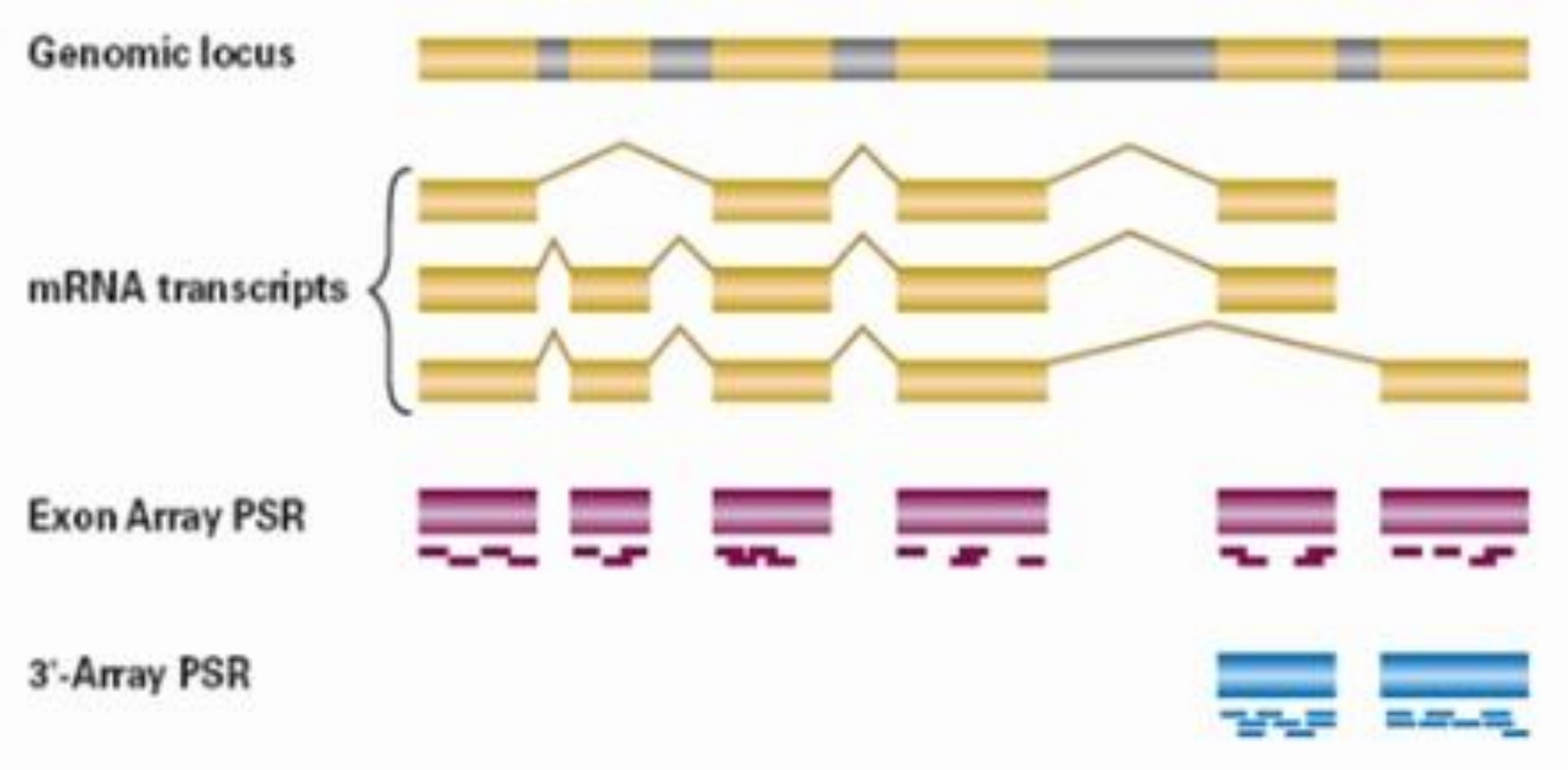
Microarrays in the Literature



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
  - Isoform structure, splicing, novel transcripts => RNA-Seq
- Many low expressed genes in a given sample type in both technologies

Exon Array Design



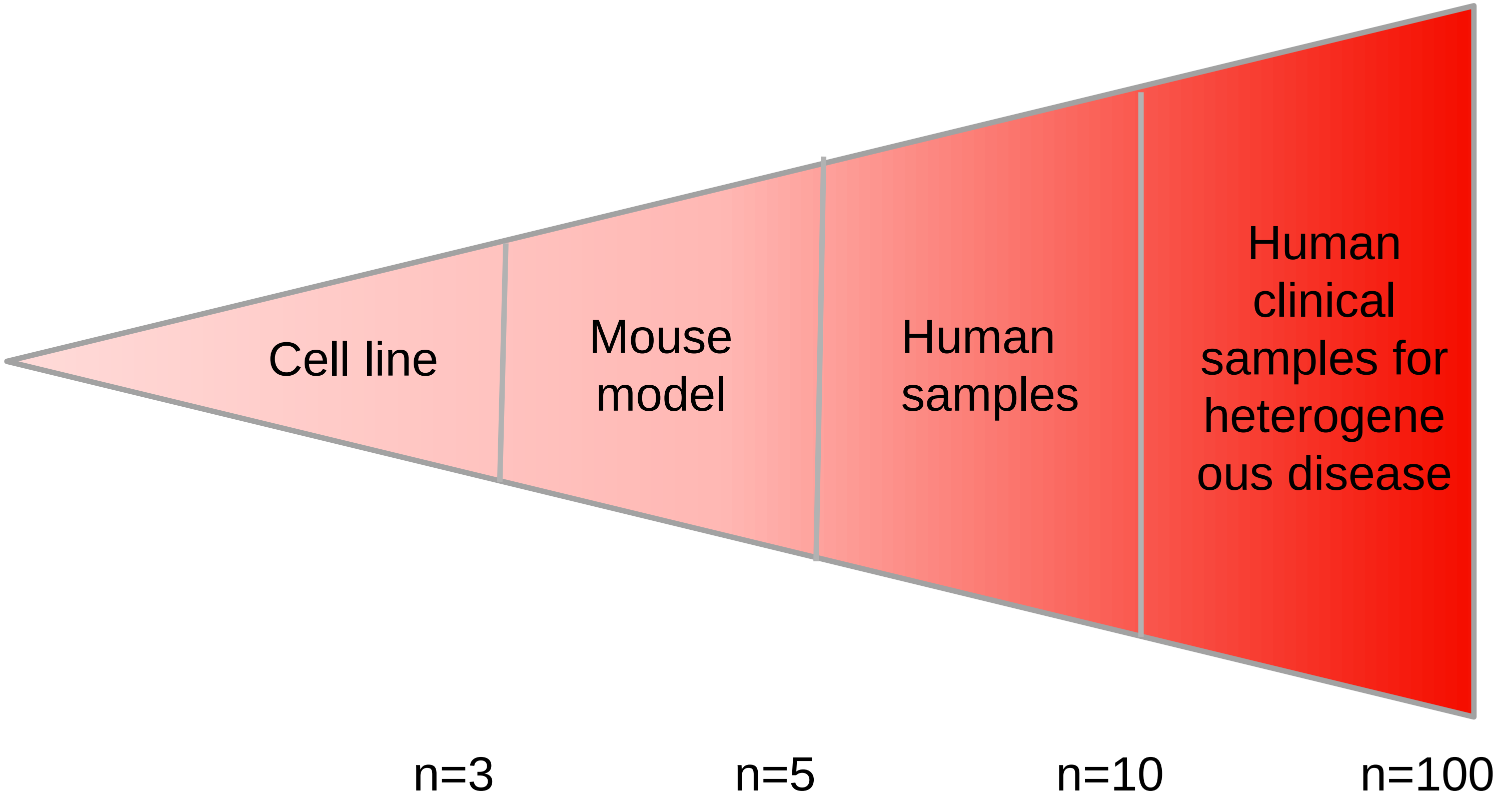
Picture from Affymetrix

Experimental Considerations

- Number of replicates
- Sequencing depth (number of reads per sample)
- Good experimental design principles

Replicates

- 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)

Single or Paired-end?

- Increased mapped reads with PE data – possibly one read maps to non-unique regions, while second read helps anchor to a specific gene/location

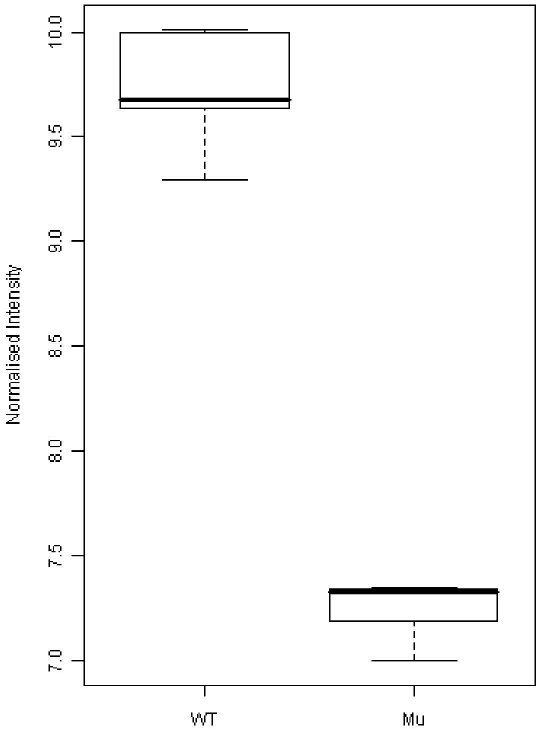
Other experimental considerations

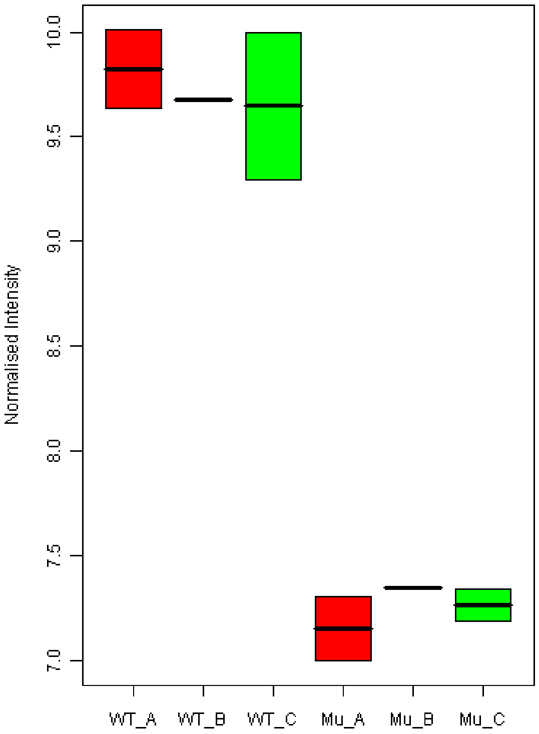
- Sufficient replication/depth for purpose

- Avoid confounding factors when obtaining/preparing samples – gene expression data highly sensitive to many factors

- Be aware of potential effects of unrelated factors on the data, which may need to be accounted for to optimise analysis

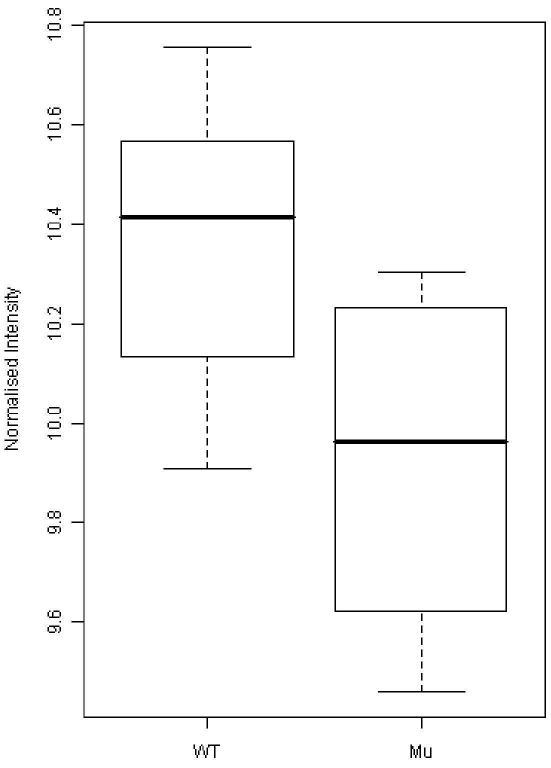
Effect of other variables

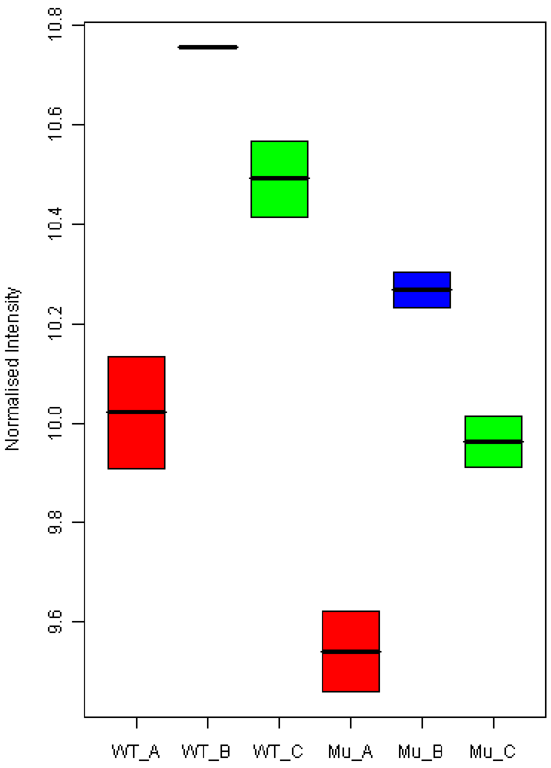




- Wt and Mut groups
- Three different litters
- Top gene ~ 5x higher expression in Wt compared to Mut
- Similarly expressed across litters in both genotypes

Strong litter effect





- Overlap between groups
- Within litters, consistent pattern of higher expression in WT vs Mut
- Within genotypes, B>C>A – expression depends on litter
- Accounting for this variance increases power