Oxford Genomics CENTRE

UNIVERSITY OF OXFORD

RNA-Seq Data Analysis
7th – 8th June 2018

Organised by

Bioinformatics Core at WHG

Helen Lockstone

Santiago Revale

Eshita Sharma

Ben Wright

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

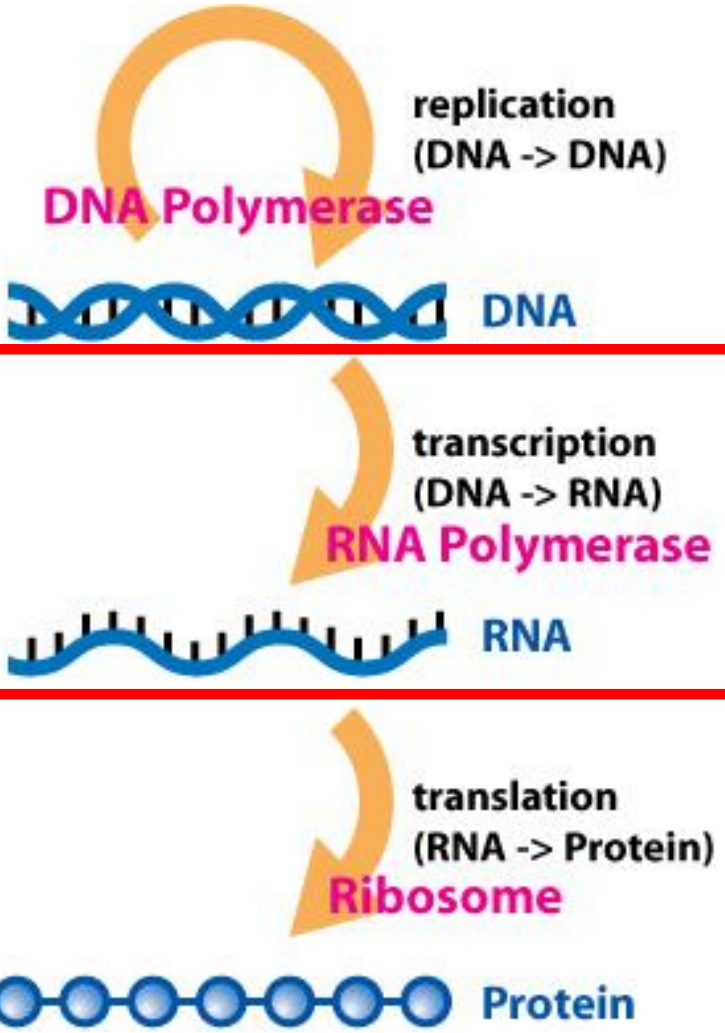
UNIVERSITY OF OXFORD

Overview

- Development of gene expression technology (RNA-Seq and microarrays) and associated methods/tools
- Experimental design considerations
- Hypothesis testing overview
- Differential expression analysis using R/BioConductor
- Practical session working with a cancer dataset

A brief history of gene expression

Transcriptome Profiling



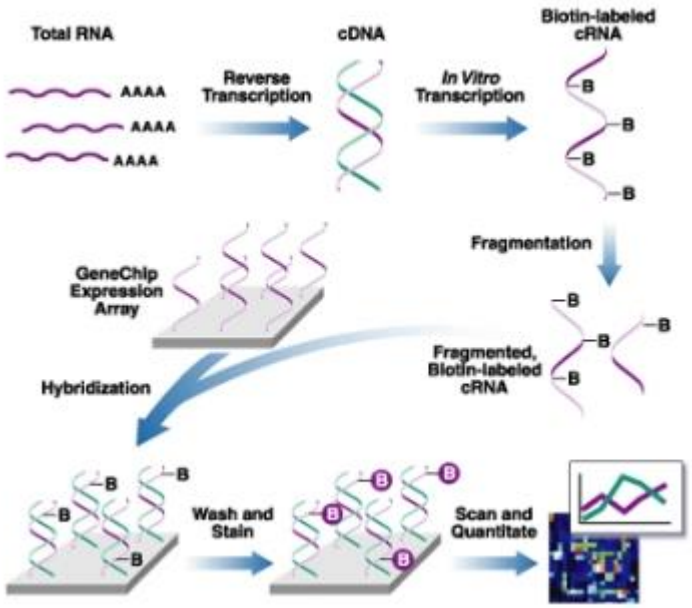
Transcriptome can be measured by microarrays or RNA-Seq

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

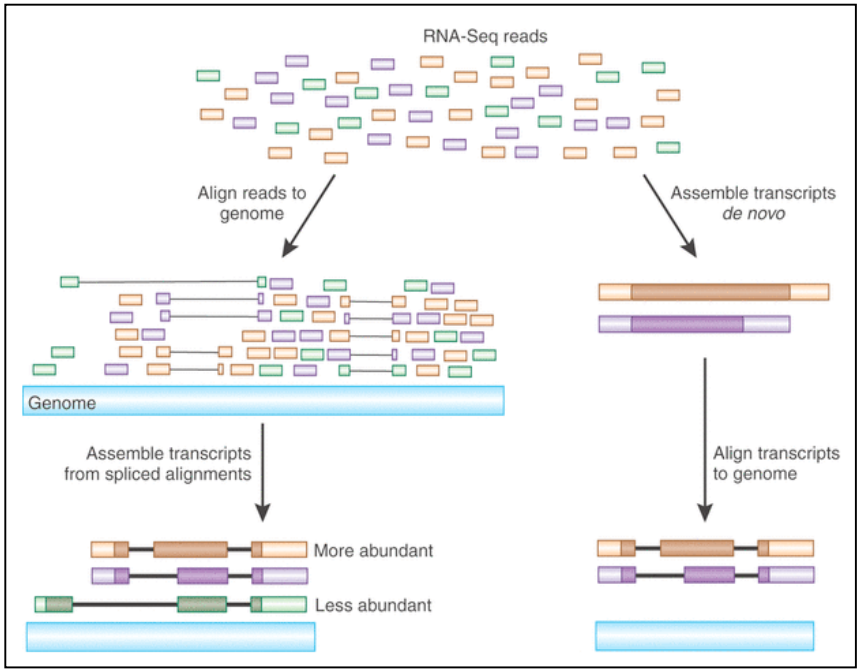
Oxford Genomics CENTRE

Microarrays



Complementary hybridisation
(early 1990s onwards)

RNA-Seq



Next-generation sequencing
(~2008 onwards)

wellcome centre human genetics

UNIVERSITY OF OXFORD

Transcriptomics Approaches

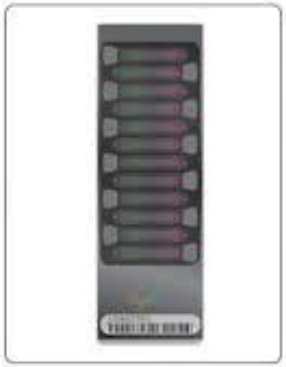
- Microarray technology (1990s onwards): based on pre-designed short oligonucleotide sequences, or probes, hybridising to complementary target sequences (genes) – generate fluorescent intensity signal (continuous data that after preprocessing can be considered approximately normally distributed).

- RNA-Seq (~2008 onwards): next generation sequencing approach. Library of cDNA fragments prepared from RNA and sequenced. Generates count data (number of reads mapping to a given gene), requiring statistical models suitable for count data (e.g. negative binomial model as implemented in edgeR or DESeq)

Key Microarray Manufacturers

Illumina

- Became main player in next generation sequencing
- Discontinued expression arrays (rat, mouse, human) over past few years but still manufacture genotyping arrays


HumanHT-12


HumanRef-8 and HumanWG-6

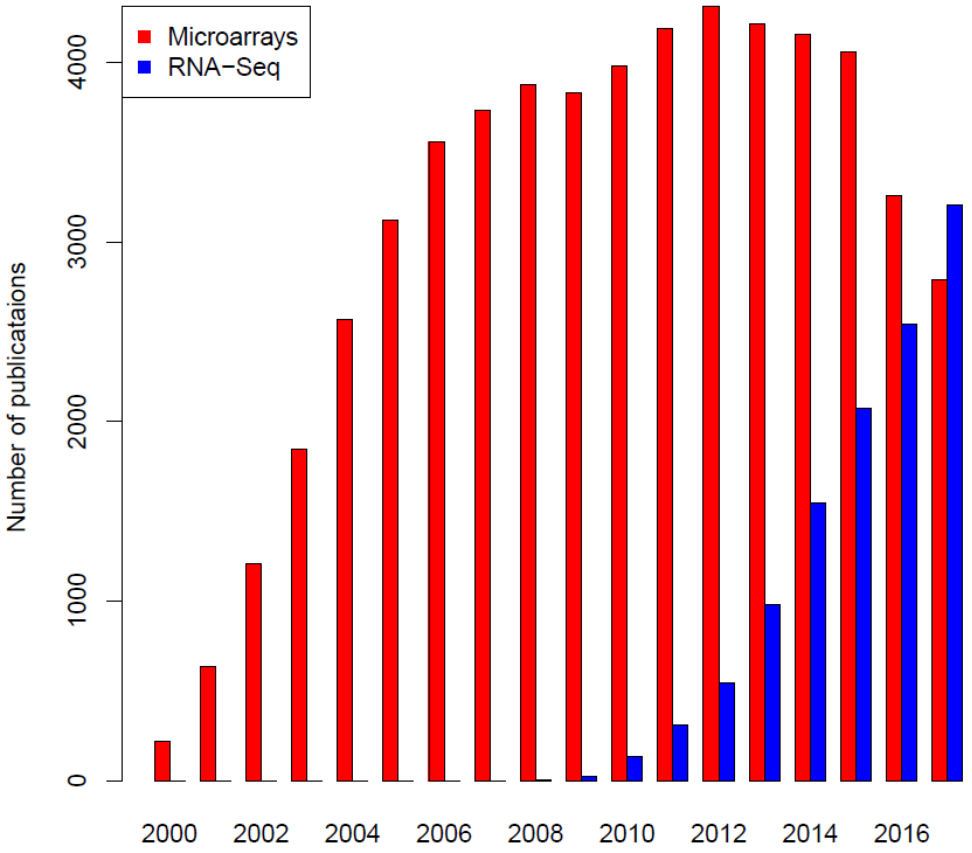
Affymetrix (now part of ThermoFisher)

- Continued with array technology and diversified (more species, able to deal with FFPE samples, gene expression/genotyping/methylation, customised content)
- Chosen as partner for many large-scale efforts including UK BioBank

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. For the same cost, can do a higher-powered expt
  - 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
- Expression estimates not necessarily similar for same sample but fold-change concordance between groups of samples reasonably high between arrays and RNA-Seq

RNA-Seq Myths and Caveats

Oxford Genomics CENTRE

✖ ‘Digital’ or absolute gene counts obtained

✖ Can detect low expressed genes better than arrays

- Would 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

✖ Larger dynamic range than arrays

- Maybe at high end but not low end, and no noticeable difference in the range of expression and fold changes seen in a typical experiment

- Furthermore, have some unique issues - what you sequence in an RNA-Seq library influences your data for all genes; very inter-dependent in a way that arrays are not

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, and require human decision-making to take the information further
    - What genes/pathways to focus follow-up experiments on?
    - Different researchers could easily identify quite different themes from the same results
    - Much is left untouched
- 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

Large-scale gene expression projects

Oxford Genomics CENTRE

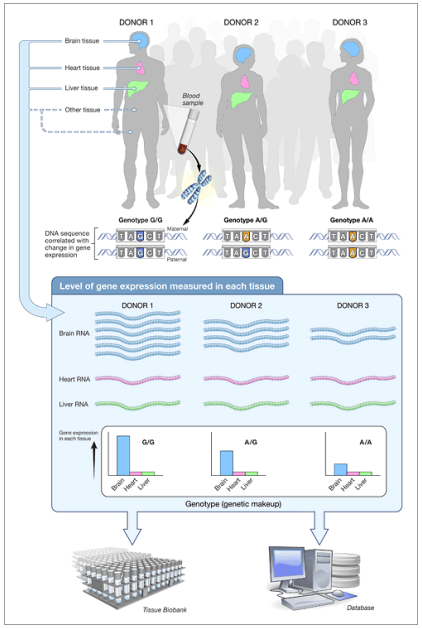
- ENCODE
- Allen brain atlas
- Genotype-Tissue Expression Project (GTEx)
- TGCA

ALLEN BRAIN ATLAS
DATA PORTAL

ENCODE

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



wellcome centre human genetics

UNIVERSITY OF OXFORD

Choice of tools

Oxford Genomics CENTRE

Any sensible pipeline will produce reasonable results

- **DON'T** worry about getting to a definitive answer (it doesn't exist anyway)
- **DO** worry about applying chosen tools carefully
  - Is it suitable for my question and data I have?
  - Have I understood how it works and how the parameter settings and options affect its behaviour?
  - Have I given the right input and made sure the output is sensible?
  - Have I checked my R code for mistakes or unintended behaviour?

It is all too easy to for untoward things to happen somewhere along the line, and also surprisingly hard to spot them in high dimensional data

RNA-Seq Generates Count Data

ENSG00000069011	211	126	168	157	146	145	168
ENSG00000069018	0	0	1	0	0	0	0
ENSG00000069020	212	359	548	134	195	193	278
ENSG00000069122	0	0	0	0	0	0	1
ENSG00000069188	54	62	73	116	136	103	83
ENSG00000069206	0	0	0	0	0	0	0
ENSG00000069248	731	748	770	632	766	582	678
ENSG00000069275	11847	12391	9959	13182	15600	12974	11946
ENSG00000069329	1586	1591	1473	1551	1801	1435	1740
ENSG00000069345	1051	988	1091	945	1072	876	1067
ENSG00000069399	152	154	279	101	94	97	120
ENSG00000069424	84	68	84	69	81	75	56
ENSG00000069431	95	116	115	86	109	107	94

- Many genes (typically half) are not sequenced at all
- Raw counts are not comparable across samples (e.g. depth, composition effects)
- Also not comparable between genes for the same sample (e.g. different lengths, amplification biases)

Methods for RNA-Seq Data

- A variety of packages for processing and analysing RNA-Seq data were developed to handle count-based expression data and a myriad of sequencing-related effects and biases (longer genes generate higher counts, amplification biases related to sequence composition, library composition/complexity and so on).

- Tailored methods required for every step including alignment, summarising gene/transcript counts, normalisation, testing for differential expression and pathway/enrichment analysis

- For differential expression analysis, the dilemma was whether to develop novel methods and work directly with the counts, or transform the data in a way to meet the assumptions of the existing methods for microarray data

BioConductor ‘limma’ package

- limma: Linear models for microarray data
- http://bioconductor.org/packages/release/bioc/html/limma.html
- Originally developed over 15 years ago to handle microarray data (including preprocessing and data analysis) and provides a comprehensive framework for gene expression data analysis
- Widely used and gold standard in the field, developed by Gordon Smyth and colleagues at Walter and Eliza Hall Institute (WEHI), Melbourne, Australia
- Some aspects now obsolete e.g. methods for 2-colour microarrays but has evolved with the field of transcriptomics and now able to analyse RNA-Seq data with limma, if suitably transformed beforehand
- Introduces some fundamental concepts for analysing gene expression data in R
- Implements standard statistical methods (linear regression) but with additional features tailored to gene expression experiments

RNA-Seq Analysis Tools

- edgeR
  - Analagous steps to limma but uses different statistical models for testing for differential expression
  - Computationally efficient and able to analyse complex experimental designs
  - Biological variability between samples increases the variance in the counts - negative binomial models fitted
  - Estimation of biological variability (dispersion) performed empirically
  - Like limma, borrows information across genes to improve estimates from small sample sizes
- limma-voom
- Alternative to count-based models is suitably transforming the counts and using existing standard statistical approaches
- ‘voom’ function does this so can use limma to analyse RNA-Seq data too!

Experimental design considerations

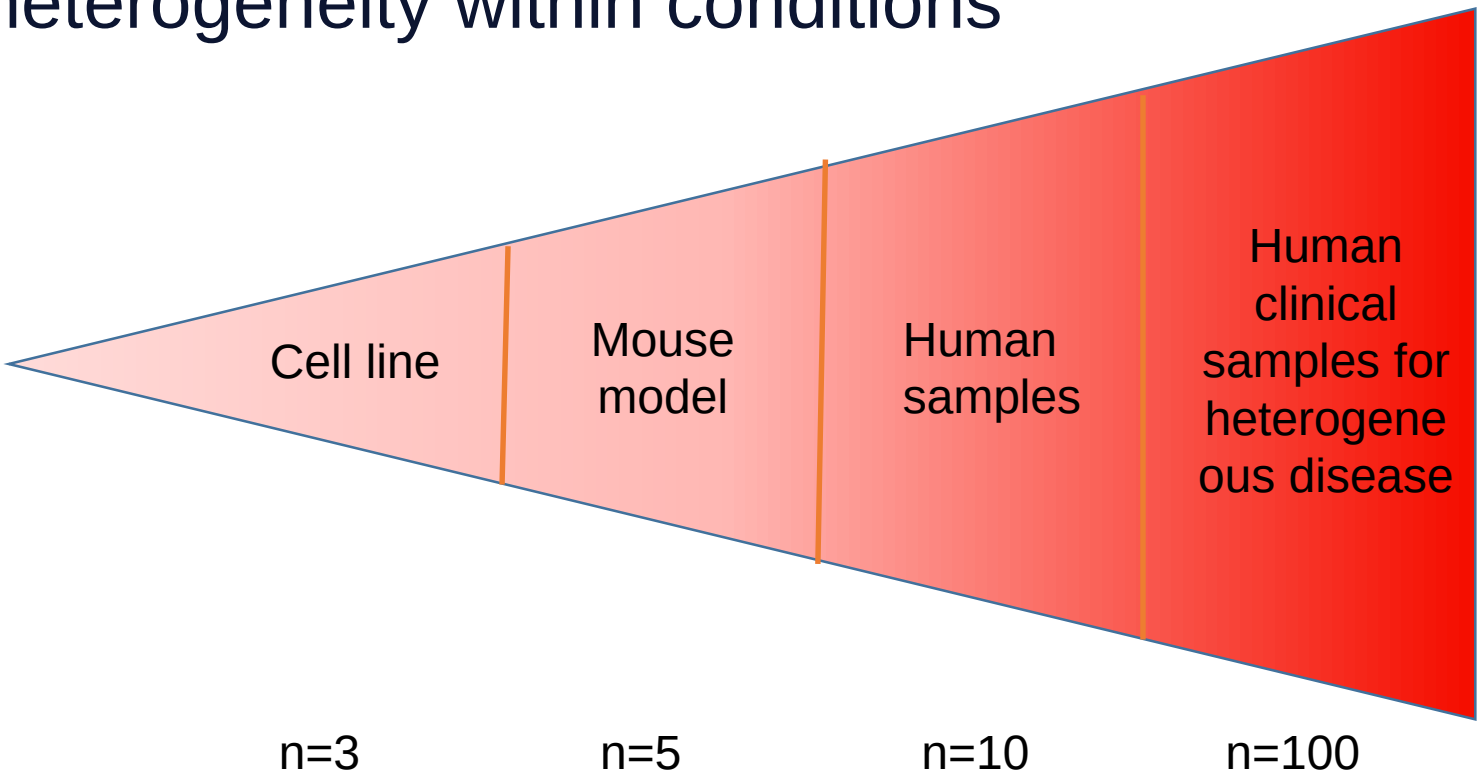
Typical experimental designs

- 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

Replication

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



Sequencing Depth

- Number of reads required per sample depends on experimental question

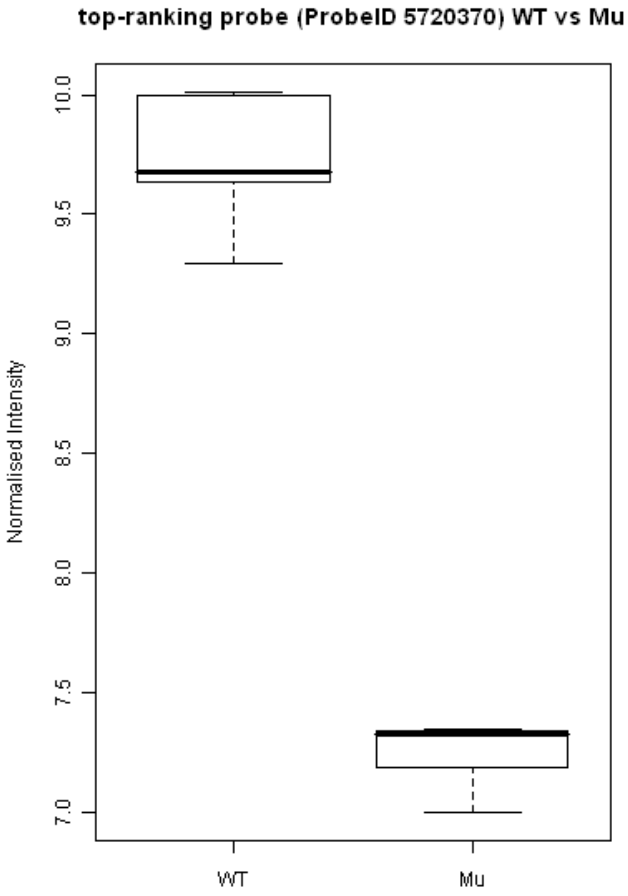
- HiSeq4000 – one lane = 250 million reads
- Multiplexing e.g. 10-plex human samples gives ~25m reads for each, plenty for quantifying gene expression (except for very low/unexpressed genes)

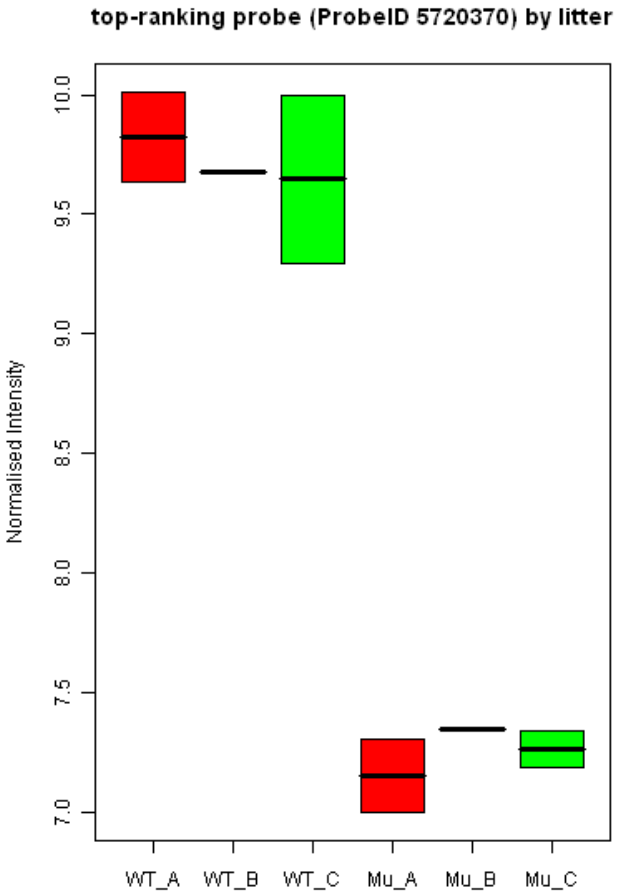
- Higher depth required in some situations e.g. for splicing analysis, certain library prep methods (ribo-depletion)

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 a far 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

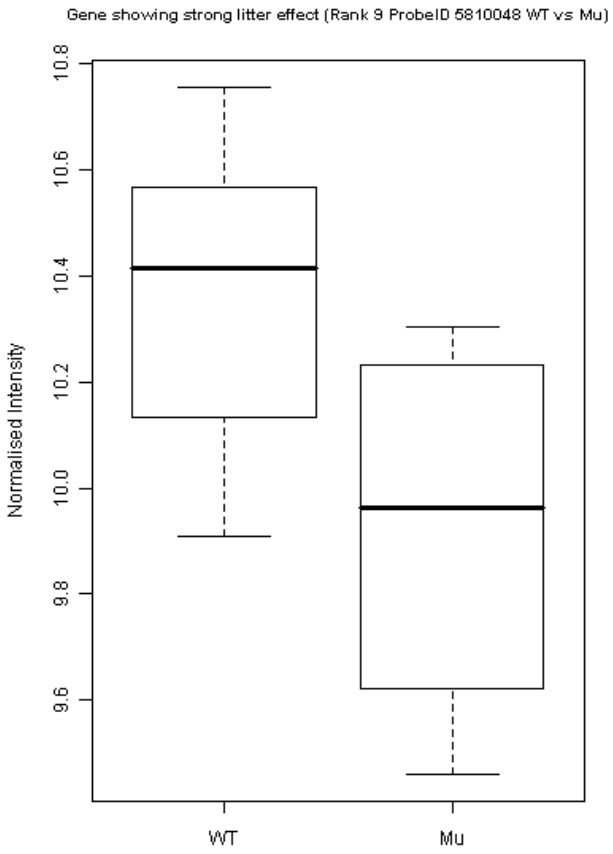
Mouse expt - example of a gene not influenced by litter

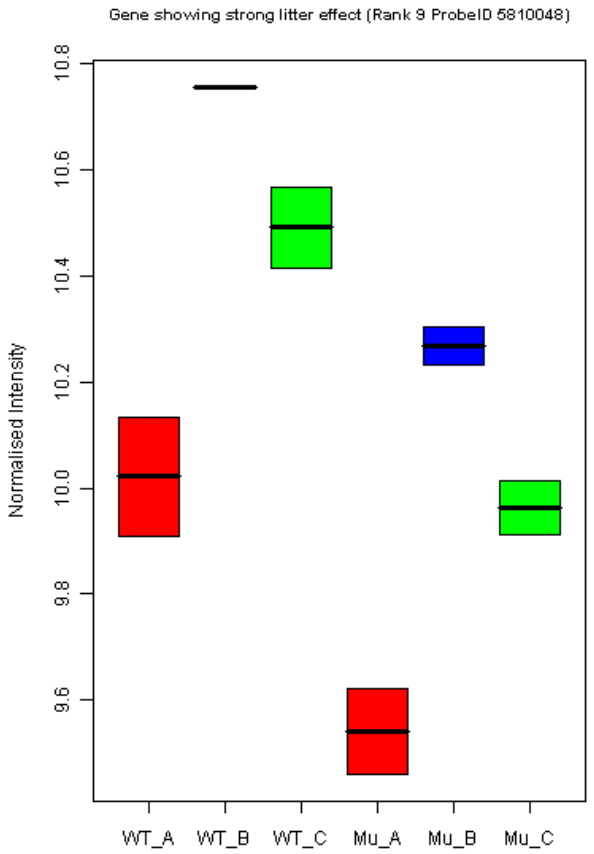




- Wt and Mut groups
- Three different litters
- Top gene ~ 5x higher expression in Wt compared to Mut
- Similarly expressed across litters in both genotypes

Another gene shows strong litter effect





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