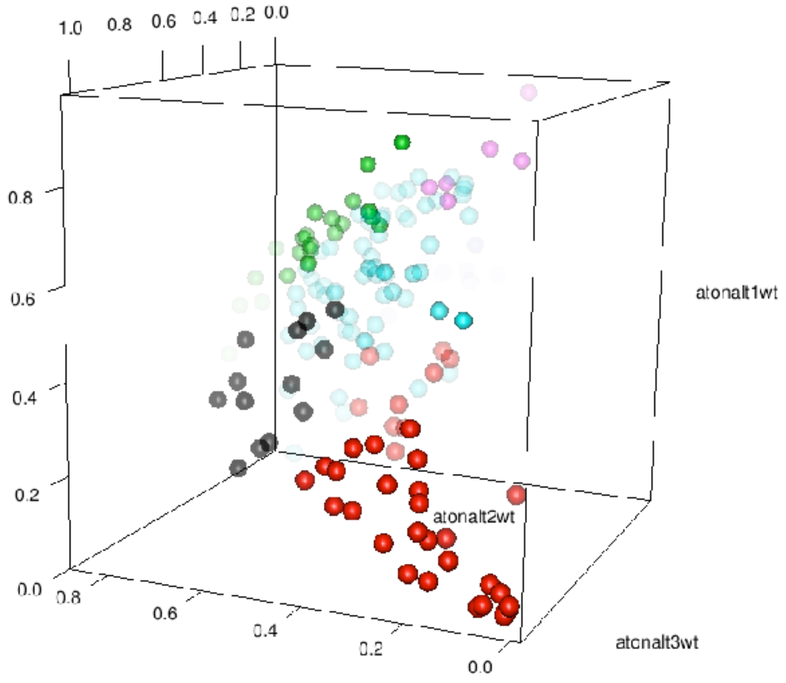
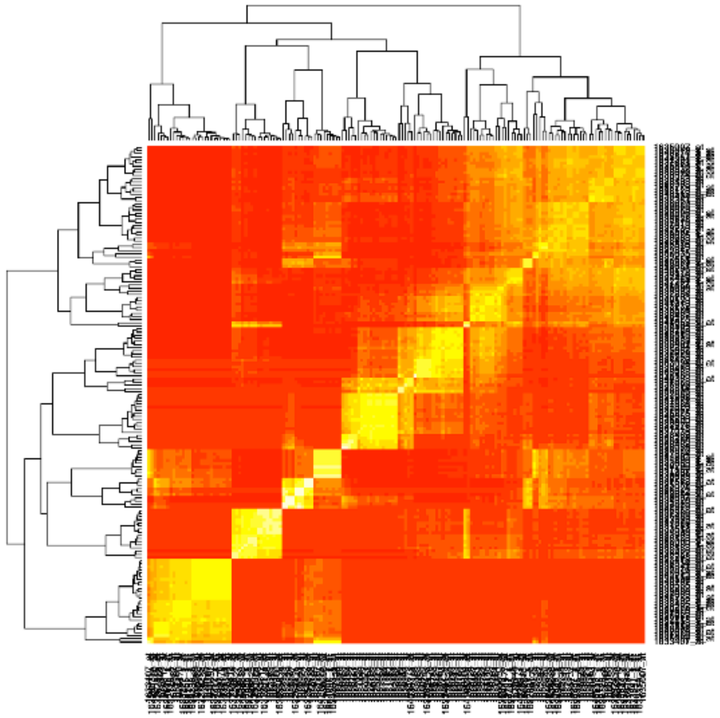
Lecture 10 - Gene Expression Analysis





Overview

Introduction to Gene

- ExpressionMicroarrays
- RNA Sequencing

Analysis of Gene Expression Data

- Pre-processing
- Differential expression analysis
- Downstream Functional Analysis

Practical Analysis in R

- microarray
- RNA-seq

Comparing Samples

Quality Control
- artefacts
- missing data
- outliers (samples and point-data)**

Batch-correction (co-variates)**

Normalisation**

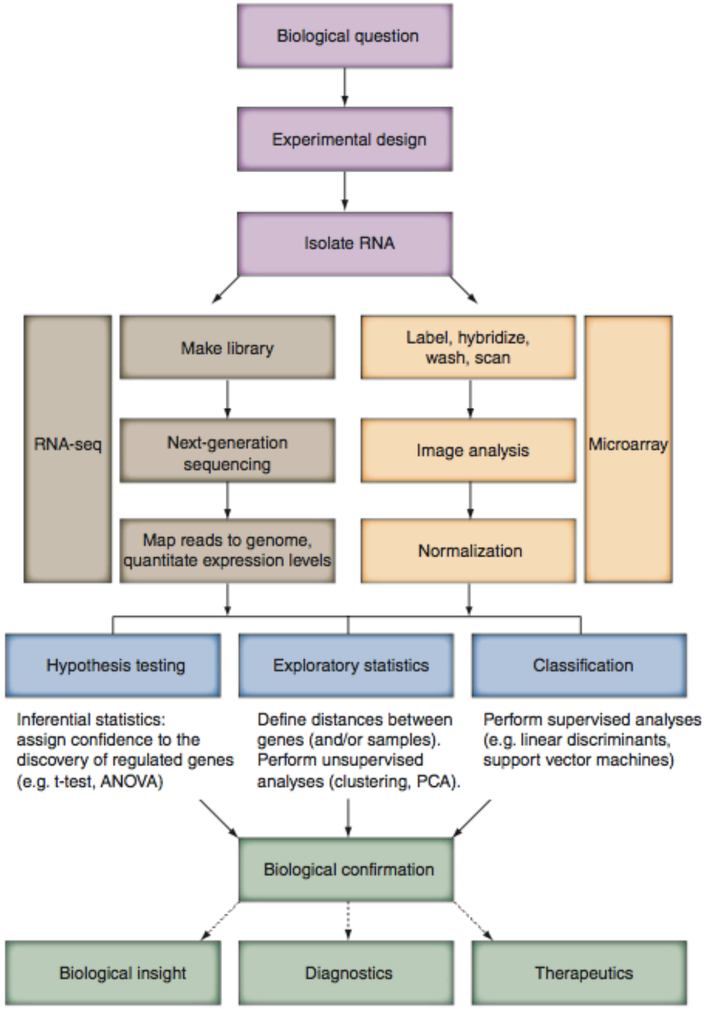
Summarisation

Statistical Testing (including multiple testing correction)

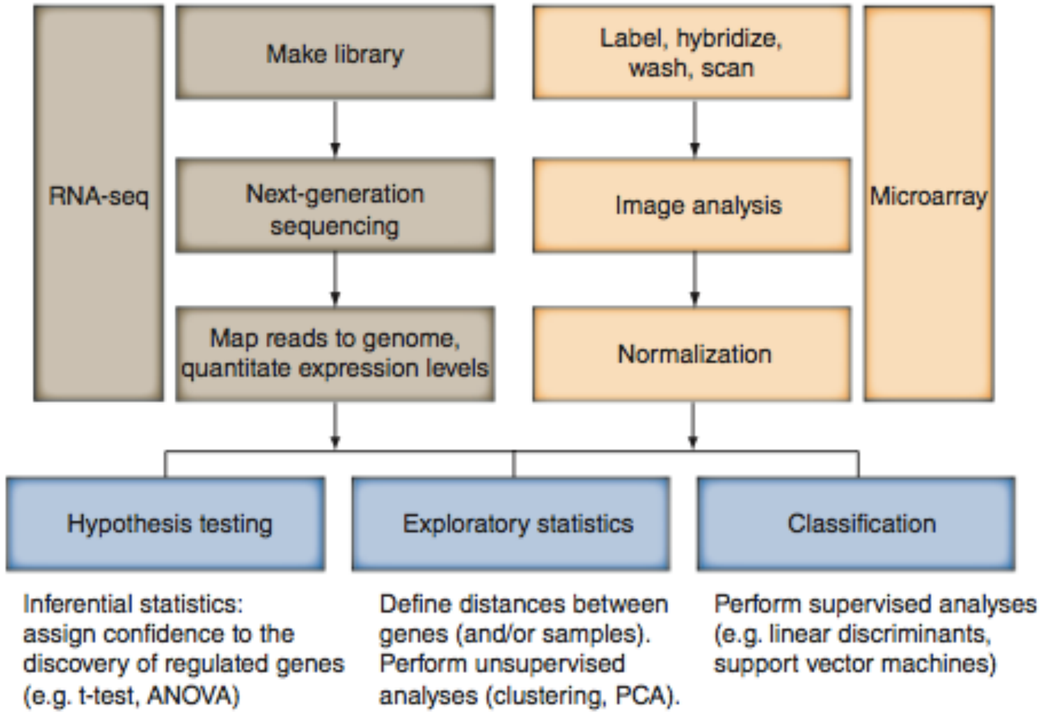
Downstream Analysis

**beware here be dragons

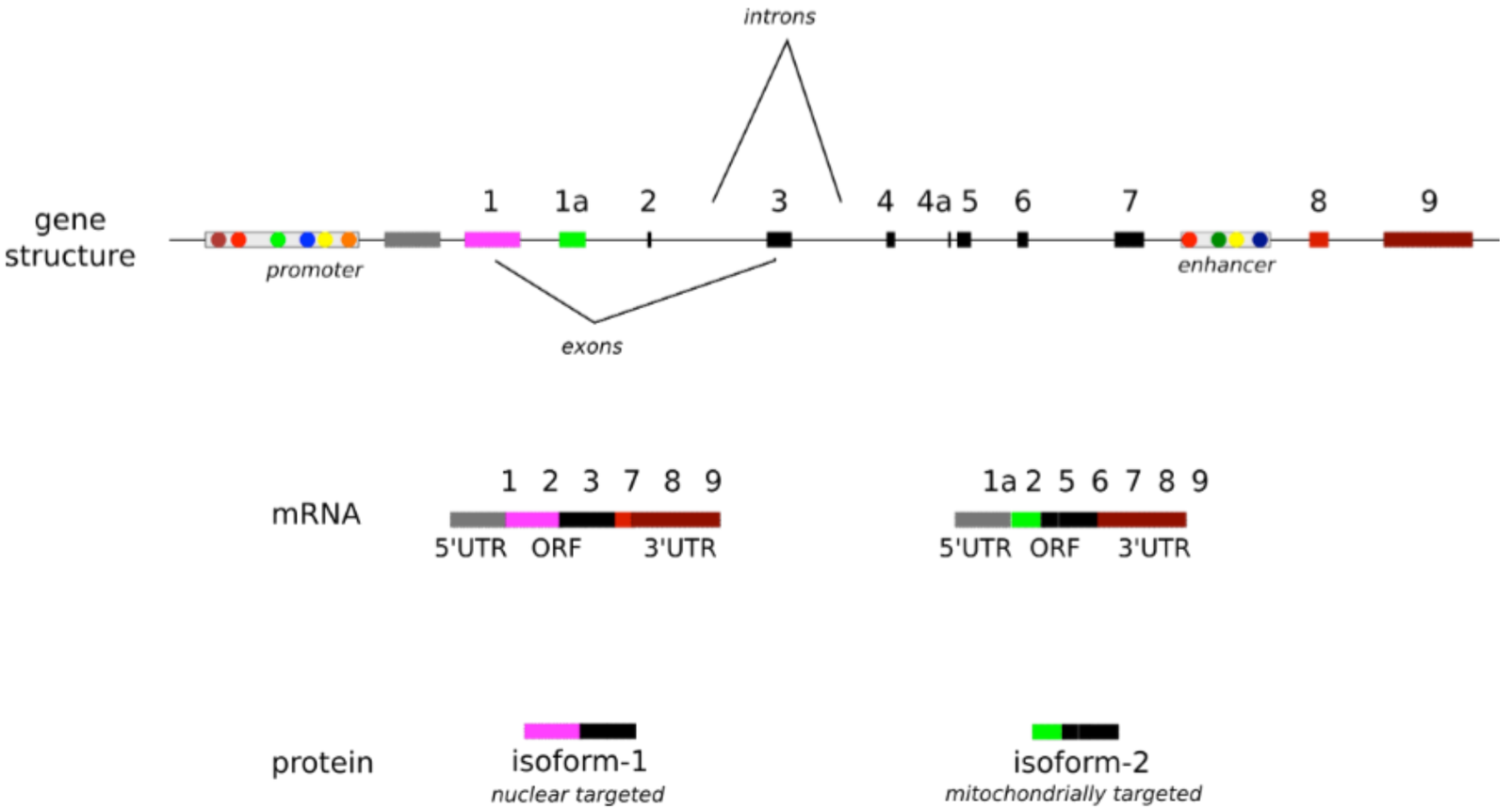
Workflow for Assessing Gene Expression Changes



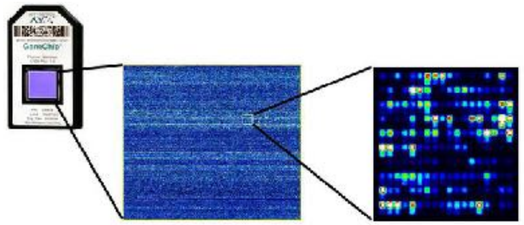
Workflow for Assessing Gene Expression Changes

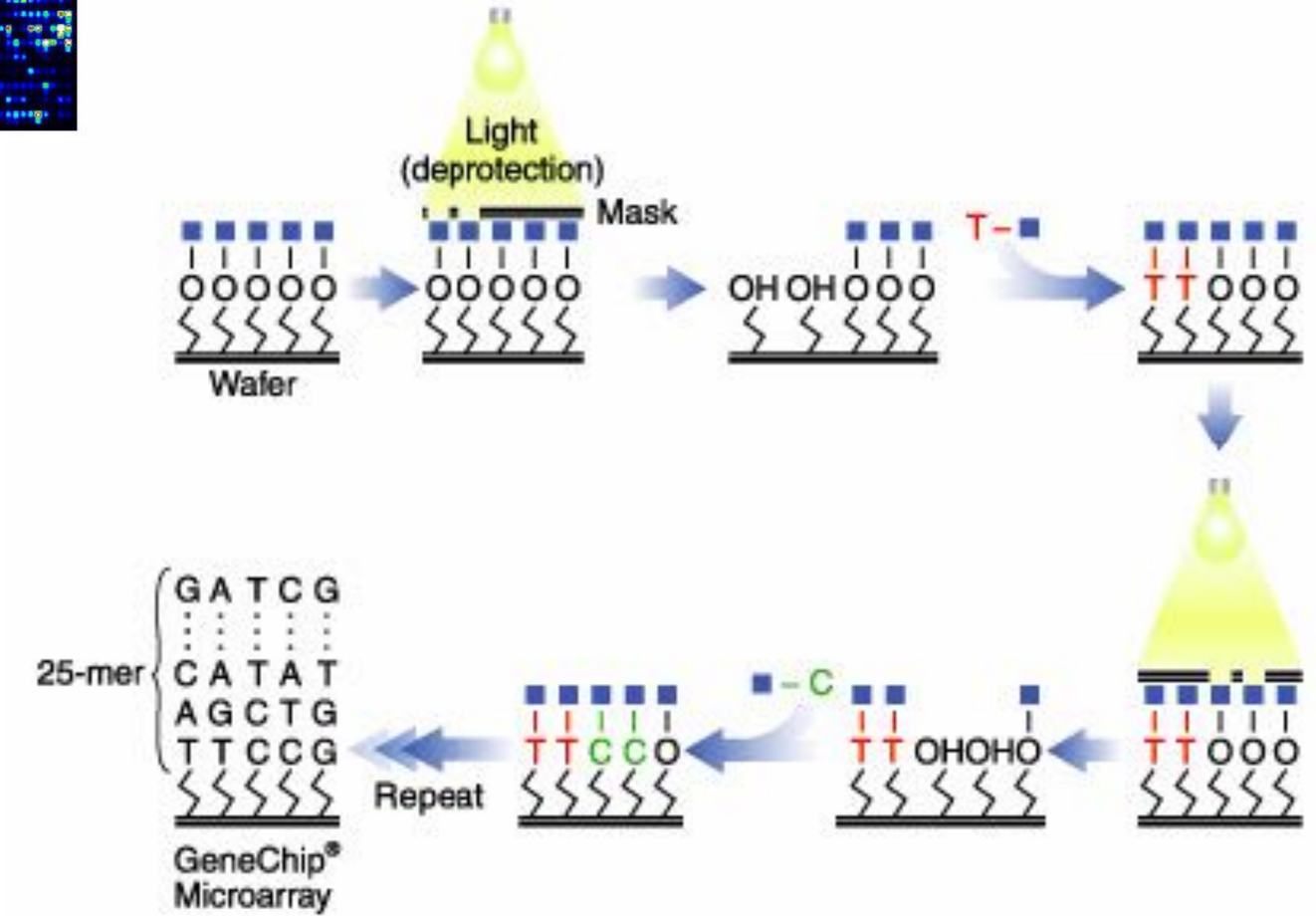


Gene structure

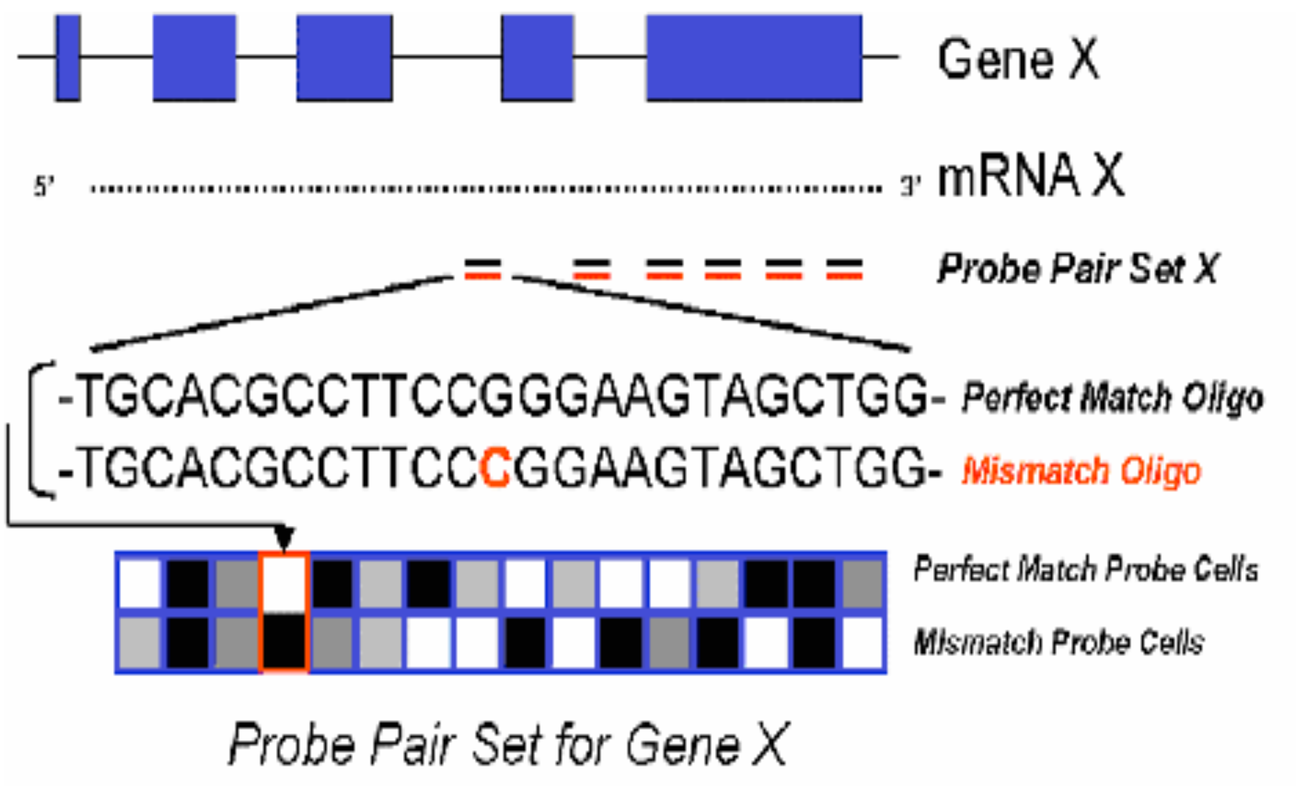


Measuring gene expression with microarrays

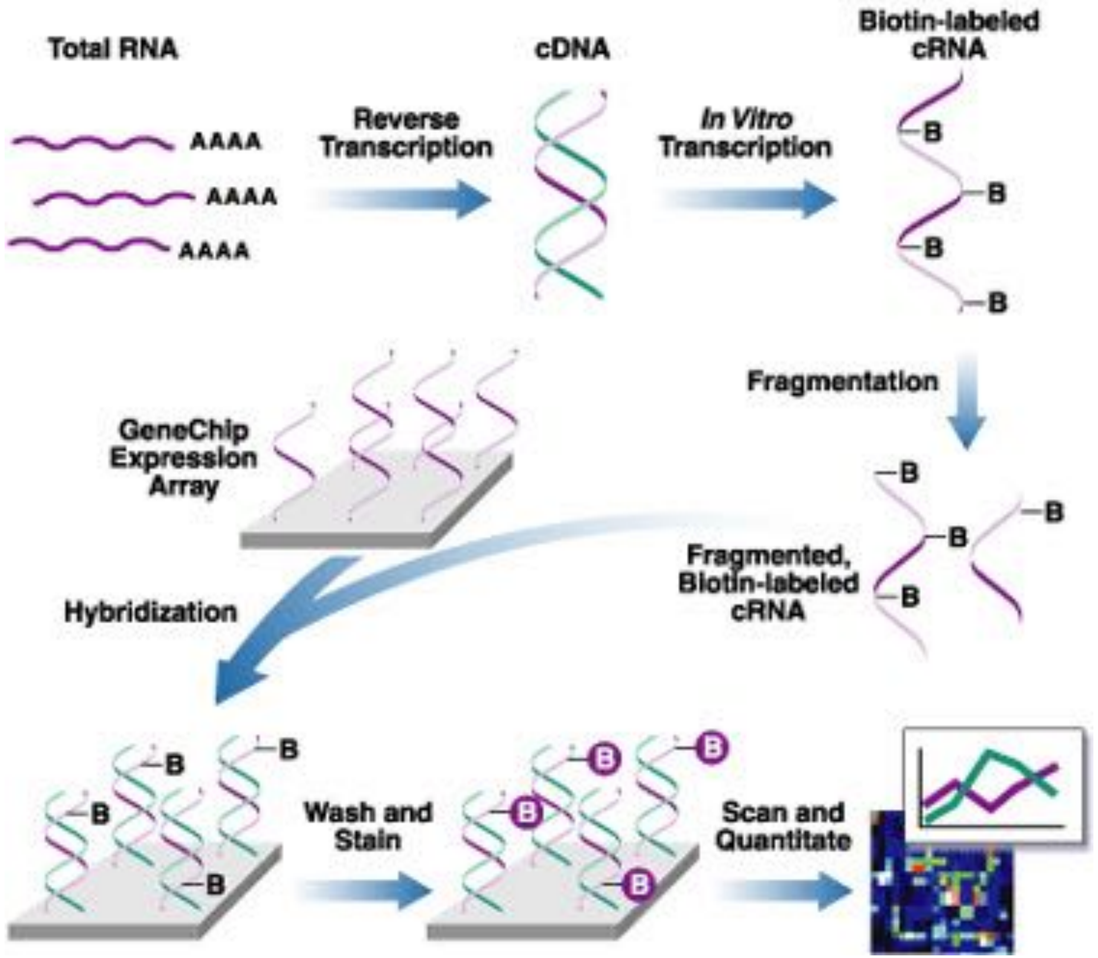




low-level chip analysis
PM/MM pairs make up probe-sets

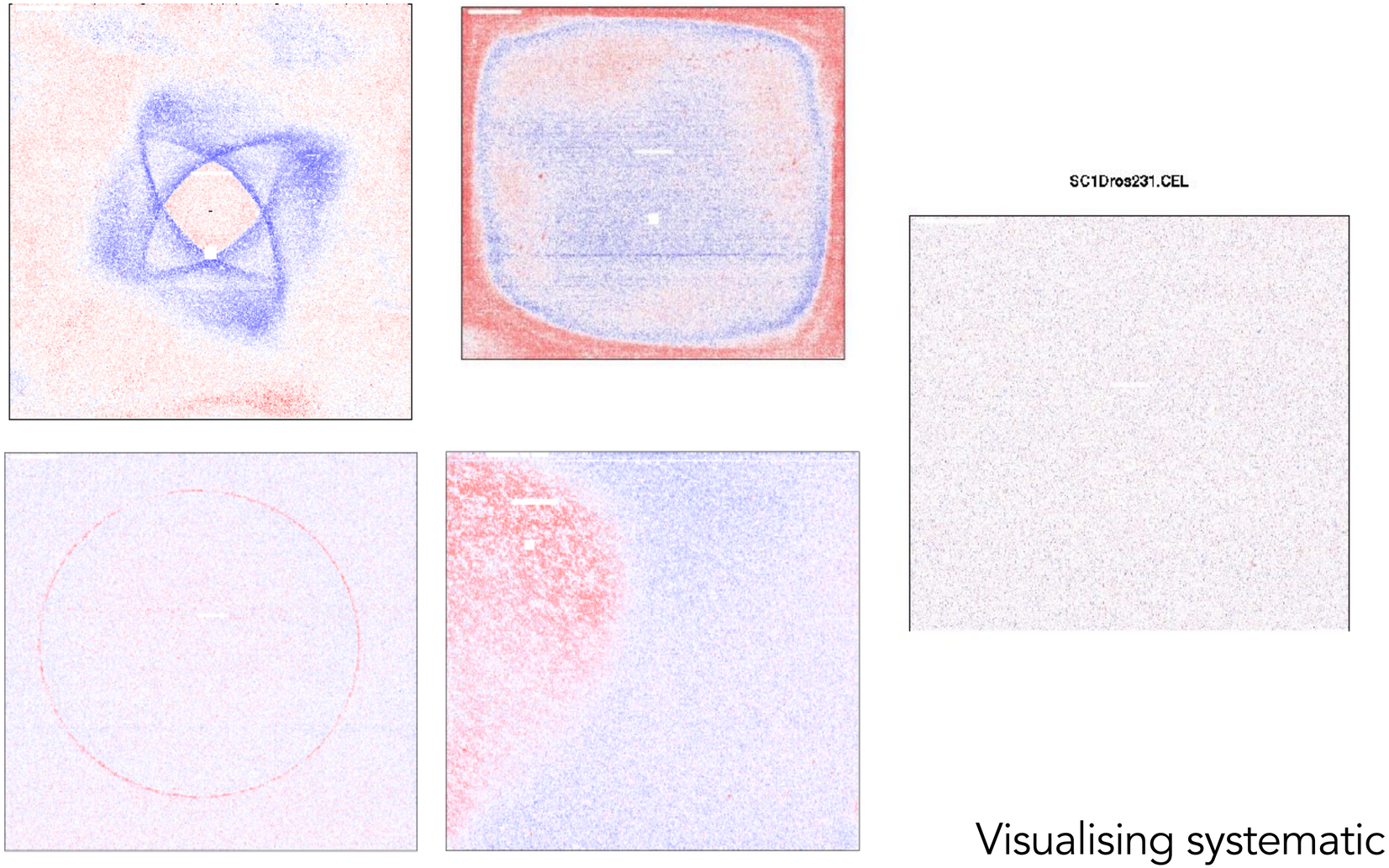


Measuring gene expression with microarrays



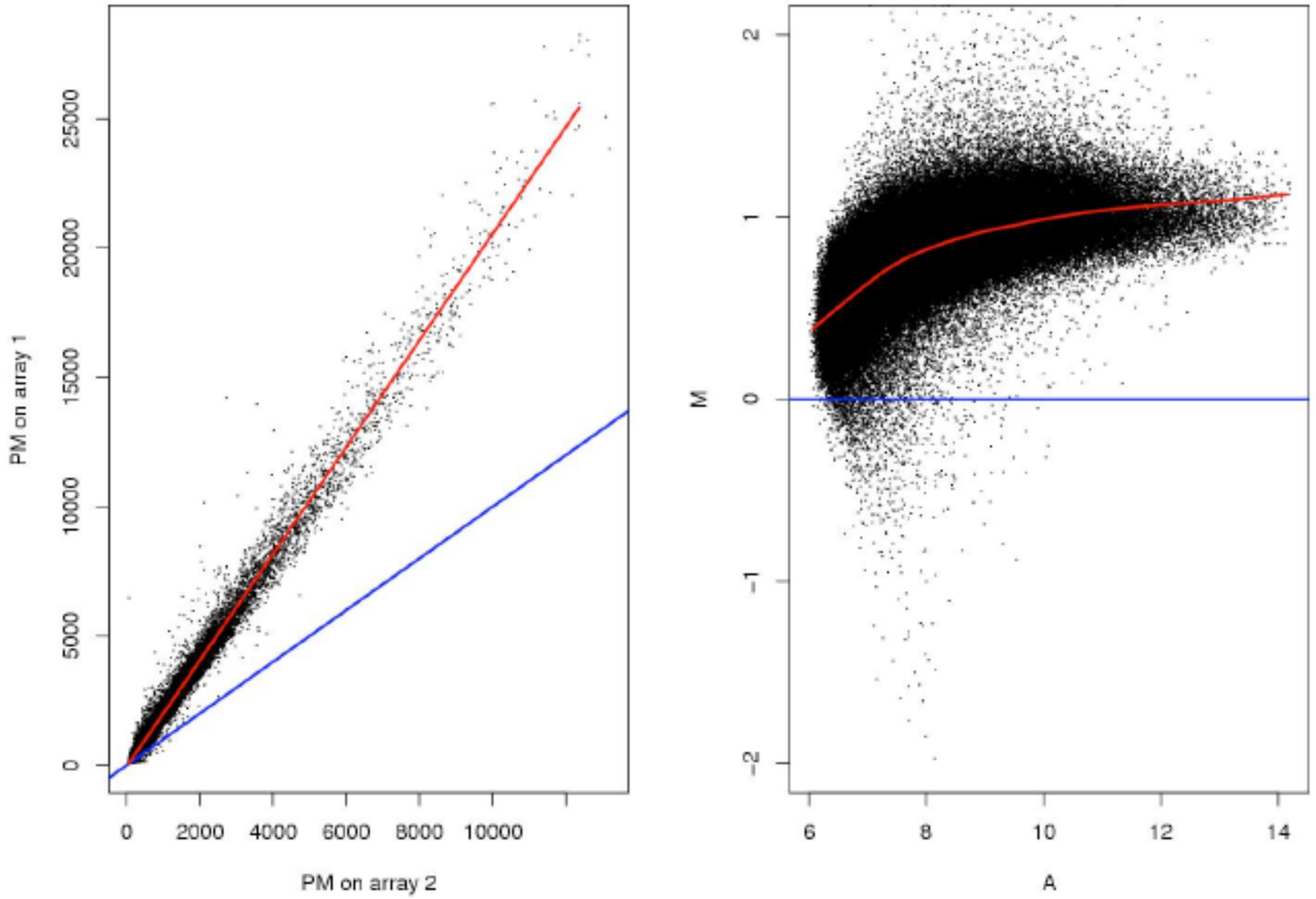
labelling and hybridisation

Plotting residuals from PLM fits



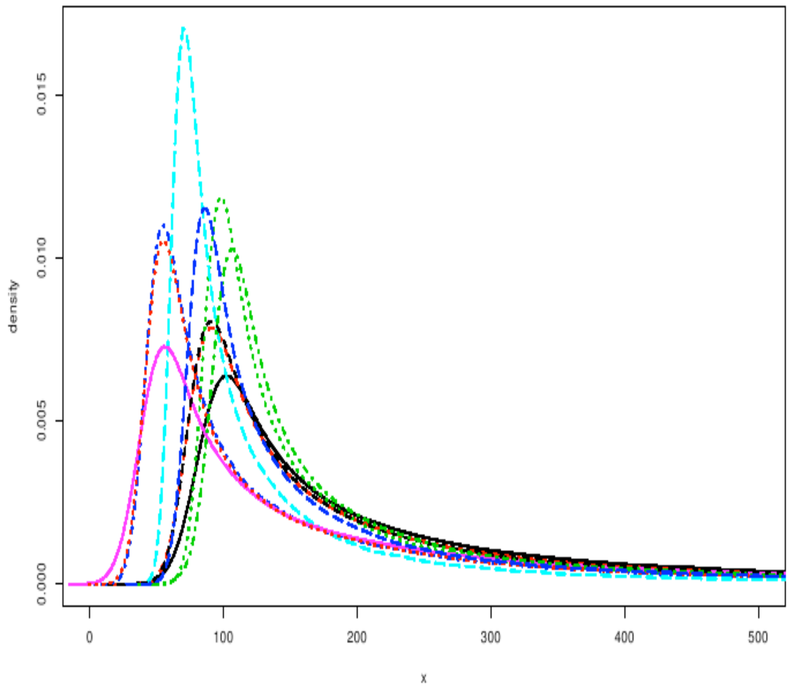
Visualising systematic error

comparing probe intensities

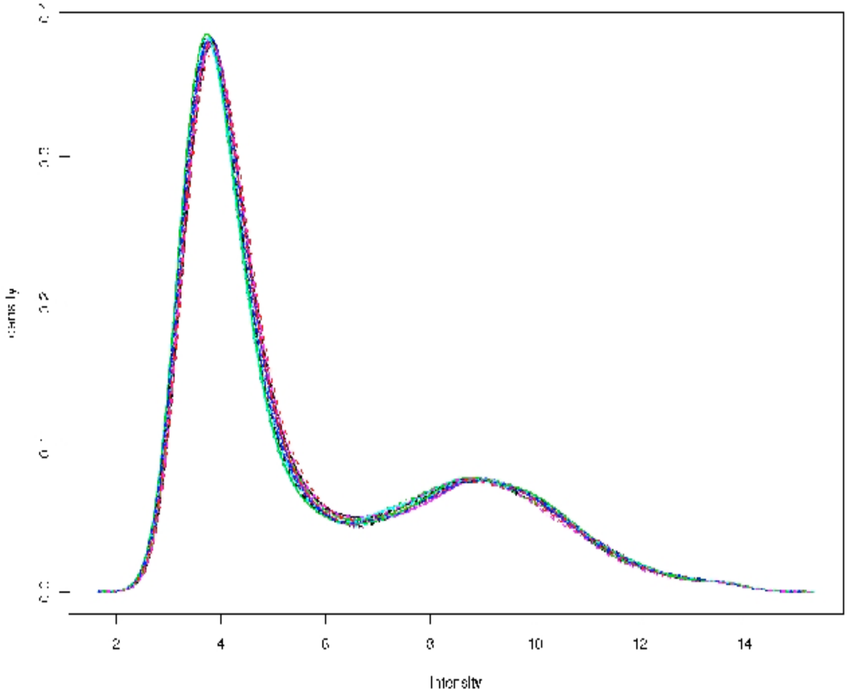


Normalisation

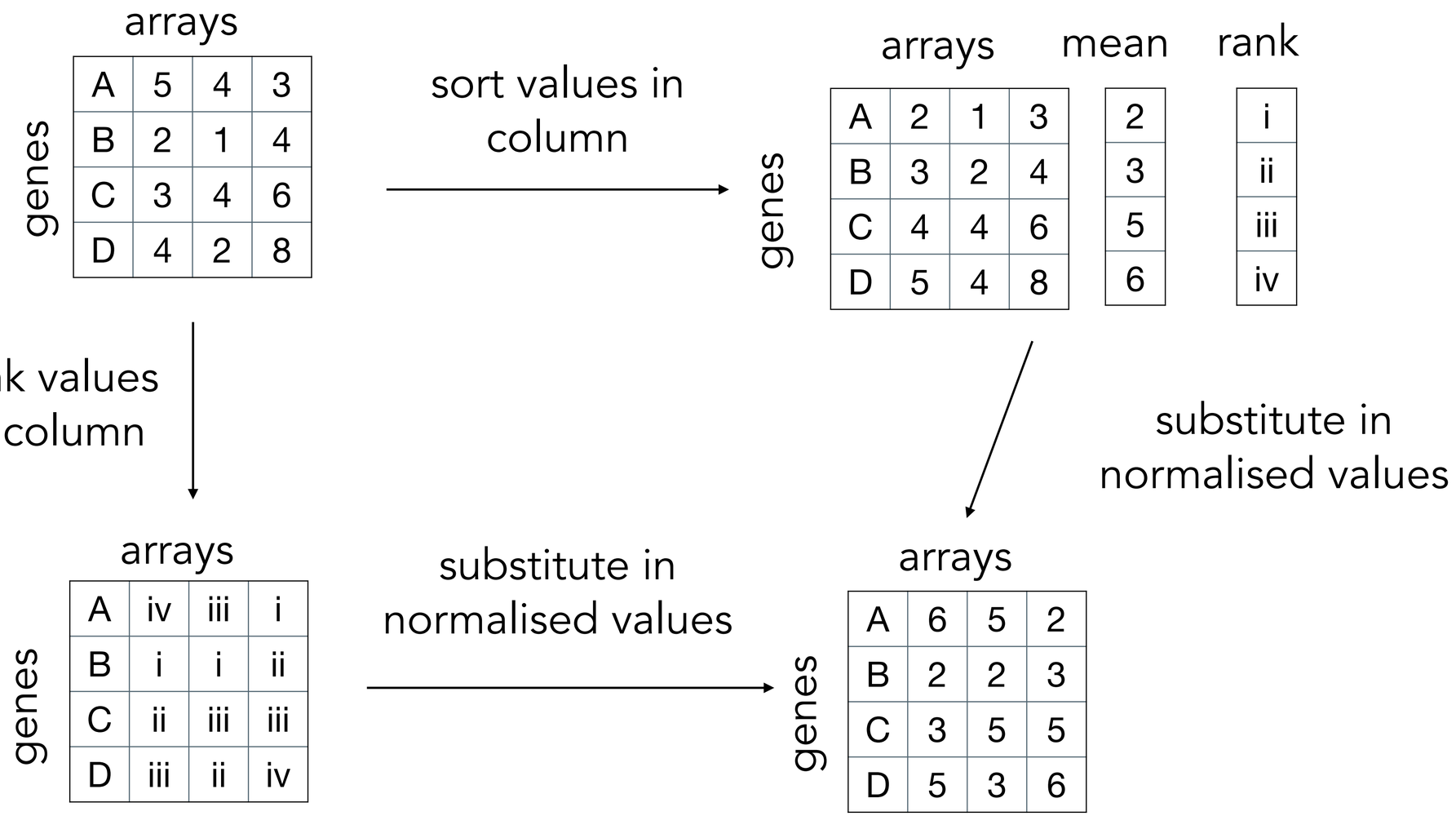
before normalisation



after normalisation



Quantile normalisation



Robust multi-array analysis (RMA)

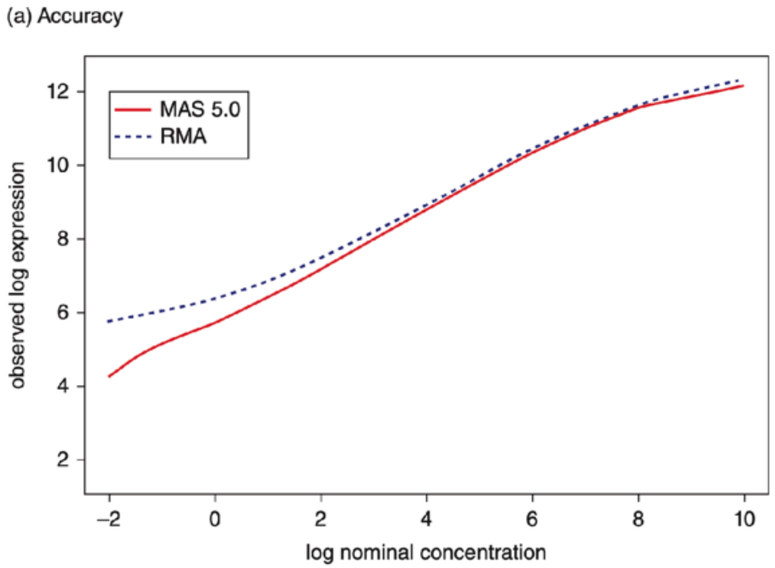
- Developed by Rafael Irizarry, Terry Speed et al.
- Available at www.bioconductor.org as an R package ('affy')

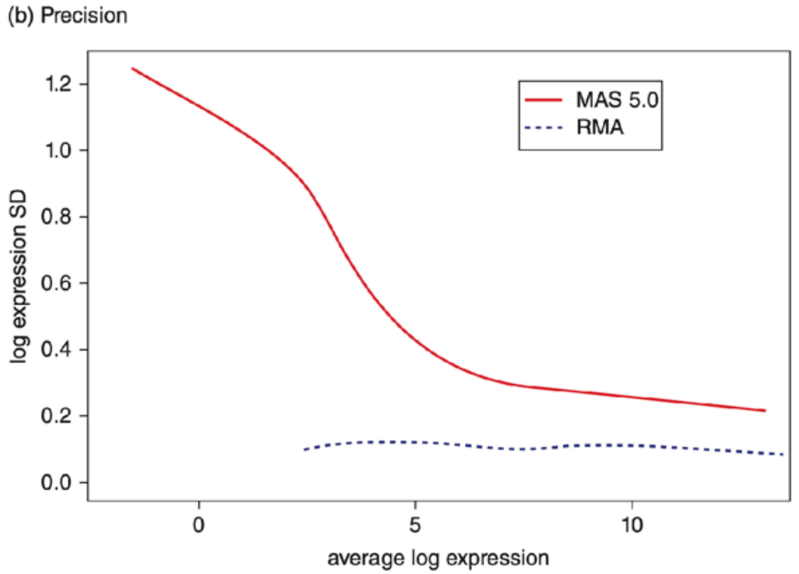
https://bioconductor.org/packages/release/bioc/html/affy.html

There are three steps:

- Background adjustment based on a normal plus exponential model (no mismatch data are used)
- Quantile normalisation (nonparametric fitting of signal intensity data to normalise their distribution)
- Fitting a log scale additive model robustly. The model is additive: probe effect + sample effect

RMA vs MAS5.0 (Affymetrix)



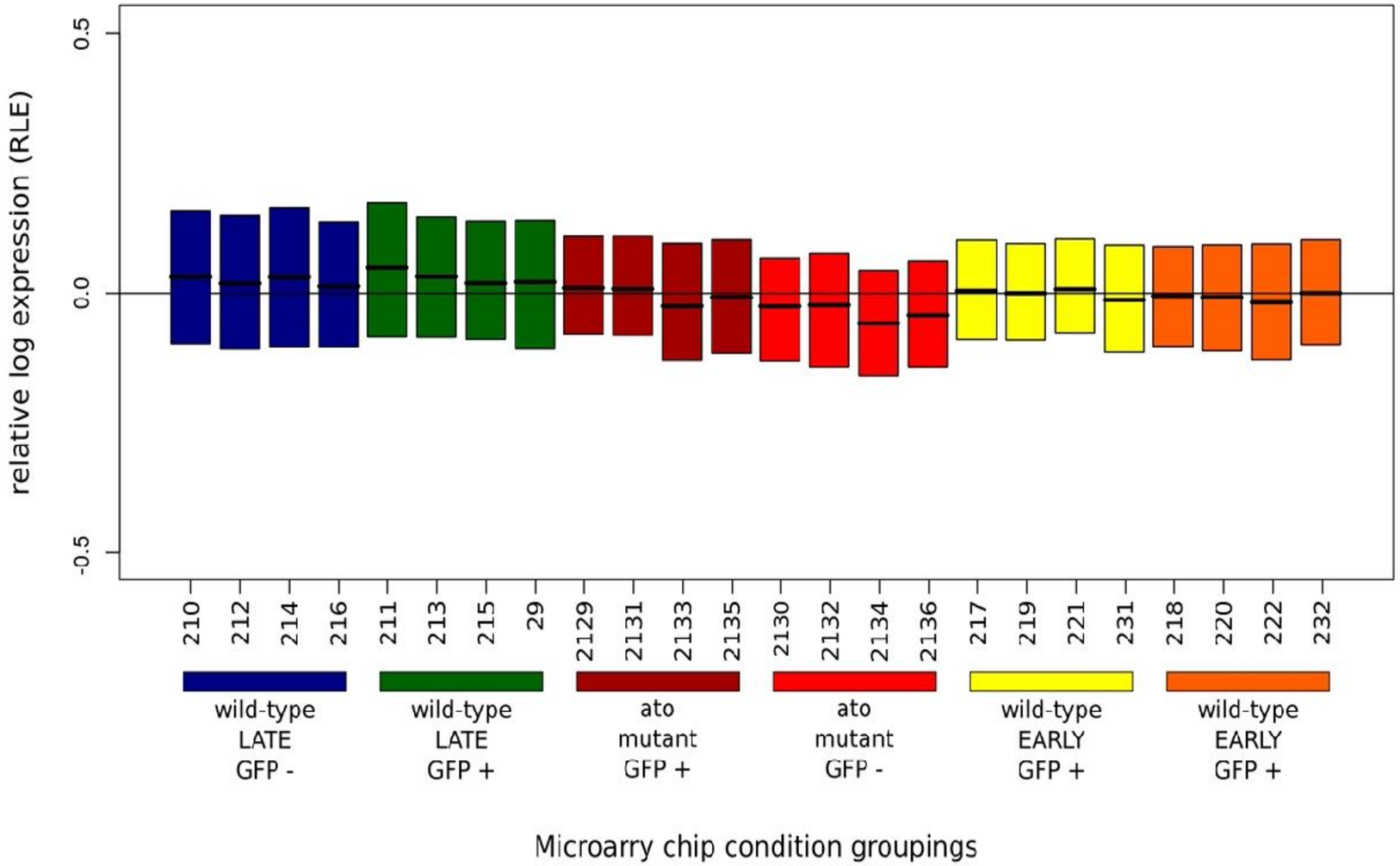


Fold change (log ratios)

- To a statistician fold change is sometimes considered meaningless. Fold change can be large (e.g. >>two-fold up- or down-regulation) without being statistically significant (e.g. based on probability values from a t-test or ANOVA).

- To a biologist fold change is almost always considered important for two reasons. First, a very small but statistically significant fold change might not be relevant to a cell's function. Second, it is of interest to know which genes are most dramatically regulated, as these are often thought to reflect changes in biologically meaningful transcripts and/or pathways.

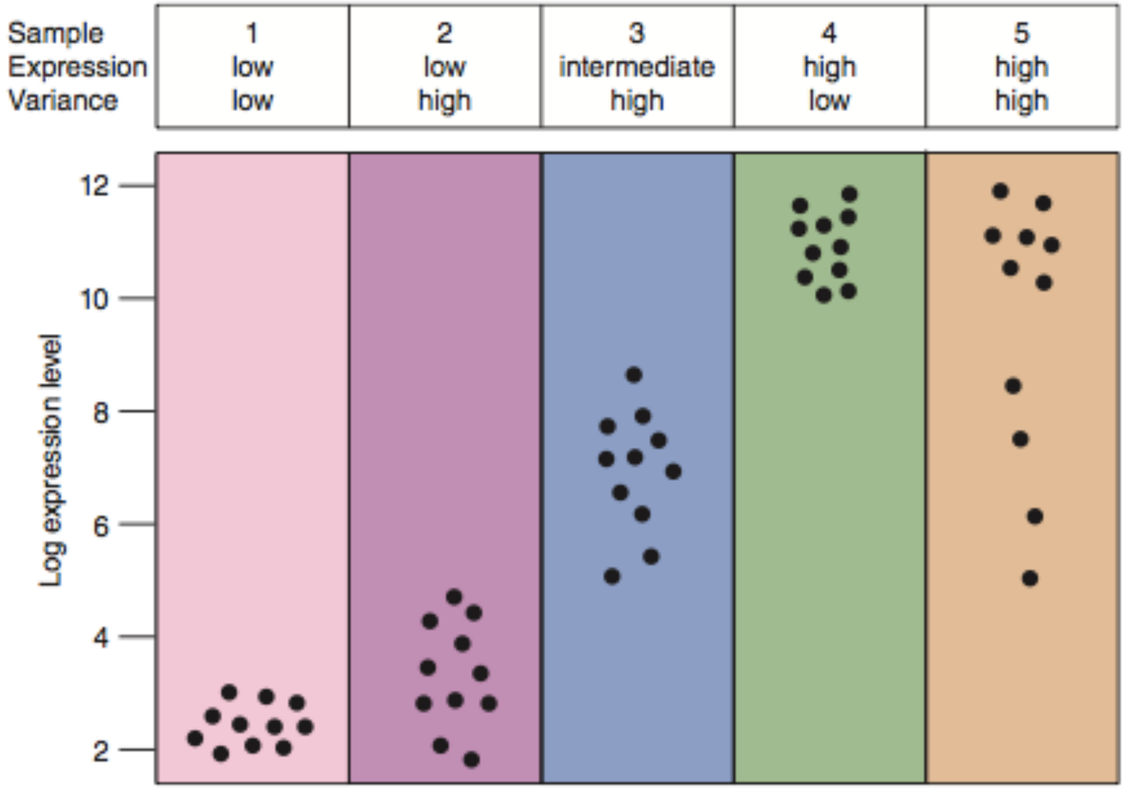
Relative log expression(RLE)



Inferential statistics

- Inferential statistics are used to make inferences about a population from a sample
- Hypothesis testing is a common form of inferential statistics. A null hypothesis is stated, such as:
- "There is no difference in signal intensity for the gene expression measurements in normal and diseased samples."
- The alternative hypothesis is that there is a difference
- We use a test statistic to decide whether to accept or reject the null hypothesis. For many applications, we set the significance level to $p < 0.05$

Transcript specific variance



Each dot is a replicate. Comparison of conditions 1 and 4 would produce a significant difference (we reject the null). Comparison of 3 versus 5 might not.

Testing differences between two groups

- Consider two groups for which you obtain measurements.
- Set a null hypothesis that there is no difference in the means of these two groups.
- Set an alternate hypothesis that there is a difference.
- In the numerator take the absolute value of the difference of the two group means.
- In the denominator calculate the noise.
- From the t-statistic obtain a probability value.

t-test

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{S_1}^2}{N_1} + \frac{{S_2}^2}{N_2}}}$$

Test statistics for gene expression data

Paradigm	Parametric test	Nonparametric test
Compare one group to a hypothetical value	One-sample *t*-test	Wilcoxon test
Compare two unpaired groups	Unpaired *t*-test	Mann–Whitney test
Compare two paired groups	Paired *t*-test	Wilcoxon test
Compare three or more unmatched groups	One-way ANOVA	Kruskal–Wallis test
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test

false discovery rate (FDR)

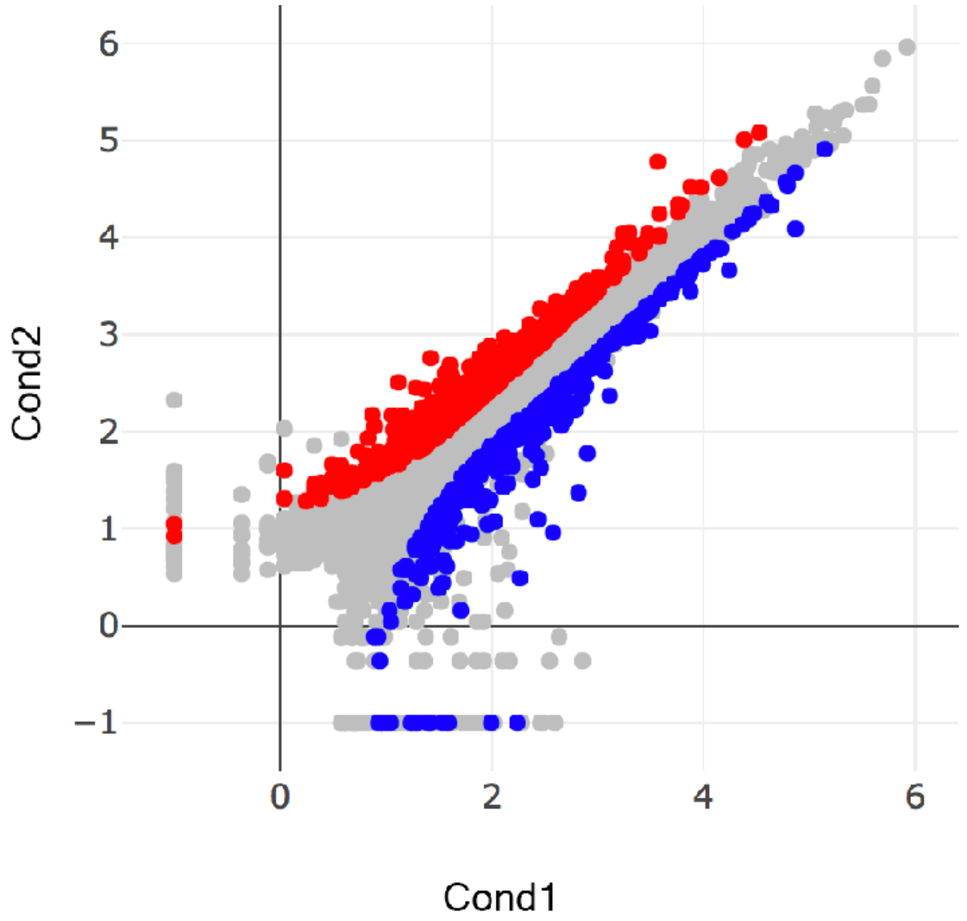
The false discovery rate (FDR) is a popular multiple testing correction. A false positive (also called a type I error) is sometimes called a false discovery.

The FDR equals the p value of the t-test times the number of genes measured (e.g. for 10,000 genes and a p value of 0.01, there are 100 expected false positives).

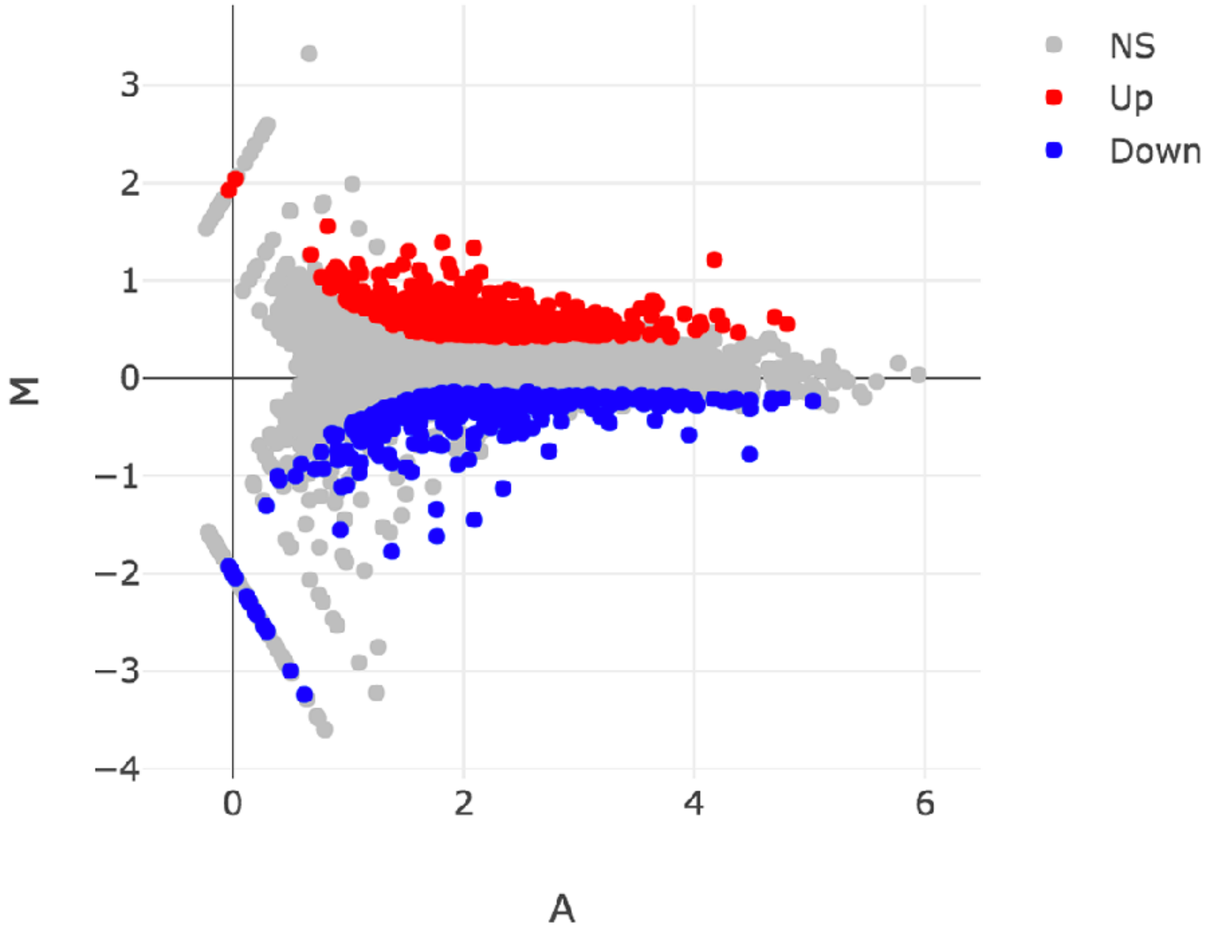
You can adjust the false discovery rate

FDR	#regulated transcripts	#false discoveries
0.10	100	10
0.05	45	3
0.01	20	1

Expression Plot



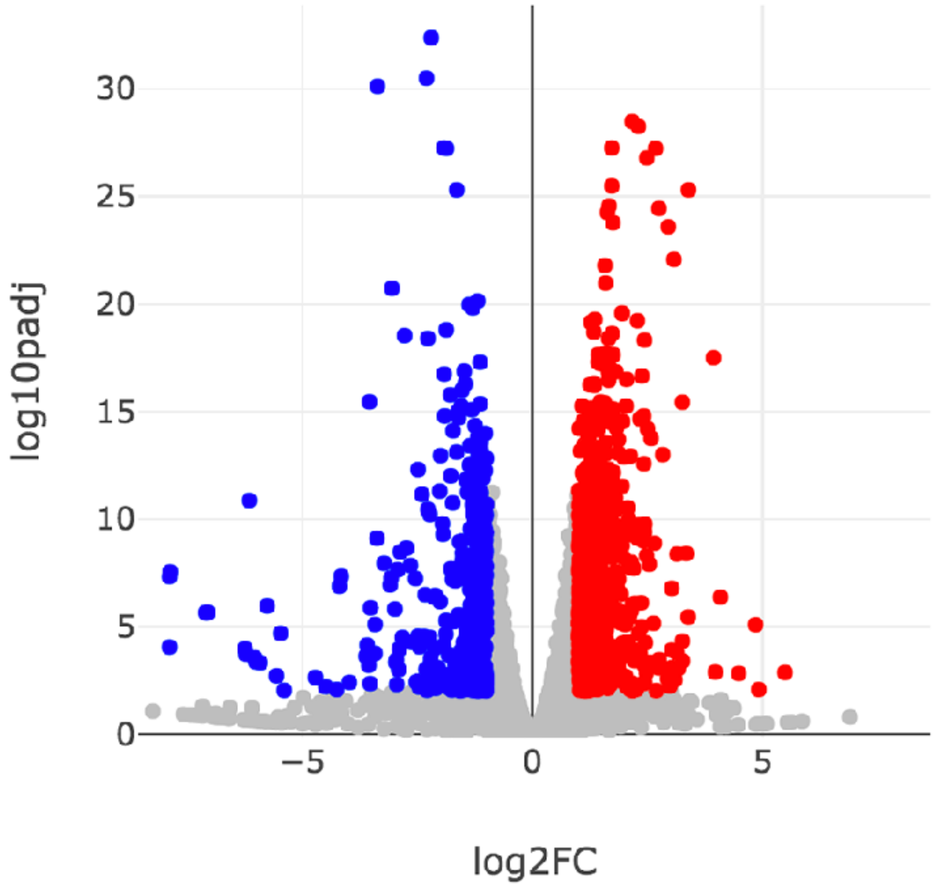
MA Plot



$$= \log_2(R/G) = \log_2(R) - \log_2(G)$$

$$= \frac{1}{2}\log_2(RG) = \frac{1}{2}(\log_2(R) + \log_2(G))$$

Volcano plot - significance of Differential Expression against fold change

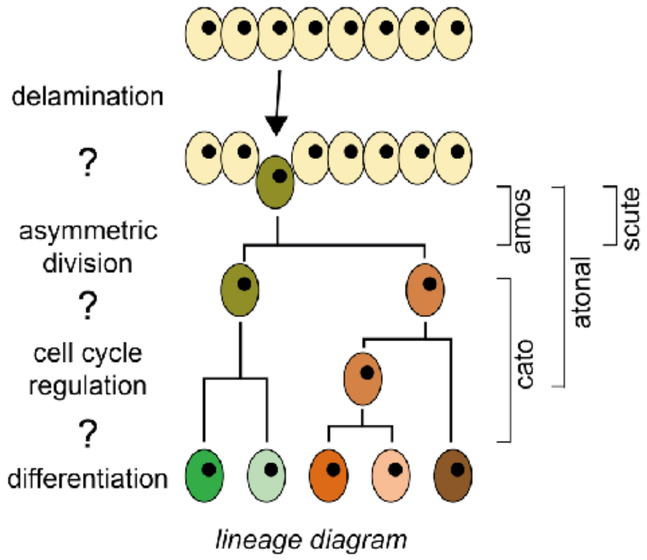


Microarray Analysis in R

- online analysis at NCBI - Geo2R
- https://www.ncbi.nlm.nih.gov/geo/geo2r/
- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21520

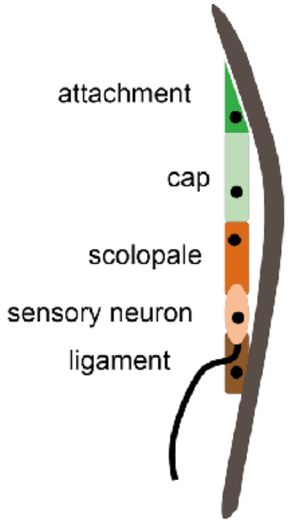
- using R locally or on a server
- RStudio (free) - https://rstudio.com/products/rstudio/
- R - https://www.r-project.org
- R/Bioconductor - https://bioconductor.org

Enriching for genes expressed during *Drosophila* neurogenesis

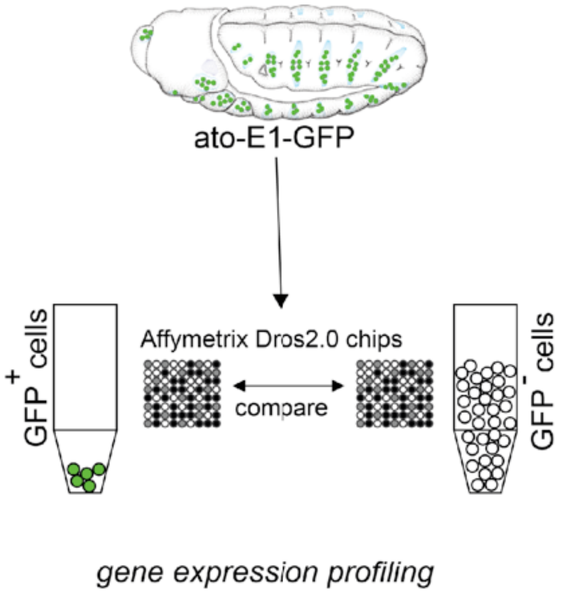


lineage diagram

timecourse



chordotonal stretch receptor



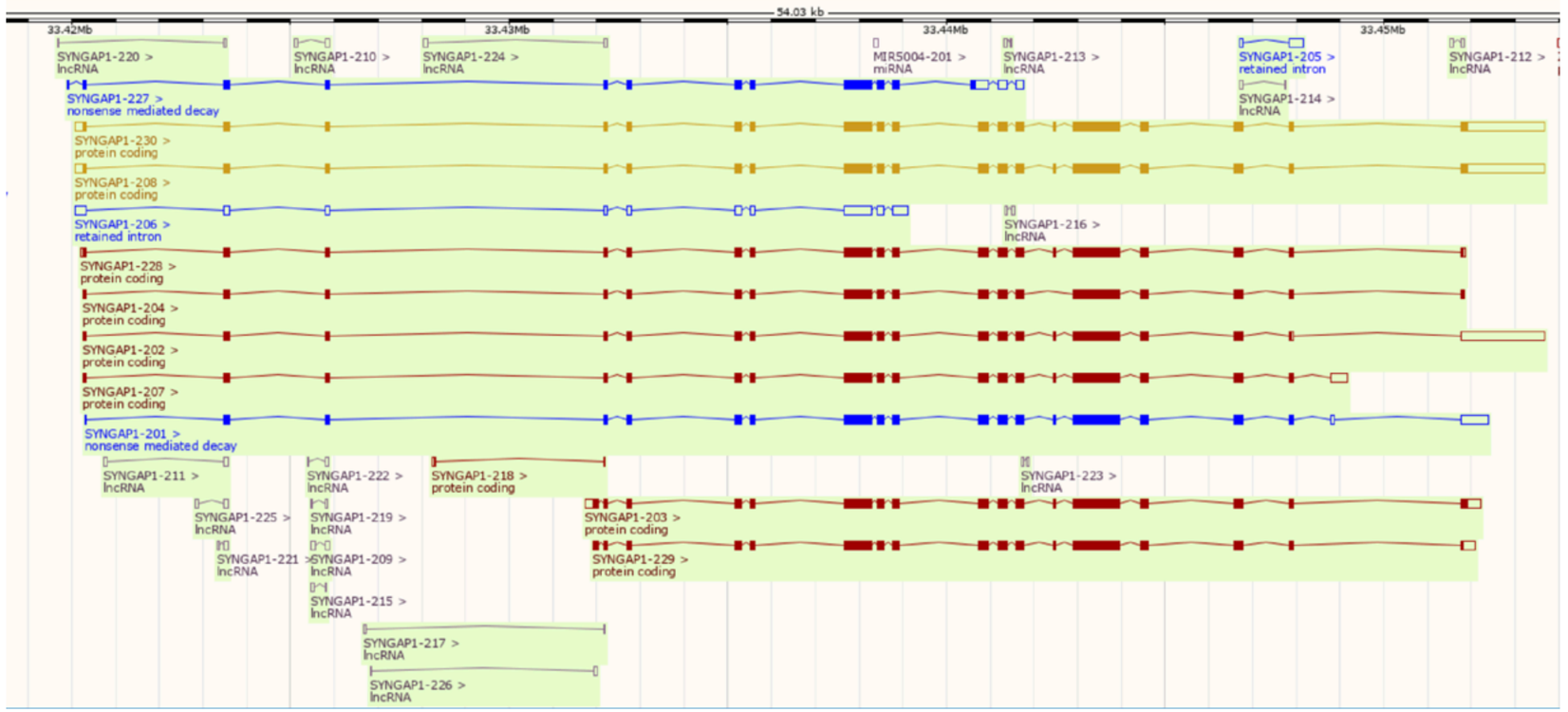
gene expression profiling

Affymetrix GeneChip Drosophila Genome2.0 Array

Critical Specifications	
Number of arrays in set	One
Number of transcripts	~18,500
Number of probe sets	18,880
Feature size	11 µm
Oligonucleotide probe length	25-mer
Probe pairs/sequence	14
Array format	100
Control sequences included:	
Hybridization controls:	*bioB, bioC, bioD* from *E. coli* and *cre* from P1 bacteriophage
Poly-A controls:	*dap, lys, phe, thr, trp* from *B. subtilis*
Housekeeping/Control genes:	Actin (Actin 42A), GAPDH (Glyceraldehyde 3 phosphate dehydrogenase 2), Eif-4a (Eukaryotic initiation factor 4a)
Detection sensitivity	1:100,000*

*As measured by detection in comparative analysis between a complex target containing spiked control transcriptions and a complex target with no spikes.

Bioinformatics 1 (INFR11160)



Common Questions for RNA-seq Analysis

What transcripts can we recover from the data?

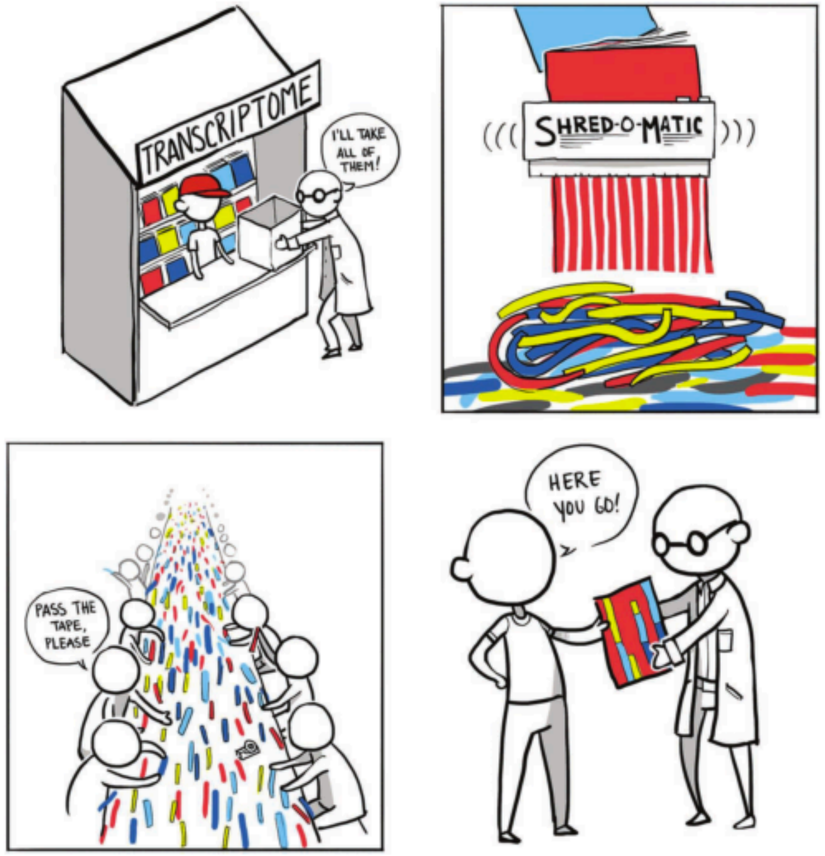
- transcriptome assembly

How much of each transcript is there?

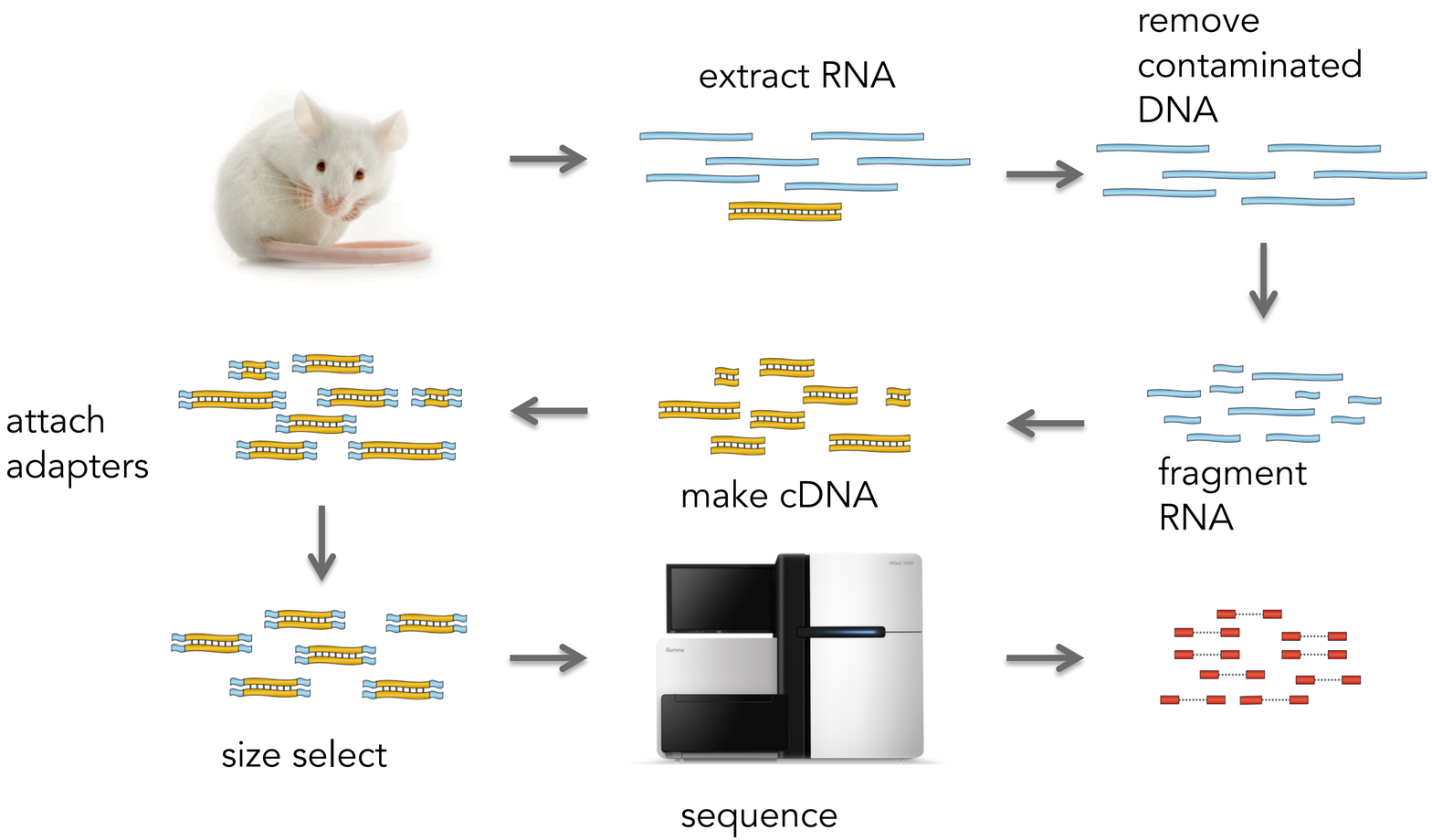
- quantification; normally relative

Are there statistically significant differences in the composition and amount of transcripts between the samples we are comparing?

- differential expression analysis (gene, transcript(s), isoform(s))



Korf, Nature Methods (2013).


extract RNA
remove contaminated DNA
fragment RNA
make cDNA
attach adapters
size select
sequence

RNA-seq Data

- Millions of short (50-150bp) sequence “reads”
- Single- or paired-end
- Stranded or unstranded

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

DNA read

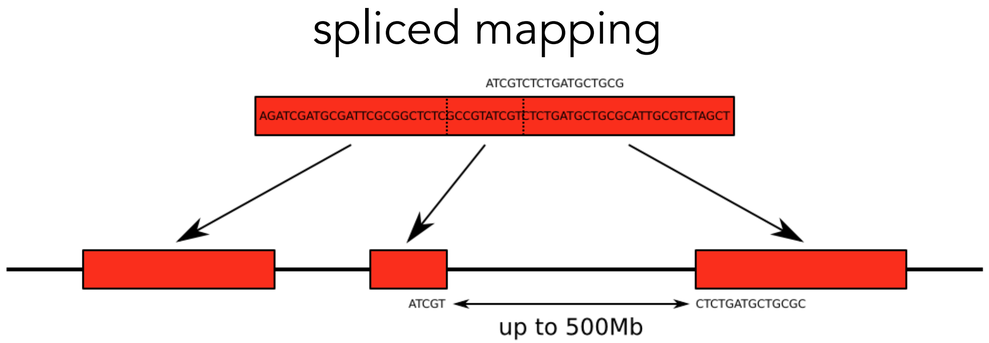
Base quality score

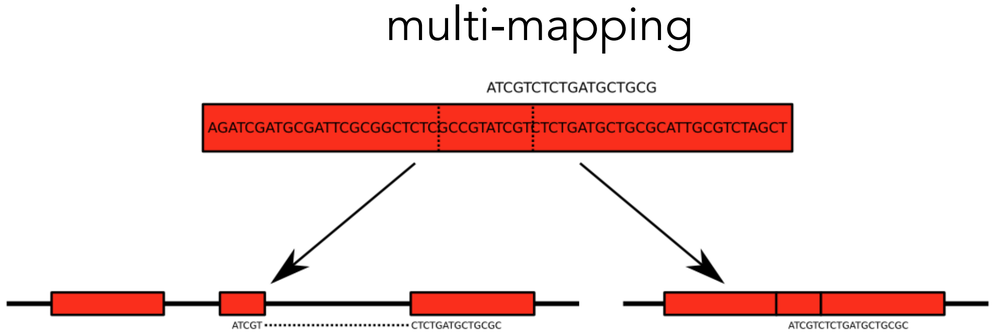
Read mapping

- Find the genomic location of millions of sequence reads
- Sequence alignment problem for each read
- Why is this hard?

Difficulties mapping RNA-seq reads

- Errors in the reads
- mismatches
- insertions & deletions
- Errors in the reference sequence
- Genetic variation
- Spliced mapping
- Multi-mapping
- Efficiency vs. Accuracy



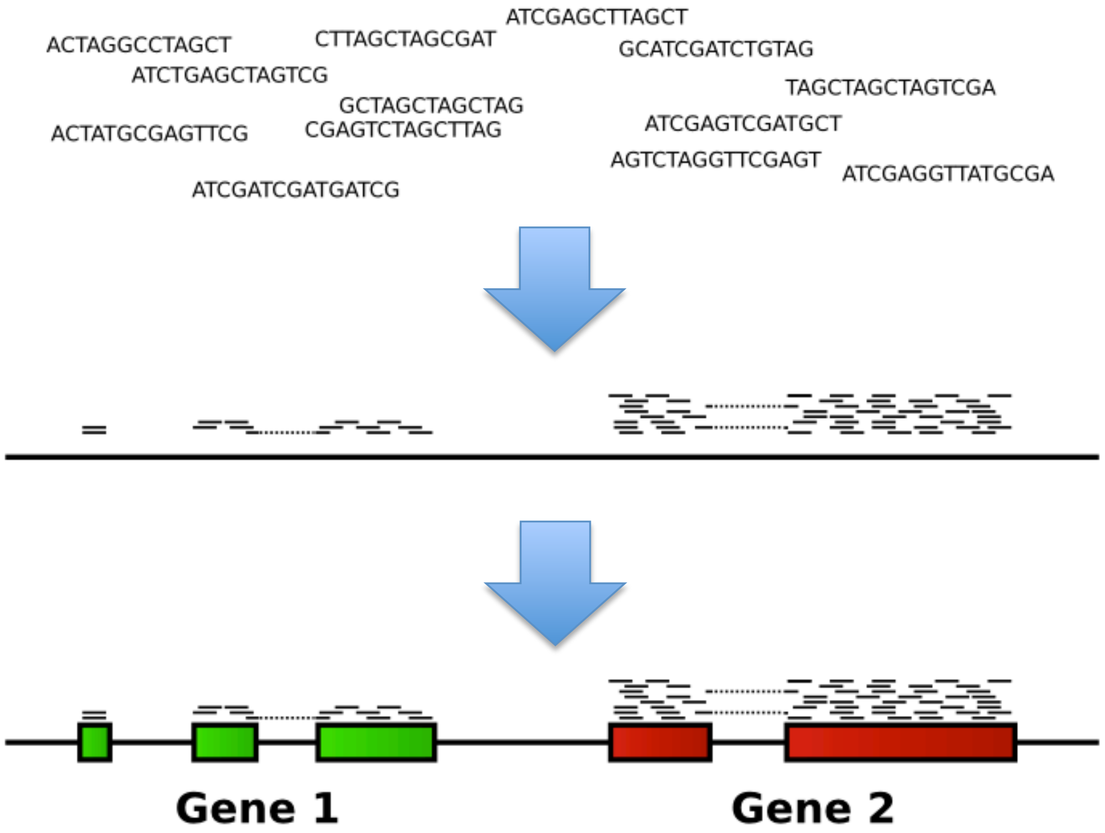


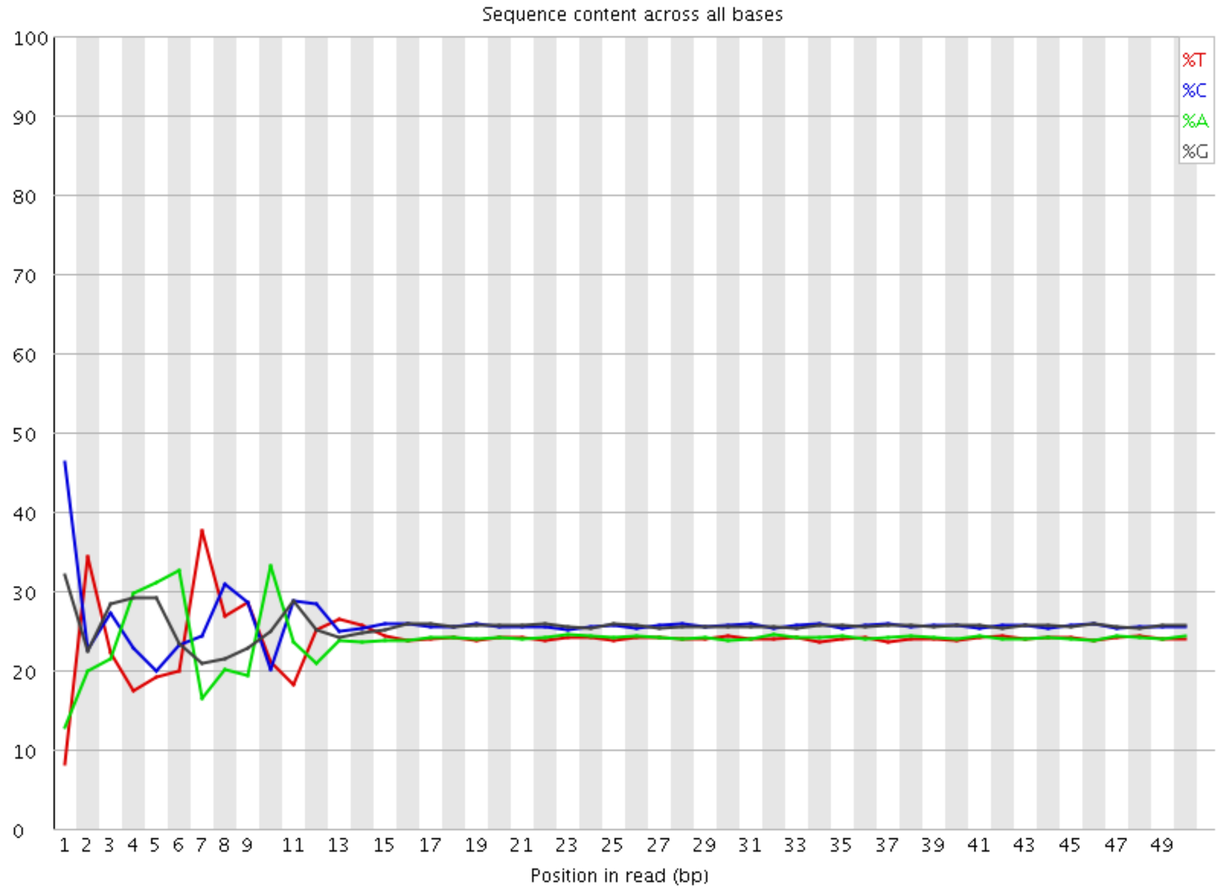
STAR: ultrafast universal RNA-seq Aligner

- Builds genome index (uncompressed suffix array) and for each read
- finds “maximal mappable” seeds
- clusters and stitches seeds together via local alignment
- Mapping speed is traded against RAM
- ~30Gb RAM for aligning to the human genome
- 550 million paired-end reads/hour

Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR

Current Protocols in Bioinformatics. 2015 Sep 3;51:11.14.1-19.


ATCGAGCTTAGCT
ACTAGGCCTAGCT
CTTAGCTAGCGAT
GCATCGATCTGTAG
ATCTGAGCTAGTCG
TAGCTAGCTAGTCGA
GCTAGCTAGCTAG
ATCGAGTCGATGCT
ACTATGCGAGTTCG
CGAGTCTAGCTTAG
AGTCTAGGTTCGAGT
ATCGAGGTTATGCGA
ATCGATCGATGATCG
Gene 1
Gene 2


Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
%T
%C
%A
%G
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
Position in read (bp)

Read ambiguity

- Sources of read ambiguity
- Mapping uncertainty: errors in sequence reads, sequence similarity
- Isoforms: different gene transcripts sharing sequence lead to uncertainty in transcript of origin
- Computational methods for transcriptome quantification attempt to account for sequence bias, read ambiguity, etc.

Huge Number of quantification tools

- Salmon, Kallisto
- Sailfish, RSEM, eXpress, Cufflinks
- MISO, Scripture, FluxCapacitor
- SLIDE, PSGInfer, IsoLasso, SpliceTrap
- ERANGE, IsoEM, iReckon, DRUT, rQuant
- etc. etc. etc.

http://seqanswers.com

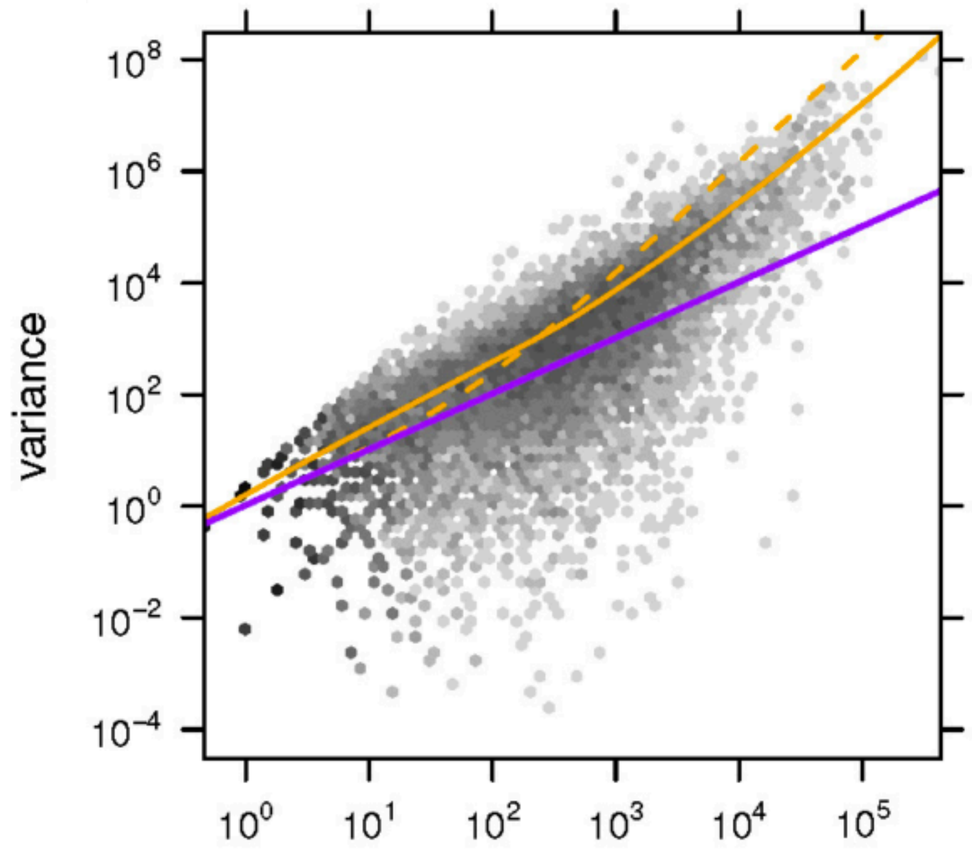
Differential expression

- Identify genes (or isoforms) expressed in significantly different quantities between conditions
- Treatment versus control, Diseased versus healthy, Different tissues, developmental stages….
- More complicated designs might include experimental factors with multiple levels
- How can we tell if a gene or isoform is differentially expressed?
- Need to understand how abundance varies within sample groups
- Need biological replicates
- RNA-seq replicates are expensive
- Need statistical methods with enough detection power to perform well with small sample sizes
- Make assumptions about the distribution of data

Feature Count Distributions

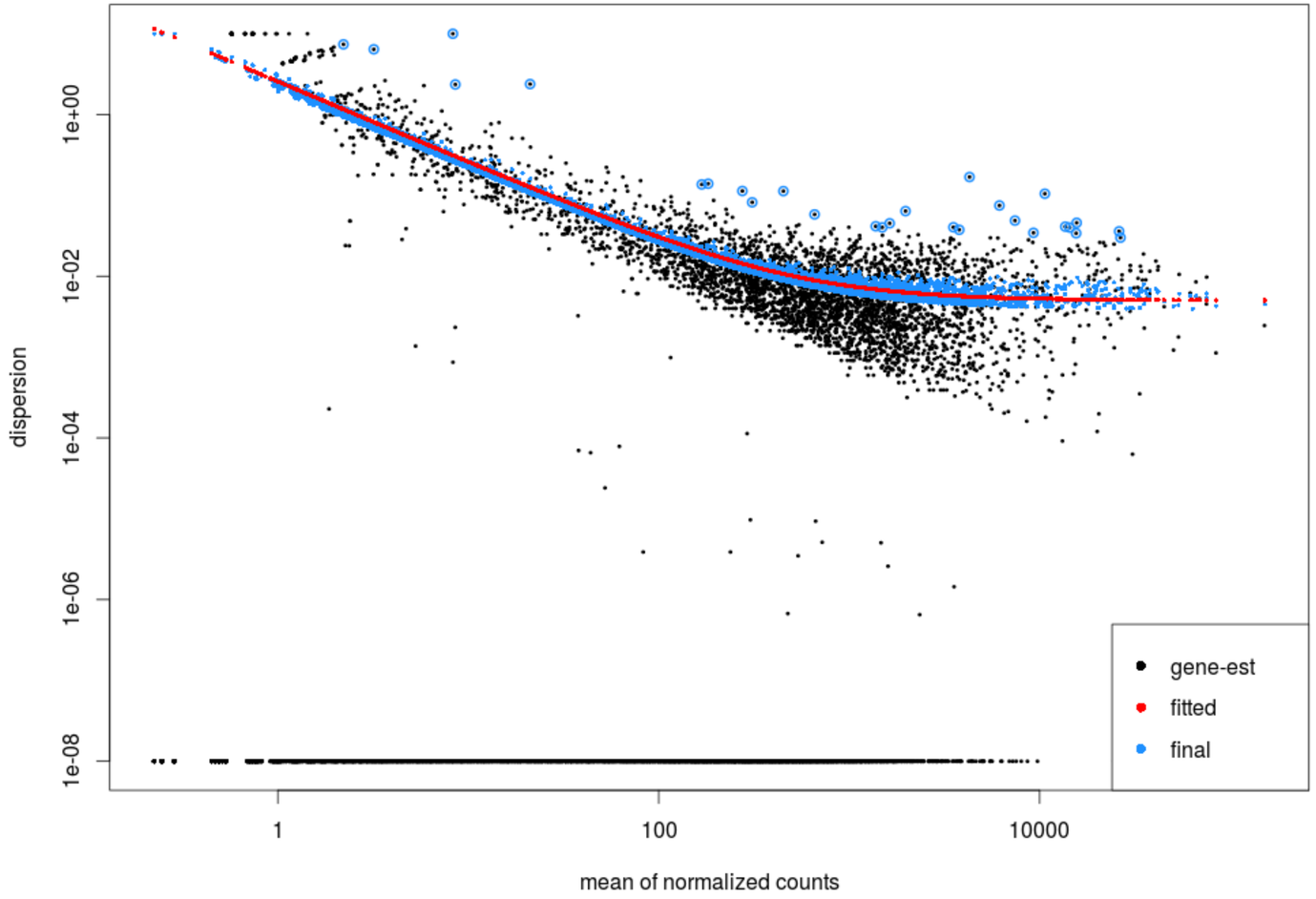
- Common approach is to count reads mapping to “features”
- Model feature counts by a particular distribution
- Identify features whose count difference between experimental conditions exceeds the variability predicted by the distribution
- calculate p-values

Negative Binomial Distribution



After Anders *et al*. (2010).

Dispersion


dispersion
1e+00
1e-02
1e-04
1e-06
1e-08
1
100
10000
mean of normalized counts
gene-est
fitted
final

RNA-seq Analysis in R

- using R locally or on a server
- RStudio (free) - https://rstudio.com/products/rstudio/
- R - https://www.r-project.org
- R/Bioconductor - https://bioconductor.org
- Overview of both the experimental and analytical sides of RNA-seq
- RNA-seqlopedia: http://rnaseq.uoregon.edu
- DESeq2 R package - https://bioconductor.org/packages/release/bioc/html/DESeq2.html
- pasilla R package - http://www.bioconductor.org/packages/ release/data/experiment/html/pasilla.html

Conservation of RNA regulation

- RNA-seq data from a paper studying the conservation of gene regulation between Drosophila (fruit fly) and mammals
- Comparing a knock-down of the Drosophila "*pasilla*" gene
- "*pasilla*" regulates alternative splicing by binding to RNA
- The gene was knocked-down using RNA-interference, this reduced the amount of protein produced (translation)
- Data is a mix of single-end and paired-end sequencing

Microarrays vs. RNA-Seq

Technology	Microarrays	RNA-seq
Reliance on genomic sequence	Yes	Sometimes
Background noise	High	Low
Dynamic range for expression quantification	Up to a few-hundredfold	>8,000 fold
Able to discover novel genes and isoforms	No	Yes
Bioinformatics	Well developed	Well Developed (now!)

After Wang *et al*. (2009).

Summary

Analysis of gene expression data from microarray and RNA-sequencing requires complex multi-method approaches

Knowledge of the target genomes and gene-models can be critical in successfully quantifying expression and changes in expression between conditions

Microarrays and RNA-sequencing are still widely used and have specific advantages & disadvantages

In order to compare samples, experiments & conditions appropriate pre-processing needs to be performed such as:

- Quality control
- Batch correction
- Outlier detection
- Normalisation
- Statistical evaluation (including multiple testing correction)

Initial downstream functional analysis is carried out using enrichment tests against annotations associated with genes as defined by the outcomes of the differential expression analysis