Microarray Technology

Dr Stevie Pederson (They/Them) stevie.pederson@thekids.org.au

Black Ochre Data Labs
The Kids Research Institute Australia

Acknowledgement Of Country

I'd like to acknowledge the Kaurna people as the traditional owners and custodians of the land we know today as the Adelaide Plains, where I live & work.

I also acknowledge the deep feelings of attachment and relationship of the Kaurna people to their place.

I pay my respects to the cultural authority of Aboriginal and Torres Strait Islander peoples from other areas of Australia, and pay my respects to Elders past, present and emerging, and acknowledge any Aboriginal Australians who may be with us today

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Technology Development

Microarray Technology

- Marked the birth of the modern transcriptomics era (~1996)
  - Thousands of genes analysed simultaneously
  - Microarrays are still in use!
- Developed alongside Human Genome Project (+ other organisms)
  - Reference genome was nearing completion
  - Gene/Transcript sequences available at scale
- Focus was quantitative analysis $\implies$ *differential gene expression*

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Microarray Technology

- Compute resources were increasing rapidly
  - CPU multi-threading becoming commonplace
  - `python 2.0` released in 2000
  - `R` v1.0.0 released in 2000
  - Bioconductor release 1.0 in 2002
  - High Performance Computing (HPC) becoming accessible

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Microarray Technology

- Probes complementary to target sequence positioned on array
- RNA $\rightarrow$ fluorescently labelled cDNA

1. Labelled cDNA binds to probe

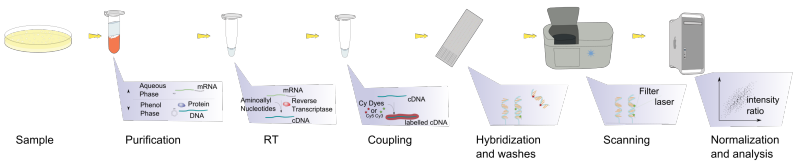


Figure 1: Image courtesy of Squidonius, Public domain, via Wikimedia Commons

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Microarray Technology

- Probes complementary to target sequence positioned on array
- RNA $\rightarrow$ fluorescently labelled cDNA

1. Labelled cDNA binds to probe
2. Microarray excited with laser

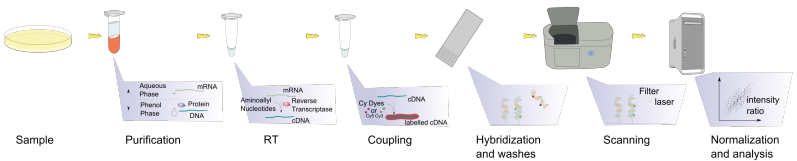


Figure 1: Image courtesy of Squidonius, Public domain, via Wikimedia Commons

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Microarray Technology

- Probes complementary to target sequence positioned on array
- RNA $\rightarrow$ fluorescently labelled cDNA

1. Labelled cDNA binds to probe
2. Microarray excited with laser
3. Fluorescence at a given probe $\propto$ target RNA abundance

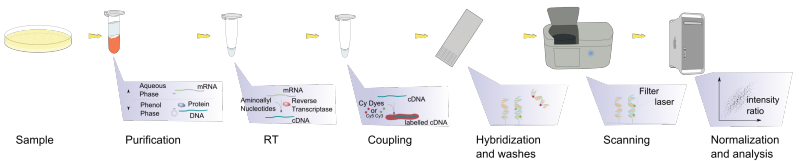


Figure 1: Image courtesy of Squidonius, Public domain, via Wikimedia Commons

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Two Colour Arrays

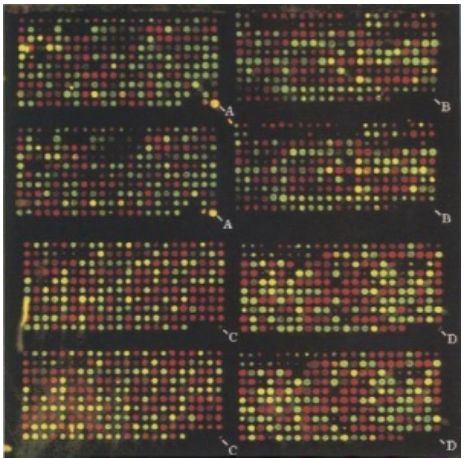


Figure 2: Section of two-colour array taken from Shalon, Smith, and Brown (1996)

- Two colour microarrays were printed microscope slides
- Known probe sequences were *printed* to the surface in defined locations
  - 60-75mer oligonucleotide probes
  - Highly customisable by project
- Two samples per array
  - Samples labelled with Cy5 (Red) or Cy3 (Green)
- Scanned at 570nm (Cy3) and 670nm (Cy5)

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

MA Plots

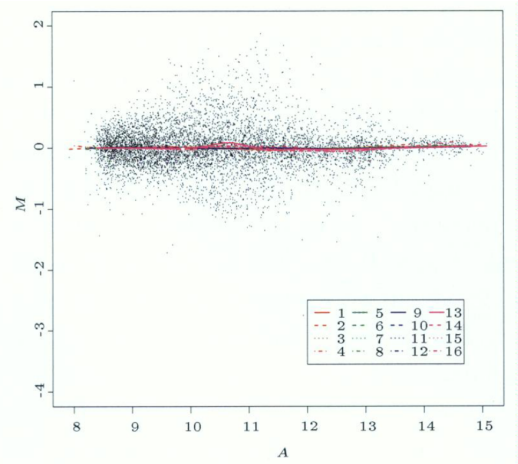


Figure 3: Image from Dudoit et al. (2002)

- Mean of Differences $M = \log_2(\frac{R}{G}) = \log_2(R) - \log_2(G)$
- Average Signal $A = \frac{1}{2}\log_2(RG) = \frac{\log_2(R)+\log_2(G)}{2}$
- Assess bias within and between arrays
- Also to show DE genes
- Term "MA Plot" still used in RNA-Seq despite no connection to formula

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Single Channel Arrays

Figure 4: No author known. Schutz assumed based on copyright claims. CC BY-SA 3.0, via Wikimedia Commons

- Affymetrix Arrays became dominant
  - Factory manufactured
- Standardised layout for each organism
- Single sample per array
  - Only scanned at one frequency $\implies$ no dye bias
- More genes/array
- 25mer probes targeting 3' end of transcript
  - Captured only intact transcripts

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Single Channel Arrays

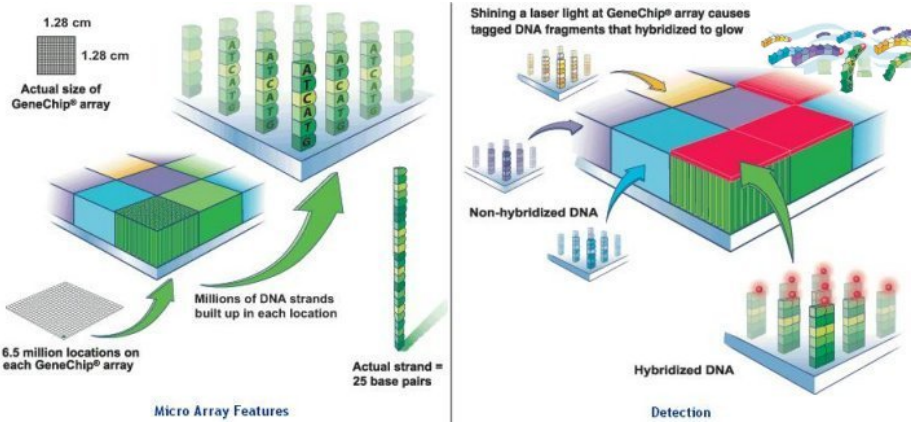


Figure 5: The basic methodology underpinning Affymetrix array design. Source Affymetrix

KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY OF ADELAIDE

3' Arrays

- Each 3' exon targeted by 11 unique 25mer probes $\implies$ *a probeset*
- Possible to detect different transcripts only if 3' exons differ
- *Perfect Match* (PM) probes $\implies$ exactly matches target sequence

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

3' Arrays

- Each 3' exon targeted by 11 unique 25mer probes $\implies$ *a probeset*
- Possible to detect different transcripts only if 3' exons differ
- *Perfect Match* (PM) probes $\implies$ exactly matches target sequence
  - Known to capture off-target signal $\implies$ non-specific binding (NSB)

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

3' Arrays

- Each 3' exon targeted by 11 unique 25mer probes $\implies$ *a probeset*
- Possible to detect different transcripts only if 3' exons differ
- *Perfect Match* (PM) probes $\implies$ exactly matches target sequence
  - Known to capture off-target signal $\implies$ non-specific binding (NSB)
- 3' arrays include paired *mismatch* probes (MM) with a change at the 13th position

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

3' Arrays

- Each 3' exon targeted by 11 unique 25mer probes $\implies$ *a probeset*
- Possible to detect different transcripts only if 3' exons differ
- *Perfect Match* (PM) probes $\implies$ exactly matches target sequence
  - Known to capture off-target signal $\implies$ non-specific binding (NSB)
- 3' arrays include paired *mismatch* probes (MM) with a change at the 13th position
  - Literally half the array

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

3' Arrays

- Each 3' exon targeted by 11 unique 25mer probes $\implies$ *a probeset*
- Possible to detect different transcripts only if 3' exons differ
- *Perfect Match* (PM) probes $\implies$ exactly matches target sequence
  - Known to capture off-target signal $\implies$ non-specific binding (NSB)
- 3' arrays include paired *mismatch* probes (MM) with a change at the 13th position
  - Literally half the array
  - Intended to quantify NSB properties of each probe

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

3' Arrays

- Each 3' exon targeted by 11 unique 25mer probes $\implies$ *a probeset*
- Possible to detect different transcripts only if 3' exons differ
- *Perfect Match* (PM) probes $\implies$ exactly matches target sequence
  - Known to capture off-target signal $\implies$ non-specific binding (NSB)
- 3' arrays include paired *mismatch* probes (MM) with a change at the 13th position
  - Literally half the array
  - Intended to quantify NSB properties of each probe
  - Sometimes returned more signal than PM probes

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Other Array Types

- 3' Arrays replaced by whole transcript (WT) arrays
  - Exon/Gene Arrays
  - Maximum of 4 probes/exon
  - Multiple probes target missing exons
- Illumina Bead arrays
  - 65-mer probes
  - Used barcoded beads instead of set probe locations

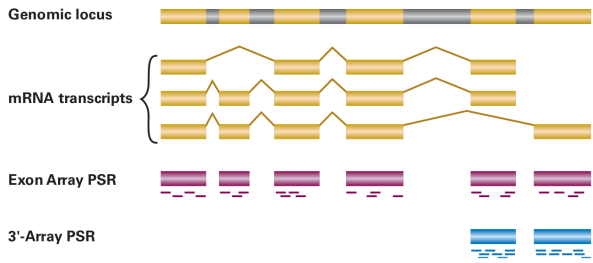


Figure 6: Source: Affymetrix Technical Note

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Differential Gene Expression

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Motivation

- The aim is to detect *differentially expressed* (DE) genes
- Classic design is Control samples vs Treated samples e.g.
  - Resting T cells vs Stimulated T cells
  - T47D + Estrogen vs T47D + Estrogen + Testosterone
- What genes are DE between conditions
- Need replicates for each condition $\implies$ t-test for each gene
  - Anything with $n < 3$ per condition is unacceptable
  - What values do we use for t-test?

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Quantile Normalisation

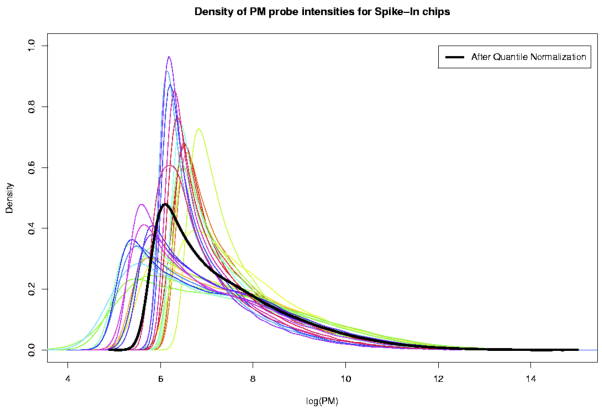


Figure 7: Example of raw $\log_2$PM probe intensities. Taken from Bolstad et al. (2003)

- Need to summarise probes to a probeset
  - Probes can be very noisy
  - BG signal + inconsistent binding
- Individual arrays give different fluorescence distributions $\Longrightarrow$ technical noise
  - Normalisation between arrays

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Quantile Normalisation

- Quantile normalisation is perfect for arrays with probes and probesets
  - Normalise probe-level signal, but estimate gene expression at the probeset level
  - Smooths out any normalisation artefacts

1. Select the lowest signal probe on each array$\longrightarrow$ Likely to be a different probe on each array

- Effectively randomises noise
- Leads to arrays with identical distributions

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Quantile Normalisation

- Quantile normalisation is perfect for arrays with probes and probesets
  - Normalise probe-level signal, but estimate gene expression at the probeset level
  - Smooths out any normalisation artefacts

1. Select the lowest signal probe on each array$\longrightarrow$ Likely to be a different probe on each array
2. Calculate the average signal across all arrays

- Effectively randomises noise
- Leads to arrays with identical distributions

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Quantile Normalisation

- Quantile normalisation is perfect for arrays with probes and probesets
  - Normalise probe-level signal, but estimate gene expression at the probeset level
  - Smooths out any normalisation artefacts

1. Select the lowest signal probe on each array$\longrightarrow$ Likely to be a different probe on each array
2. Calculate the average signal across all arrays
3. Give each of the selected probes the average signal

- Effectively randomises noise
- Leads to arrays with identical distributions

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Quantile Normalisation

- Quantile normalisation is perfect for arrays with probes and probesets
  - Normalise probe-level signal, but estimate gene expression at the probeset level
  - Smooths out any normalisation artefacts

1. Select the lowest signal probe on each array$\longrightarrow$ Likely to be a different probe on each array
2. Calculate the average signal across all arrays
3. Give each of the selected probes the average signal
4. Move to the next lowest signal probe until finished

- Effectively randomises noise
- Leads to arrays with identical distributions

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Quantile Normalisation

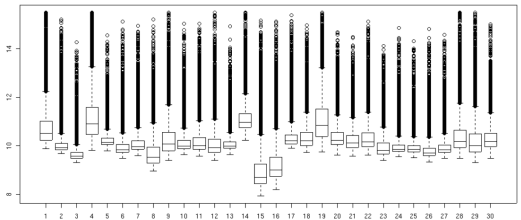


Figure 8: Before quantile normalisation

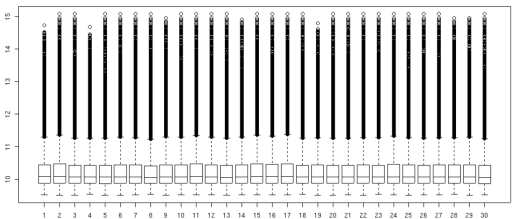


Figure 9: After quantile normalisation[1]

- Now we have identical distributions of signal across all arrays
- Equivalent to having identical amounts of source material (mRNA)
- Reduces technical noise across dataset $\implies$ more statistical power

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Background Correction

- Background Correction performed *simultaneously with estimation of signal*
- Robust Multichip Average (RMA) (Irizarry et al. 2003)
  - Estimates signal for each array (μ_i) $\implies$ used for differential expression
  - Model includes probe affinities (α_j)
  - Doesn't include MM probes
  - Fitted using robust statistics to reduce impact of outlier probes

$$\log_2 PM_{ij} = \mu_i + \alpha_j + \epsilon_{ij}$$

- Extended to GC-RMA (Wu et al. 2004) to include GC content of probes

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Linear Regression

$$\log_2 PM_{ij} = \mu_i + \alpha_j + \epsilon_{ij}$$

- This formula is a linear regression equation:
  - Fitted values for expression and probe affinity
  - An error term $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$
- The expression estimate μ_i is the array-specific intercept

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Linear Regression

$$\log_2 PM_{ij} = \mu_i + \alpha_j + \epsilon_{ij}$$

- This formula is a linear regression equation:
  - Fitted values for expression and probe affinity
  - An error term $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$
- The expression estimate μ_i is the array-specific intercept
- Probe affinities are common and consistent across arrays

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Differential Gene Expression

- For each gene $\rightarrow$ take the array-level estimates of gene expression (μ_i)
- Perform a t-test grouped by treatment group
  - Estimates variance (σ_g^2)
  - Estimate change in expression: $logFC = \beta_g$
- logFC is on $\log_2$ scale:
  - logFC $= 0 \implies$ unchanged expression
  - logFC $= 1 \implies$ doubling in RNA abundance

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Differential Gene Expression

- logFC estimates can also be from fitting *linear regression models*
  - Might use a covariate (e.g. age) alongside key predictor (i.e. diagnosis)

$$\mu_{ij} = \beta_0 + \beta_1 x_j + \epsilon_{ij}$$

- Commonly β_0 is the average expression
- The group variable x_j would be $x_1 = 0$ for control ($j = 1$) and $x_2 = 1$ for treated
  - β_0 becomes the *expected expression in control samples*
  - β_1 becomes the *expected difference in expression after treatment* $\implies$ logFC
- The variability around the predicted values is $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Differential Gene Expression

$$T = \frac{\beta_1}{\sigma/\sqrt{n}}$$

- Some estimates of variance will be over/under-estimates
  - $\hat{\sigma}_g$ too low $\implies T \uparrow \implies$ significant result where no change
  - $\hat{\sigma}_g$ too high $\implies T \downarrow \implies$ no significant result where there is change
- Variance estimates moderated by *taking distribution of σ across all genes*
  - Bayesian posterior estimate of variance $\implies$ moderated t-statistic (Smyth 2004)

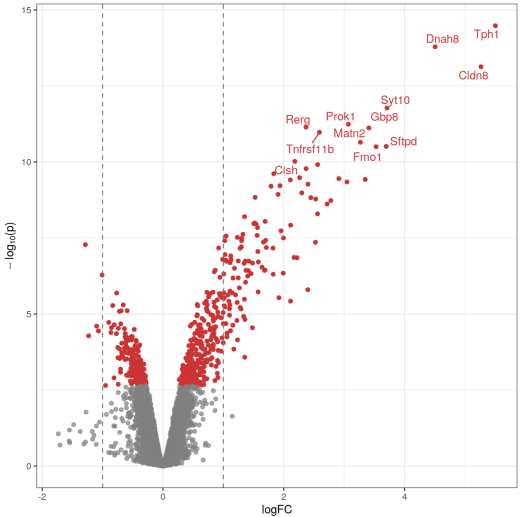
The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Differential Gene Expression

- t-test for each gene $\implies$ p-value for each gene
- Use False Discovery Rate (Benjamini and Hochberg 1995) to adjust p-values
  - Allows an expected α as a proportion of false discoveries
  - Usually $\alpha = 0.05 \implies \leq 5\%$ of genes accepted as DE are *false discoveries*
  - Assumes a bit of noise will be drowned out by true discoveries

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Presentation Of Results From DGE Analysis



- MA plots
  - Often used to check bias
  - Show logFC vs Average signal
- Volcano plots:
  - logFC vs Significance (e.g. -$\log_{10}$p)
  - Label genes as chosen

The KIDS RESEARCH INSTITUTE AUSTRALIA

BLACK OCHRE DATA LABS

THE UNIVERSITY of ADELAIDE

Closing Comments

- The microarray era defined the analytic approaches still taken
  - FDR-adjustment, statistical modelling, enrichment testing etc
  - Most methods use $\log_2$ transformed + normally-distributed data (t-tests)
  - Variance moderation
  - Bioconductor packages (e.g. `limma`) (Ritchie et al. 2015)

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

Additional Reading

- StatQuest: Linear Regression, Clearly Explained!!!
- A guide to creating design matrices for gene expression experiments

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE

References

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300. http://www.jstor.org/stable/2346101.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed. 2003. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." *Bioinformatics* 19 (2): 185–93.

Dudoit, Sandrine, Yee Hwa Yang, Matthew J. Callow, and Terence P. Speed. 2002. "STATISTICAL METHODS FOR IDENTIFYING DIFFERENTIALLY EXPRESSED GENES IN REPLICATED cDNA MICROARRAY EXPERIMENTS." *Statistica Sinica* 12 (1): 111–39. http://www.jstor.org/stable/24307038.

Irizarry, Rafael A., Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. 2003. "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data." *Biostatistics* 4 (2): 249–64. https://doi.org/10.1093/biostatistics/4.2.249.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Shalon, D, S J Smith, and P O Brown. 1996. "A DNA Microarray System for Analyzing Complex DNA Samples Using Two-Color Fluorescent Probe Hybridization." *Genome Research* 6 (7): 639–45. https://doi.org/10.1101/gr.6.7.639.

Smyth, G. K. 2004. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Stat Appl Genet Mol Biol* 3: Article3.

Wu, Zhijin, Rafael A Irizarry, Robert Gentleman, Francisco Martinez-Murillo, and Forrest Spencer. 2004. "A

The KIDS RESEARCH INSTITUTE AUSTRALIA
BLACK OCHRE DATA LABS
THE UNIVERSITY of ADELAIDE