Lecture 3: Microarray Technology

BIOINF3005/7160: Transcriptomics Applications

Dr Stephen Pederson

Bioinformatics Hub,
The University of Adelaide

March 23rd, 2020

THE UNIVERSITY of ADELAIDE

Microarray Technology

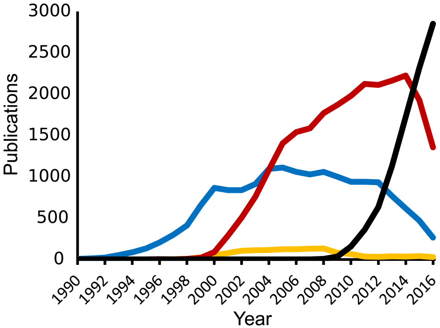
Two Colour Microarrays

Single Channel Microarrays

Whole Transcript Arrays

Microarray Technology

Microarrays



EST (blue); SAGE / CAGE (yellow); Microarrays (red); RNA Seq (black)[1]

[1] Rohan Lowe et al. "Transcriptomics technologies". In: *PLOS Computational Biology* 13.5 (May 2017), pp. 1–23. DOI: 10.1371/journal.pcbi.1005457. URL: https://doi.org/10.1371/journal.pcbi.1005457.

Microarrays

- Microarrays effectively ushered in the modern era of transcriptomics
- Purely interested in *relative abundances*
- Could measure expression levels for 1000's of genes simultaneously, for *the first time*
- Were essentially glass slides with probes affixed to them

Microarrays

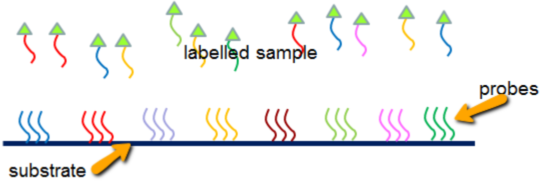
- Once again depends on reverse transcriptase for mRNA $\rightarrow$ cDNA
- **No reliance on Sanger Sequencing**
- Used probes (like a Northern blot) but the **cDNA is labelled and the probes are spatially fixed**
    - Probes must be designed beforehand
    - Probes are fixed to the array in *known locations*

Microarrays

1. Fluorescent labelling during mRNA conversion to cDNA
2. Complimentary probes bind target sequences (hybridisation)
3. Fluorescence detection at each probe

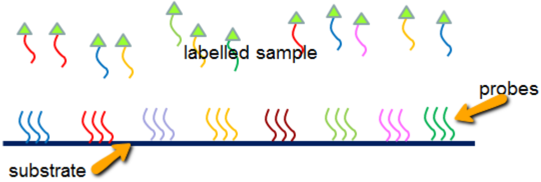
Fluoresence Intensity $\propto$ mRNA abundance

Microarrays

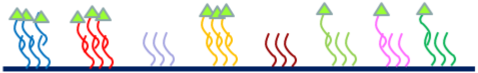


Source: `https://dev.stat.vmhost.psu.edu/stat555/book/export/html/635`

Microarrays

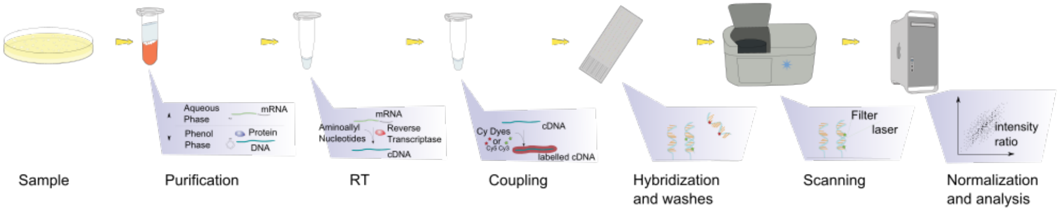


Highly abundant targets will yield more signal after hybridisation



Source: `https://dev.stat.vmhost.psu.edu/stat555/book/export/html/635`

Microarrays



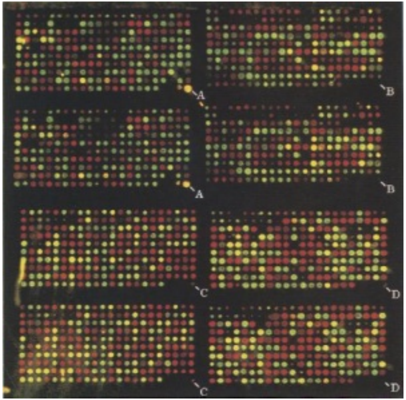
Source: `https://commons.wikimedia.org/wiki/File:Microarray_exp_horizontal.svg`

Two Colour Microarrays

Two Colour Microarrays

- Sometimes called "Low-Density Oligo Microarrays"
- Probes with known sequences are at known locations
    - Probes were 60-75mer complimentary cDNA
    - Originally printed in local facilities
- Samples are labelled with *either* Cy3 (Green @ 570nm) or Cy5 (Red @ 670nm)
- Two samples are hybridised to each array
    - Competitive hybridisation
    - Relative Red/Green intensities were of interest
    - Gave an estimate of logFC within each array

Two Colour Microarrays



A section of a two colour array[2]

[2] D Shalon, S J Smith, and P O Brown. "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization.". In: *Genome Research* 6.7 (1996), pp. 639–645. DOI: 10.1101/gr.6.7.639. eprint: http://genome.cshlp.org/content/6/7/639.full.pdf+html. URL: http://genome.cshlp.org/content/6/7/639.abstract.

Two Colour Microarrays

- Probes are "printed" to the array
  - Print tips can get clogged and be uneven
- Able to be customised for your own experiment
  - A mapping file for probe location to target sequence is required
- Both colours were scanned separately
  - One scan detects red only, the next detects green only
  - Each individual scan would have to be aligned spatially with the other

Two Colour Microarrays

- Spots were detected using astronomical software
  - Sizes were variable / irregular
- Detection of true signal above background (DABG)
  - Required "identified" (foreground) pixels and surrounding (background) pixels
  - Used surrounding pixels to estimate BG
  - Assumed BG was additive, e.g. $R = R_{bg} + R_{fg}$
- Dye bias was also noted $\implies$ experiments often used dye swaps
  - A sample from "group 1" might be labelled with red on one array, then labelled with green on the next

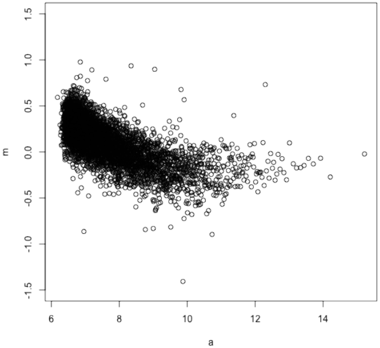
Two Colour Microarrays

- All intensities are transformed to the $\log_2$ scale
- Dye bias was checked using "MA Plots"
    - M was the *difference in intensity* across both channels
    - A was the *average intensity* across both channels

$$M = \log_2 R - \log_2 G$$

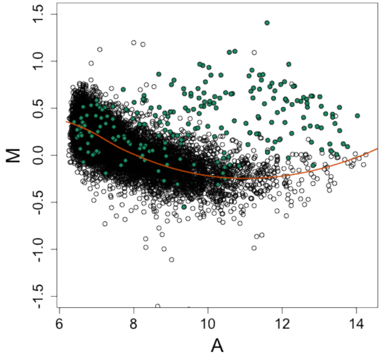
$$A = \frac{\log_2 R + \log_2 G}{2}$$

Two Colour Microarrays



Source: `https://genomicsclass.github.io/book/pages/normalization.html`

Two Colour Microarrays



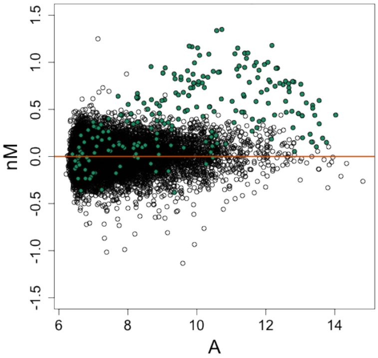
We can fit a **loess** curve through the data
(Here, spike-in controls are also highlighted)

Source: `https://genomicsclass.github.io/book/pages/normalization.html`

Two Colour Microarrays

- `loess`: Locally estimated scatterplot smoothing
  - We use a sliding window and fit a polynomial line
  - Usually polynomial of order 1 (linear) or 2 (quadratic)
- Once we have the loess curve: we subtract it from the data
  - Explicitly assumes that the bulk of the difference is bias, i.e. *most genes are not differentially expressed*
  - No modification to the A values, or any R/G intensities

Two Colour Microarrays



No more dye bias . . .

Source: `https://genomicsclass.github.io/book/pages/normalization.html`

Two Colour Microarrays

- We use these normalised M values across arrays to estimate logFC
- Dye-swap complications $\Longrightarrow$ *Experimental Design*
- Robust suite of statistical tools developed from here
- The `R` package `limma` set the standard

Single Channel Microarrays

Single Channel Microarrays

- Affymetrix 3' Arrays became the dominant technology (until RNA seq)
- Probes target the 3' end of transcripts $\implies$ intact transcripts
- Single channel (i.e. single colour) $\implies$ one sample per array
- $\sim$1,000,000 $\times$ 25-mer probes

Fluoresence Intensity $\propto$ mRNA abundance

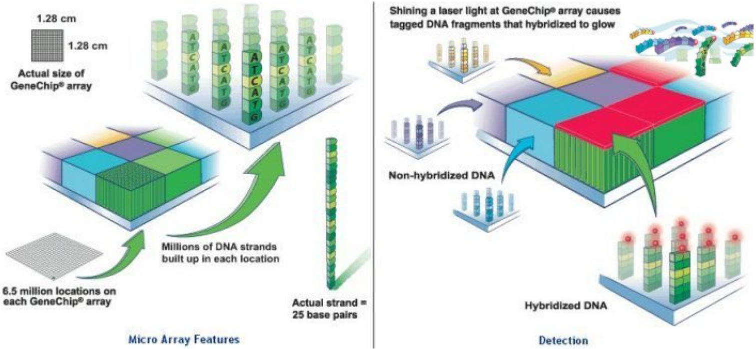
Single Channel Microarrays

Source: `https://commons.wikimedia.org/wiki/File:Affymetrix-microarray.jpg`

Single Channel Microarrays

- Manufacture used photolithography
- Far greater density of probes than two-colour arrays
    - Shorter probes but far more of them
- Also need a mapping file from location to probe sequence

Single Channel Microarrays



Source: `https://universe-review.ca/R11-16-DNAsequencing.htm`

Single Channel Microarrays

- Each 3' exon would be targeted by 11 unique probes
  - The set of **11 probes** would be collected together as a single **probeset**
- Alternate isoforms with different 3' exons could be detected easily as they would have distinct probesets
- Need a *Chip Description File* to map probes to array coordinates and probesets

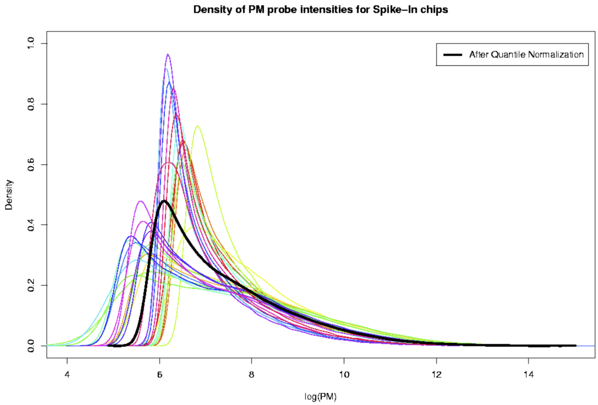
Single Channel Microarrays

Key Technical Issues:

1. Differences between **arrays**
   - Hybridisation afterfacts, cDNA/RNA concentration artefacts
2. Background Correction at the **probe** level
   - 25-mer probes $\implies$ *non-specific binding*
   - Optical Background
3. Expression estimates at the **probeset** level
   - Some probes *unresponsive*, other probes *promiscuous*
   - Do you just *average them*?

THE UNIVERSITY *of* ADELAIDE

Normalisation



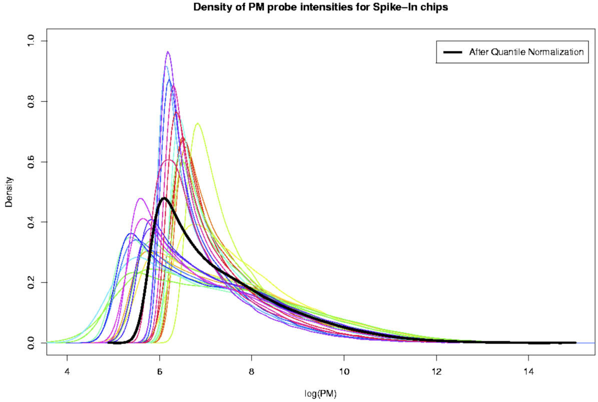
Taken from Bolstad et al., "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias"

Quantile Normalisation

1. Find the probe with the lowest intensity on each array
    - This will be from different probesets and unrelated to each other
2. Find the average intensity across these probes
3. Assign this value to each probe
4. Repeat for the probes with the next lowest intensity until done
5. All arrays now have the same intensity distribution

Under this approach, **we are adjusting the raw intensities**

Quantile Normalisation



Taken from Bolstad et al., "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias"

Background Correction

- Probes targeting 3' exons: Perfect Match (*PM*) probes
- Probes with middle base changes: MisMatch (*MM*) probes
- *MM* probes were expected to capture similar *NSB* behaviours to paired *PM* probe
  - Were often **brighter** than *PM* probes in pair
- Literally **half** of the array was *MM* probes

Background Correction

For a given PM/MM probe pair

$$PM = B + S$$

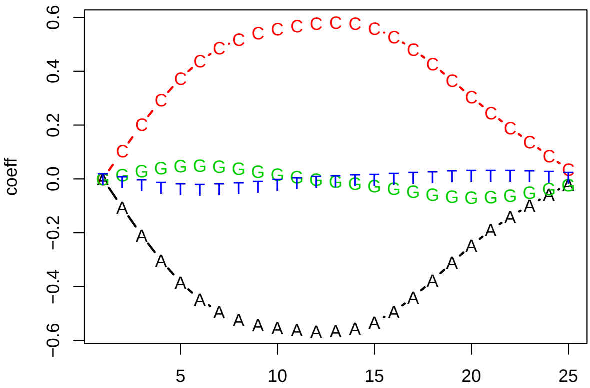
$$\text{but} \ldots MM \neq B$$

- How do we estimate S?
- $S \geq 0$

Background Correction

- Found $\hat{S} = E[S|PM]$ using a convolution of normal and exponential distributions (RMA)
- GC content and position in probe also impacted $NSB \implies GC\text{-}RMA$
- No need for the MM probe as a pair
    - MM probes still used in estimation of parameters

Background Correction



Probeset Summarisation

- Probes $j = 1, 2, \ldots, 11$ need to be combined (summarised) within a **probeset**
  - This gives the **gene-level expression estimates** for **each array**
  - Poor performing probes were generally poor on all arrays
  - Promiscuous probes were general similar on all arrays
- Probe-level modelling gave μ_i for each array i
  - The model was fit robustly $\implies$ outlier signal is down-weighted
  - Using $Y_{ij} = \log_2 \hat{S}_{ij}$:

$$Y_{ij} = \mu_i + \alpha_j + \varepsilon_{ij}$$

Now we have a single, gene-level estimate of expression for each array: $\hat{\mu}_i$

Analysis

- For each gene we take $\hat{\mu}_i$ and fit a linear model, conduct a t-test etc
- We will deal with the statistics very soon (FUN!)

Analysis

The basic process for single channel arrays:

1. Normalise for technical differences
2. Find probe-level estimates of *true signal*
3. Obtain gene-level estimates of signal
4. Statistical Analysis across all genes

Whole Transcript Arrays