Introduction to R

Microarray Analysis Group

- The University of Adelaide
  - Patty Solomon
  - Gary Glonek
  - Jonas Lloyd
- Hanson Centre for Cancer Research/IMVS
  - Anna Tsykin

  - Acknowledgements:
    - Terry Speed, UC Berkeley and WEHI
    - Greg Goodall, IMVS

Overview

- 9am–10.30am
  - Image analysis: role of Spot and R.
  - Why R, and how do you use it?
- 11am–12.30pm
  - Computing Practical I: An introduction to R.
- 1.30pm–2.30pm
  - Looking at some of R's more advanced features.
  - Why, and how, do we normalize microarray data?
- 2.50pm–4.00pm
  - Computing Practical II: `sma`, normalization; some statistical analysis.

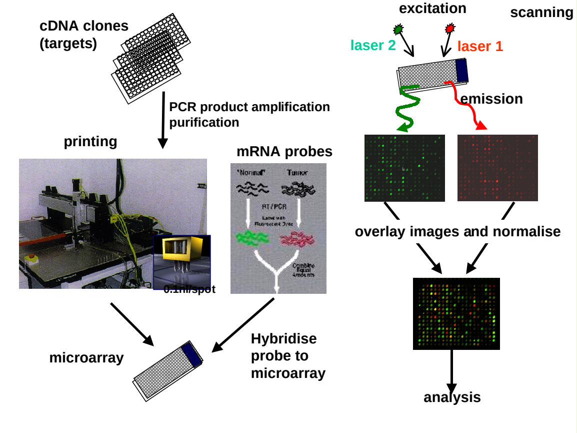
What are microarrays?

- *Powerful new technique for measuring the expression levels of many thousands of genes simultaneously.*
- There are many different technologies:
  - High-density nylon membrane arrays.
  - Short oligonucleotide arrays (Affymetrix).
  - Spotted long oligonucleotide arrays.
  - * Spotted cDNA arrays: Brown & Botstein (1999).
  - ...
- There are common themes to all these technologies.
  - *Challenges for statistics: design, analysis and interpretation of data from microarray experiments.*

cDNA Microarray Experiments

- Aim to identify genes which are *differentially expressed* in two or more cell populations.
  - Simplest experiments seek to identify changes in gene expression between
    - different tissue types e.g. normal vs tumour
    - different drugs e.g. treatment vs control
    - different locations within an organ: spatial effects.
- Time course experiments to monitor *expression profiles* over time.
- More complex experiments seek to identify *patterns in groups of genes*.
- *No single method of statistical analysis can be appropriate for all experiments.*

Microarray and scanner process



cDNA clones (targets)

PCR product amplification purification

printing

0.1nl/spot

mRNA probes

microarray

Hybridise probe to microarray

excitation

scanning

laser 2

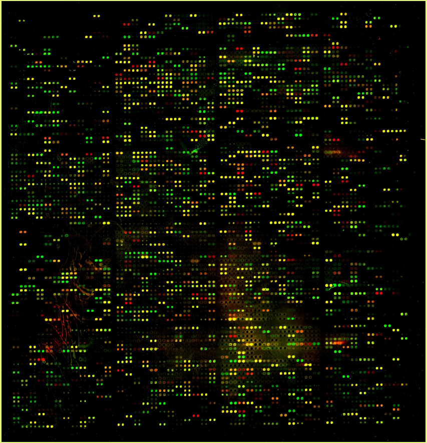
laser 1

emission

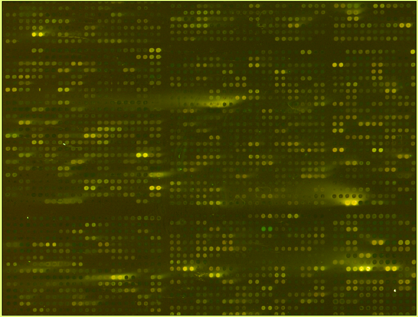
overlay images and normalise

analysis

Human clones with MCF-7 and Jurkatt probes

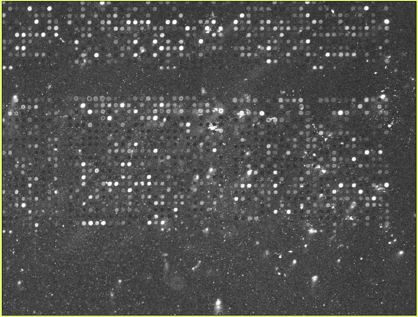


Comet tails, high background, spot overlap, ...



Woodcock 1_Ratio.tif

Precipitation, dust ....



Steps in image processing

1. Addressing: finding the spots.

◇ *Automating this part of the procedure permits high throughput analysis.*

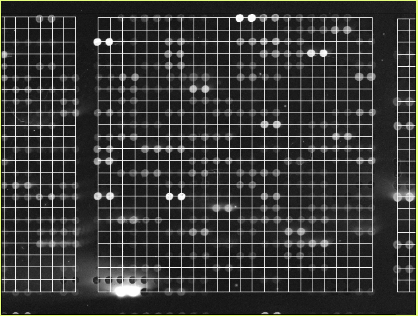
◇ Problems with automatic addressing include

- ⋆ translation of grids
- ⋆ overall position of the array in the image
- ⋆ mis-registration of the red and green channels
- ⋆ rotation of the array in the image
- ⋆ skew in the array.

(Forward)

◇ Most software systems provide both manual and automatic gridding procedures.

Gridding



Steps in image processing

2. Segmentation: finding the DNA.

- Each pixel is classified as **foreground** or background.
  - ◇ Fixed circle
    - ⋆ Limitations: one size does not fit all.
    - ⋆ *GenePix, ScanAlyze, QuantArray,* ...
  - ◇ Adaptive circle
    - ⋆ e.g. *GenePix, Dapple*
    - ⋆ Limitations: spots not circular; small spots.

◇ Adaptive shape

- ⋆ Watershed
- ⋆ **Seeded region growing** : regions grow outwards from seed points preferentially according to the difference between a pixel's value and the running mean of values in an adjoining region.
- ⋆ → Spot[1] is a prototype system for the image analysis of microarrays. It is built on R, which is a powerful environment for data analysis.
- ⋆ *SRG not available in widely used software.*

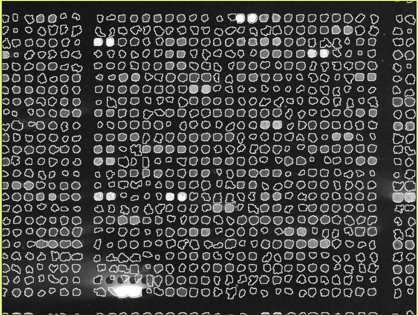
◇ Histogram segmentation

- ⋆ Adaptive threshold, e.g. *QuantArray*.

(Forward)

[1] Buckley et al, 2000

Seeded region growing



seg.overlay <memory>

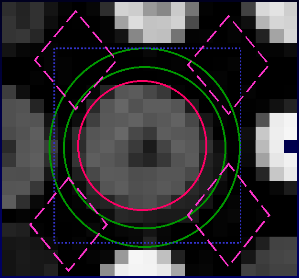
Steps in image processing

3. Information extraction: for each spot on the array, calculate

- ◇ Signal intensity pairs R, G, background Rb, Gb
  - ⋆ usually means, alternatively medians.
- ◇ Background is a big issue
  - ⋆ *'Bias–variance trade-off'.*
- ◇ Methods for background adjustment:
  - ⋆ Local.
  - ⋆ None.
  - ⋆ Constant.
  - ⋆ Morphological opening available in Spot generates an image of the estimated background intensity for the entire slide. (Forward)

Some local backgrounds

Single channel grey scale



We use something different again: a smaller, less variable value.

◇ *An intermediate approach looks best.*

- ⋆ SRG/morphological opening generates a data frame for each array → R.
- ⋆ Rows in data frame correspond to the spots and the columns to different spot statistics e.g. red and green foreground intensities, red and green background intensities for different background adjustment methods, spot area, ...
- ⋆ Spot handles batches of arrays.

Quality measures

- ◇ Control spots important
  - ⋆ **Spots:**
  - ⋆ Relative signal/background intensity.
  - ⋆ Variation in pixel values within each spot mask.
  - ⋆ Spot size (area in pixels).
  - ⋆ Circularity measure.
  - ⋆ Identification of ‘bad’ spots.
  - ⋆ **Arrays:**
  - ⋆ Correlation between spot intensities.
  - ⋆ Percentage of spots with no signals.
  - ⋆ Distribution of spot signal area.
  - ⋆ **Ratios (2 spots combined).**

Quantifying expression

For each spot on the array, calculate the background-adjusted intensities:

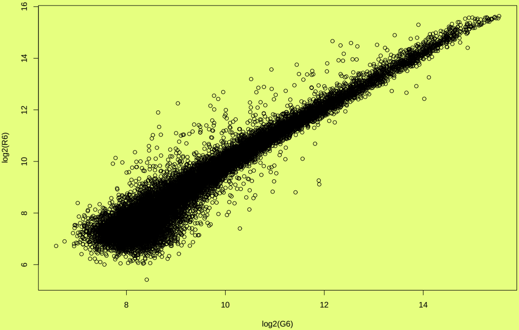
red intensity $R \rightarrow R - Rb$

green intensity $G \rightarrow G - Gb$

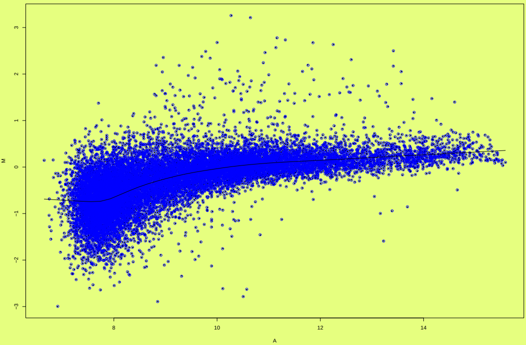
and combine them in the log base 2 ratio:

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

A typical plot

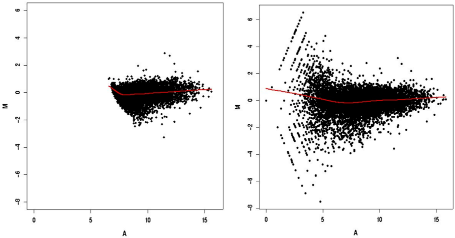


M is intensity dependent



$$M = \log_2 R/G \quad \text{vs} \quad A = \log_2(RG)/2$$

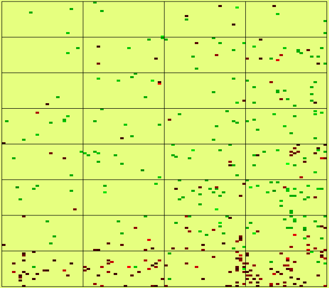
Background matters



From Spot

From GenePix

M can be spatially dependent



Print-tip effects

Some reading and web sites

- PO Brown, D Botstein. Exploring the new world of the genome with DNA microarrays. In *The Chipping Forecast* **21** 33–37. Supplement to *Nature Genetics*, January 1999.
- M. Eisen and P. Brown. *Methods in enzymology*. **303**, 1999.
- R. Ihaka & R. Gentleman. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314, 1996.
- M Schena (Ed.) *Microarray biochip technology*. Eaton, 2000.
- MJ Buckley. The Spot user's guide. CSIRO Mathematical and Information Sciences, August 2000. http://www.cmis.csiro.au/IAP/Spot.htm

- Microarray Analysis Group. The University of Adelaide http://maths.adelaide.edu.au/MAG
- The Chipping Forecast. *Supplement to Nature Genetics*, **21**, 1999.
- TP Speed. Preprints, information, software (sma) and Slides 5, 15, 21:

  www.stat.berkeley.edu/users/terry/zarray
- YH Yang, MJ Buckley, S Dudoit, TP Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics* **11**, 2002.