Microarray Data Analysis

Data quality assessment and normalization
Outline

• **Microarray data quality: diagnostic plots**
• **Pre-processing** cDNA chips
  – Image analysis
  – **Normalization**
  – Non-specific filtering
• Preprocessing Affymetrix chips
  – **Normalization**
  – Non-specific filtering
Microarray studies life cycle

- Biological question
- Experimental design
- Microarray experiment
- Quality Measurement
- Image analysis
- Normalization

Here we are

- Analysis (Estimation, Testing, Clustering, Discrimination)
- Biological verification and interpretation
Looking at microarray data
Diagnostic Plots

Was the experiment a success?
Diagnostic plots: *Look at the data!*

- Plots can be useful to
  - Check microarray data quality
  - Give hints on how to pre-process the data
  - Verify how the preprocessing has worked

- Must keep in mind that raw data differs considerably between cDNA or Affy
  - Although the “spirit” of the process is similar
  - Specific procedures or steps differ
Raw data (1): cDNA

- ... → Scanning → Images → .gpr files (ASCII)
- One .gpr file per chip, containing
  - One row per gene but many columns with
    - Intensity values for each channel
    - Summary values for intensities
    - Quality controls, such as FLAGS
- Intensity values are converted into a single expression matrix containing:
  - One column per chip with log(R/G) values
  - One row per gene (same rows as .gpr files)
- Gene information stored in the .gal file
Raw data (1): Examples

- See an example of a .gpr file
- See an example of a .gal file
- See an expression matrix
Raw data (2): Afffy

- ... → Scanning → Images → .CEL files (binary)
- One .CEL file per chip, containing
  - PM and MM values for each probe in the chip
  - Presence/Absence calls (one per probeset)
- Separate PM/MM values are converted into a single expression matrix containing:
  - One column per chip with absolute intensity values
  - One row per probeset
  - See an example of an expression matrix
- Gene information stored in the .cdf file
Diagnostic plots for cDNA chips
Red / Green overlay images

- Start by looking at the slides

Co-registration and overlay offers a quick visualization, providing information on:
- colour balance,
- uniformity of hybridization,
- spot uniformity,
- background, and
- artifacts such as:
  - dust or
  - scratches

Bad: high bg  Good: low bg
Scatterplots: always log*, always rotate

\[ \log_2 R \text{ vs } \log_2 G \]

\[ M = \log_2 R/G \text{ vs } A = \log_2 \sqrt{RG} \]

* Other transformations can provide improvement
MA-plot for spotted arrays (2 colors)

Mutant (MT)

Cy3/5-cDNA or aRNA

MT and WT intensity for each probe

Wild Type (WT)

Spot

M

Log$_2$(MT/WT)

A

Log$_2$(MT*WT) / 2
(signal strength)
Images with high background tend to have lower $\log_2(\text{signal/noise})$ ratios.
Spatial plots for slide backgrounds
Spatial plot of high intensity log ratios

If there are no spatial effects → high intensity spots should be uniformly distributed.

Top (black) and bottom (green) 5% of log ratios
Pin-group effects

Lowess lines through points from pin groups

Boxplots of log ratios by pin group
Highlighting pin group effects

Scatterplot and boxplots show a clear spatial bias.
Quality between slides
Alltogether:
Diagnostic plots for affy chips
Image plots for affymetrix chips

read from file: HIVControl4A.CEL.gz

read from file: HIVControl4A.CEL.gz

read from file: HIVControl4B.CEL.gz

read from file: HIVControl4B.CEL.gz
MA-plot for GeneChip arrays (1 color)

MT intensity for each probe set

WT intensity for each probe set

\[ M = \log_2(MT) - \log_2(WT) \] (signal strength)

\[ A = \frac{\log_2(MT \times WT)}{2} \] (signal strength)
Box plots
Density plots
Digestion plots

RNA digestion plot

Mean intensity shifted and scaled

5' --- 3'
Probe Number

Mean intensity shifted and scaled

5' --- 3'
Probe Number
Pre-processing cDNA chips

Non-specific Filtering
Normalization
Preprocessing: Filtering

- There may be errors during Hybridization and/or Scanning which yield bad spots
  - These are automatically flagged
- Many spots may show very low signals
  - Problems with spotting
  - No hybridization in this spot
- Bad spots may be removed from the analysis to avoid unnecessary noise
Filtering spots & adjust signals

• We may filter the data on intensity by excluding values where both the red and green channels are less than 100.
• We may set the value of an intensity to the minimum in the event only one of the two channel intensities is below the minimum of 100.
• In addition, we may use the flag column imported with the data, and exclude intensities with a flag value not equal to 1.
Must we filter the data?

- Filtering is intended to remove spots whose images or signals were wrong due to different possible reasons
  - Small quantity of cDNA in the array
  - Errors during the scanning process
- Some people prefer not to filter to avoid eliminating good spots unintentionally.
- *In case of doubt be conservative and reduce the filter operation to the minimum.*
We filter following the tutorial’s indications.

2998 genes pass the filtering criteria.
Preprocessing: normalization

• The word *normalization* describes techniques used to *suitably* transform the data before they are analysed.

• Goal is to correct for *systematic differences*
  – between samples on the same slide, or
  – between slides,

  *which do not represent true biological variation* between samples.
The origin of systematic differences

- Systematic differences may be due to …
  - Dye biases which vary with spot intensity,
  - Location on the array,
  - Plate origin,
  - Printing quality which may vary between
    - Pins
    - Time of printing
  - Scanning parameters,…
How to know if it’s necessary?

• Option 1: to perform self-self normalization
  – If we hybridize a sample with itself instead of sample vs control intensities should be the same in both channels
  – All deviations from this equality means there is systematic bias that needs correction

• Option 2: Look at diagnostic plots for dye, slide or spatial effects
Self-self hybridizations

False color overlay  Boxplots within pin-groups  Scatter (MA-)plots
Some non self-self hybridizations

From the NCI60 data set

Early Ngai lab, UC Berkeley

Early Goodman lab, UC Berkeley

Early PMCRI, Melbourne Australia
Normalization methods & issues

• Methods
  – Global adjustment
  – Intensity dependent normalization
  – Within print-tip group normalization
  – *And many other…*

• Selection of spots for normalization
Global normalization

- Based on a global adjustment
  \[ \log_2 \frac{R}{G} \to \log_2 \frac{R}{G} - c = \log_2 \frac{R}{(kg)} \]
- Choices for \( k \) or \( c = \log_2 k \) are
  - \( c \) = median or mean of log ratios for a particular gene set
    - All genes or control or housekeeping genes.
  - Total intensity normalization, where
    - \( K = \sum R_i / \sum G_i \).
Example: (Callow et al 2002)
Global median normalization.
Intensity-dependent normalization

- Run a line through the middle of the MA plot, shifting the M value of the pair \((A,M)\) by \(c=c(A)\), i.e.
  \[
  \log_2 \frac{R}{G} \rightarrow \log_2 \frac{R}{G} - c(A) \\
  = \log_2 \frac{R}{(k(A)G)}.
  \]
- One estimate of \(c(A)\) is made using the LOWESS function of Cleveland (1979): LOcally WEighted Scatterplot Smoothing.
Example: (Callow et al 2002) loess vs median normalization.
Example: (Callow et al 2002) Global median normalization.

- Global normalization performs a global correction but it cannot account for spatial effects
  - See next slide boxplots for the same situations in only one mouse, showing all sectors
Global normalisation does not correct spatial bias (print-tip-sectors)
Within print-tip group normalization

- To correct for spatial bias produced by hybridization artefacts or print-tip or plate effects during the construction of arrays.
- To correct for both print-tip and intensity-dependent bias perform LOWESS fits to the data within print-tip groups, i.e.

  $\text{Log}_2 \frac{R}{G} \rightarrow \log_2 \frac{R}{G} - c_i(A) = \log_2 \frac{R}{(k_i(A)G)}$, where $c_i(A)$ is the LOWESS fit to the MA-plot for the $i$th grid only.
Local print-tip normalisation corrects spatial bias (print-tip-sectors)
Normalization, which spots to use?

• LOWESS can be run through many different sets of points,
  – All genes on the array.
  – Constantly expressed genes (housekeeping).
  – Controls.
  – Spiked controls (genes from distant species).
  – Genomic DNA titration series.
  – Rank invariant set.
Strategies for selecting a set of spots for normalization

- Use of a global LOWESS approach can be justified by supposing that, when stratified by mRNA abundance,
  a) Only a minority of genes expected to be differentially expressed,
  b) Any differential expression is as likely to be up-regulation as down-regulation.
- Pin-group LOWESS requires stronger assumptions: that one of the above applies within each pin-group.
Example

• Sometimes it becomes obvious that all genes cannot be used for normalization – e.g. if most genes are expected to be differentially expressed

• The example shows normalizations performed on some arrays with only 37 different genes (replicated 16 times) belonging to a study with *E. Coli* mutant genes.
Normalization based on calibration spots
The effect of within-slide normalization
The effect of between-slide normalization

M box plot for all arrays. Normalization within slides only

M box plot for all arrays. Normalization within & between slide
Preprocessing affy chips
Preprocessing steps

• Computing expression values for each probe set requires 3-steps
  – Background correction
  – Normalization
  – Probe set summaries
Most popular approaches

• Affymetrix’s own MAS 5 or GCOS 1.0 algorithms
• RMA (Robust Multichip Analysis)
MAS

• **Background correction**
  
  \[ E_j = PM_j - MM_j^* \]

  where \( MM_j^* \) is chosen so that \( E_j \) is non-negative

• **Normalization**
  
  – Scale so that mean \( E_j \) is same for each chip

• **Probe Set Summary**
  
  \[ \log(\text{Signal Intensity}) = \text{TukeyBiweight}(\log E_j) \]
RMA- Background correction

• Ignore MM, fit model to PM
  – PM = Background \( (N(0, \sigma^2) + \text{Signal \ (Exp(\alpha)}) \)
RMA-Normalization

- Force the empirical distribution of probe intensities to be the same for every chip in an experiment
- The common distribution is obtained by averaging each *quantile* across chips: Quantile Normalization
One distribution for all arrays: the black curve
RMA: Probe set summary

- Robustly fit a two-way model yielding an estimate of log2(signal) for each probe set
- Fit may be by
  - median polish (quick) or by
  - Mestimation (slower but yields standard errors and good quality)
- RMA reduces variability without losing the ability to detect differential expression
MA plots before and after RMA
Summary

• Microarray experiments have many “hot spots” where errors or systematic biases can appear.
• Visual and numerical quality control should be performed.
• Usually intensities will require normalisation.
  – At least global or intensity dependent normalisation should be performed.
  – More sophisticated procedures rely on stronger assumptions → Must look for a balance.