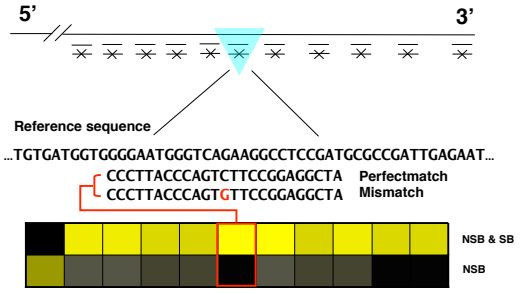
Preprocessing Affymetrix GeneChip Data

Credit for some of today's materials: Ben Bolstad, Leslie Cope, Laurent Gautier, Terry Speed and Zhijin Wu

Affymetrix GeneChip Design



Terminology

- Each gene or portion of a gene is represented by 11 to 20 oligonucleotides of 25 base-pairs.
- Reporter/Feautre/Probe: an oligonucleotide of 25 base-pairs, i.e., a 25-mer.
- Perfect match (PM): A 25-mer complementary to a reference sequence of interest (e.g., part of a gene).
- Mismatch (MM): same as PM but with a single homomeric base change for the middle (13^{th}) base (transversion purine <-> pyrimidine, G <->C, A <->T) .
- Probe-pair: a (PM,MM) pair.
- Probe-pair set: a collection of probe-pairs (11 to 20) related to a common gene or fraction of a gene.
- Affy ID: an identifier for a probe-pair set.
- The purpose of the MM probe design is to measure non-specific binding and background noise.

Affymetrix files

- Main software from Affymetrix company *MicroArray Suite* - *MAS*, now version 5.
- `DAT` file: Image file, ~10^7 pixels, ~50 MB.
- `CEL` file: Cell intensity file, probe level PM and MM values.
- `CDF` file: Chip Description File. Describes which probes go in which probe sets and the location of probe-pair sets (genes, gene fragments, ESTs).

Expression Measures

- 10-20K genes represented by 11-20 pairs of probe intensities (PM & MM)
- Obtain expression measure for each gene on each array by summarizing these pairs
- We already discussed background adjustment and normalization. We assume this has been done.
- There are many methods

Data and notation

- PM_{ijg}, MM_{ijg} = Intensity for perfect match and mismatch probe in cell *j* for gene *g* in chip *i*.
  - $i = 1,\ldots, n$ -- from one to hundreds of chips;
  - $j = 1,\ldots, J$ -- usually 11 or 20 probe pairs;
  - $g = 1,\ldots, G$ -- between 8,000 and 20,000 probe sets.
- Task: summarize for each probe set the probe level data, i.e., PM and MM pairs, into a single expression measure.
- Expression measures may then be compared within or between chips for detecting differential expression.

MAS 4.0

- **GeneChip® MAS 4.0 software used AvDiff up until 2001**

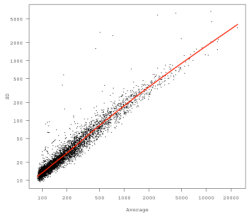
$$AvDiff = \frac{1}{|\mathrm{A}|}\sum_{j\in \mathrm{A}}(PM_j - MM_j)$$

where A is a set of "suitable" pairs, e.g., pairs with $d_j = PM_j - MM_j$ within 3 SDs of the average of $d_{(2)}, \ldots, d_{(J-1)}$

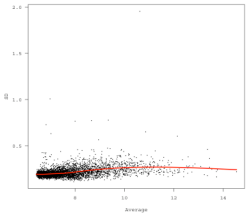
- **Obvious problems:**
  - **Negative values**
  - **No log scale**

Why use log?

Original scale



Log scale

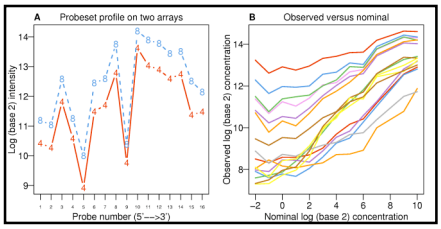


Li and Wong's observations

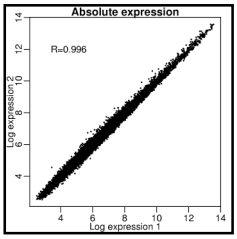
- **There is a large probe effect**
- **There are outliers that are only noticed when looking across arrays**
- **Non- linear normalization needed (discussed in previous lecture)**

PNAS vol. 98. no. 1, **31-36**

Probe effect

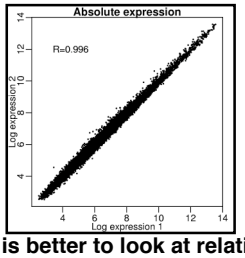


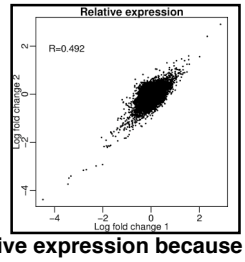
Probe effect makes correlation deceiving



Correlation for absolute expression of replicates looks great! But…

Probe effect makes correlation deceiving





- **It is better to look at relative expression because probe effect is somewhat cancelled out.**
- **Later we will see that we can take advantage of probe effect to find outlier probes.**

Li & Wong

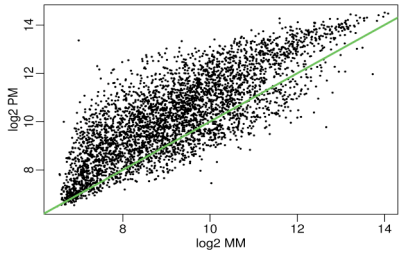
- **Li & Wong (2001) fit a model for each probe set, i.e., gene**

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \propto N(0, \sigma^2)$$

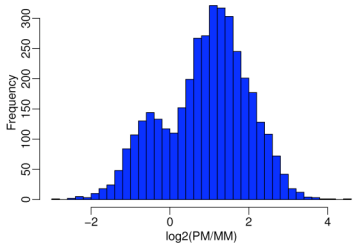
where

  - **θ_i: model based expression index (MBEI),**
  - **ϕ_j: probe sensitivity index.**
- **Maximun likelihood estimate of MBEI is used as expression measure for the gene in chip *i*.**
- **Non-linear normalization used**
- **Ad-hoc procedure used to remove outliers**
- **Need at least 10 or 20 chips**

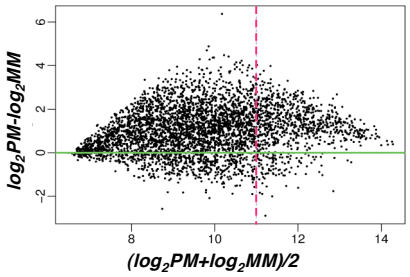
There is one more reason why PM-MM is undesirable



Especially for large PM

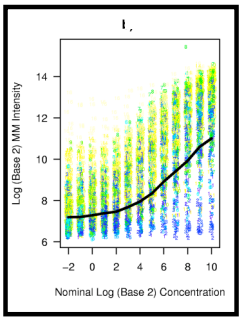


We see bimodality



Two more problems with MM

- MM detect signal
- MM cost $$$



MAS 5.0

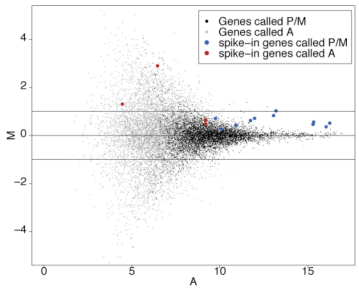
- Current version, MAS 5.0, uses Signal

$$signal = \text{Tukey Biweight}\{\log(PM_j - MM_j^*)\}$$

- Notice now log is used
- But what about negative PM-MM ?

- MM* is a new version of MM that is never larger than PM.
- If MM < PM, MM* = MM.
- If MM >= PM,
  - SB = Tukey Biweight (log(PM)-log(MM)) (log-ratio).
  - log(MM*) = log(PM)-log(max(SB, +ve)).
- Tukey Biweight: B(x) = (1 – (x/c)^2)^2 if |x|<c, 0 ow.

Can this be improved?

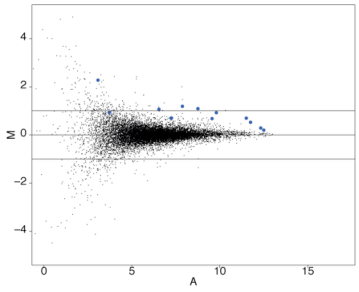


Rank of Spikeins (out of 12626)

141
250
364
368
480
586
686
838
945
1153
1567
NA
NA
NA
NA

We will discuss P/M/A calls later

Original MBEI



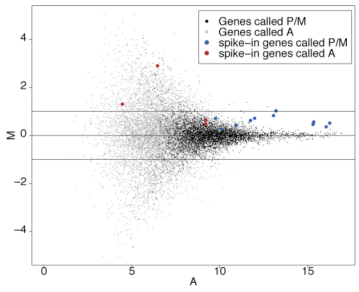
RMA

- **Robust regression method to estimate expression measure and SE from PM* (background adjusted normalized PM)**
- **Use quantile normalization**
- **Assume additive model**

$$\log_2(PM^*_{ij}) = a_i + b_j + \varepsilon_{ij}$$

- **Estimate RMA = a_i for chip *i* using robust method, such as median polish (fit iteratively, successively removing row and column medians, and accumulating the terms, until the process stabilizes).**
- **Works with *n=2* or more chips**
- **This is a robust multi-array analysis (RMA)**

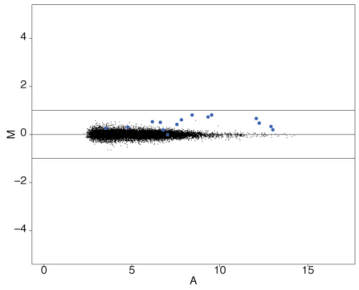
Can this be improved?



Rank of Spikeins (out of 12626)

141
250
364
368
480
586
686
838
945
1153
1567
NA
NA
NA
NA

RMA



Rank of Spikeins (out of 12626)

1
2
3
4
7
11
15
21
35
122
1182
230
450
1380
11700

Irizarry et al. (2003) *NAR* 31:e15

Detection

Detection

- The detection problem:
  "Given the probe-level data, which mRNA transcripts are present in the sample?"
- Biologists are mostly interested in expression levels, and so detection has received less attention
- To date only Affymetrix has tackled this, with
  - Rank-based tests
  - Implemented in MAS5.0

MAS Rank-based Detection

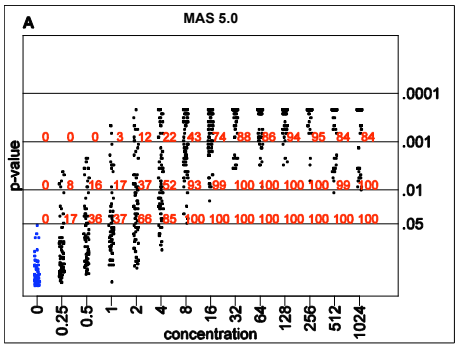
The test used in MAS 5.0 compares the following two hypotheses

H_0: median $(PM_j - MM_j)/(PM_j + MM_j) = \tau$;
H_1: median $(PM_j - MM_j)/(PM_j + MM_j) > \tau$.

Significance levels: $0 < \alpha_1 < \alpha_2 < 0.5$. If p is the p-value for the (rank) test, MAS 5.0 calls a transcript
absent: if $p > \alpha_2$,
marginal: if $\alpha_1 \leq p \leq \alpha_2$, and
present: if $p < \alpha_1$.

Typically tests are carried out with $\tau = 0.15$, $\alpha_1 = .04$ and $\alpha_2 = .06$.

Expression Detection



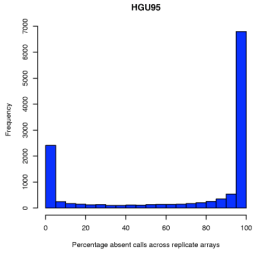
Remember uncertainty

- Some data analysts remove probesets called absent from further analysis
- This creates false negatives:

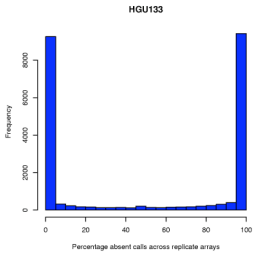
HG95	P	M	A
Present	82%	1%	17%
Absent	0%	0%	100%
HGU133	**P**	**M**	**A**
Present	77%	3%	20%
Absent	0%	0%	100%

From spike-in experiments

Consistency across reps



Consistency across reps



Current work

- **We need better estimates of means and variances of bivariate normal background noise**
- **Use observed MM intensities along with sequence information**
- **We also have a solution that does not use the MM**