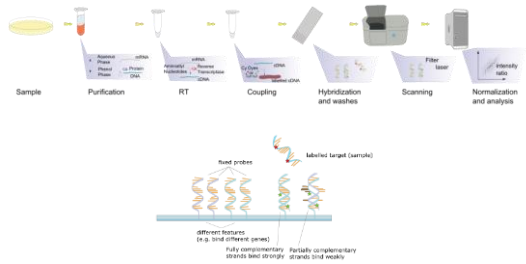
Biology 644

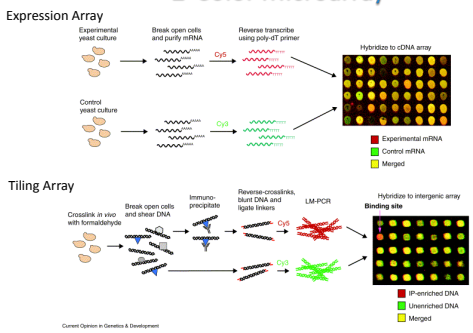
Old Title: Bioinformatics for Molecular Biologists

Potential New Title: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

1-Color Microarray

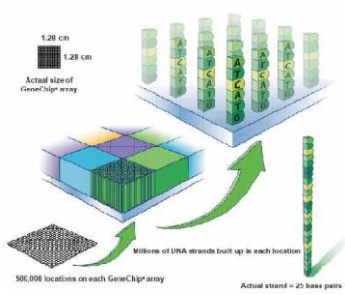


2-Color Microarray



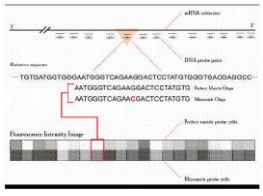
Affymetrix GeneChip

Affymetrix Microarray



Oligonucleotides (oligos), usually 25-mers (25 bases long), are directly synthesized onto a glass wafer. Each array contains up to 900,000 different oligos and each oligo is present in millions of copies.

Affymetrix Probe Set



- **Probe Set - Each gene is represented on the array by a series of different oligonucleotide probes.**
- **Probe Pair - Consists of a perfect match oligonucleotide and a mismatch oligonucleotide.**
- **The perfect match probe has a sequence exactly complimentary to the particular gene and thus measures the expression of the gene.**
- **The mismatch probe differs from the perfect match probe by a single base substitution at the center base position, disturbing the binding of the target gene transcript.**
- **This helps to determine the background and nonspecific hybridization that contributes to the signal measured for the perfect match oligo. The GeneChip Operating Software MAS algorithm subtracts the hybridization intensities of the mismatch probes from those of the perfect match probes to determine the absolute or specific intensity value for each probe set.**

Terms used in Affymetrix GeneChips

- **Target** - Fragmented, biotinylated anti-sense cRNA prepared from mRNA to be analysed. Target molecules are hybridized to the probe array and the levels of hybridization are measured with the GeneArray scanner after the array is stained with biotin - streptavidin-phycoerythrin (SAPE).
- **Probe** - Single-stranded DNA oligonucleotide synthesized directly on the surface of the GeneChip array using photolithography and combinatorial chemistry. The 25 base oligonucleotide is designed to be complementary to a specific gene transcript.
- **Probe Cell** - Single square-shaped feature on an array containing probes with a unique sequence. The size can vary depending on the array type, typically 20 µm or 18 µm. Each probe cell contains millions of probe molecules.
- **Perfect Match (PM)** - Probes that are designed to be complementary to a reference sequence.
- **Mismatch (MM)** - Probes that are designed to be complementary to a reference sequence except for a homomeric mismatch at the central position (e.g., 13th position of 25 base probe, A->T or G->C). Mismatch probes serve as a control for cross-hybridization.
- **Probe Pair** - Two probe cells, a PM and its corresponding MM. On the probe array, a probe pair is arranged with a PM cell directly above a MM cell.
- **Probe set** - A set of probes designed to detect one transcript. A probe set usually consists of 11-20 probe pairs. For example, an 11 probe pair set is made up of 11 PM probes and 11 MM probes for a total of 22 probe cells. Newer array designs from Affymetrix, e.g., HG-U133, contain probe sets with 11 probe pairs. Older designs have average probe set numbers of 16 or 20 probe pairs.
- **Target Sequence** - The portion of a transcript reference sequence that is interrogated by a probe set on the array. The target sequence extends from the first base of the most 5' probe to the last base of the most 3' probe.

The MLL.B Dataset

- "MLL.B" object is an AffyBatch R Dataset
- HGU133b Affymetrix Microarray
- MLL.B has 20 samples
- **exprs()** returns the actual expression values in a [genes, samples] matrix
- **Probenames(), pm(), and mm()** return the probenames, perfect match probes, and mismatch probes, respectively.
- **AffyRNAdeg()** models RNA degradation within each the sample
- **MAplot()** visualizes intensity-dependent ratio of raw microarray data
- **Image()** plots the actual microarray intensities on the array to look for artifacts
- **genefilter()** is a powerful function that allows you to filter (search) for gene expressions that match some criteria

Modeling RNA Degradation

- Individual PM (perfect match) probes in each probe set are ordered by location relative to the 5' end of the targeted mRNA molecule.
- RNA degradation typically starts at the 5' end, so we would expect probe intensities to be lower near the 5' end than near the 3' end.
- The affy package of BioConductor includes functions to summarize and plot the degree of RNA degradation in a series of Affymetrix experiments.
- Averages over the Nth probe in an Affymetrix probe set over all probe sets on the array.

Processing Affymetrix Data

BioConductor breaks down the low-level processing of Affymetrix data into four steps. The design is highly modular, so you can choose different algorithms at each step. It is highly likely that the results of later (high-level) analyses will change depending on your choices at these steps.

- Background correction
- Normalization
  - on the level of features = probes
- PM-correction
  - not always done
- (Probe) Summarization
  - Conversion of probe level values to probeset expression values in a robust, outlier resistant manner

Background Correction

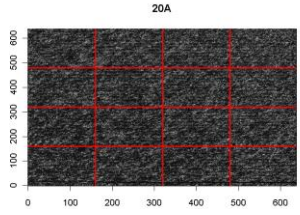
- Arguably the most crucial step for probe level processing
- The list of available background correction methods is available from the function:

```
> bgcorrect.methods()
[1] "mas" "none" "rma" "rma2 "
```

  - none - Do nothing
  - mas - Use the algorithm from MAS 5.0
  - rma - Use the Robust Multichip Analysis algorithm
  - rma2 - Use the Robust Multichip Analysis 2 algorithm

Background Correction in MAS 5.0

- MAS 5.0 divides the microarray (more precisely, the CEL file) into 16 regions.
- In each region, the intensity of the dimmest 2% of features is used to define the background level.
- Each probe is then adjusted by a weighted average of these 16 values, with the weights depending on the distance to the centroids of the 16 regions.
- MAS 5.0 background correction is a spatial detrending technique



Background Correction in RMA

- The background correction used in RMA is a non-linear correction, done on a per-chip basis.
- It is based on the distribution of PM values amongst probes on an Affymetrix array. (MM values are not considered)
- PM values are a mixture of a background signal, caused by optical noise and non-specific binding, plus a correct-hybridization signal, which is what we are trying to detect.
- The background is estimated as expectation of the signal (S) conditioned on observed PM values (O) , using a kernel density estimation.

Background Correction

JD-ALD051-v5-U133B.CEL

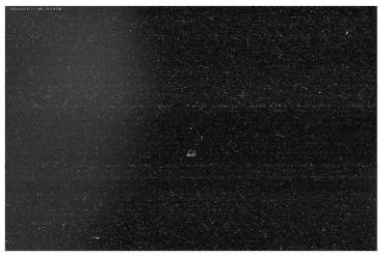


Image of the microarray reveals a bad spatial artifact, probably due to a fingerprint. Spatial detrending can at least partially remove the artifact.

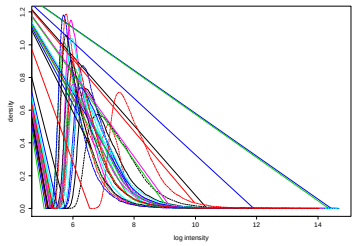
Normalization

- Normalization is necessary so that multiple chips can be compared to each other, and analyzed together.
- The normalization procedure is aimed at making the distributions identical across arrays.
- The list of available normalization methods is available from the function:

> normalize.methods(MLL.B)
- [1] "constant" "contrasts" "invariantset" "loess"
- [5] "methods" "qspline" "quantiles" "quantiles.robust"

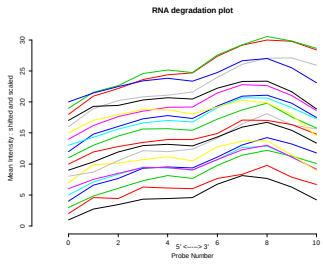
- RMA uses quantiles normalization

Normalization



The log2 intensities for the 20 samples are not on the same scale, and that is the main reason why normalization needs to be performed.

Normalization



The RNA degradation plot also shows that the 20 samples are not on the same level of expression scales.

MA Plot

- MA plots are used to visualize intensity-dependent ratio of raw microarray data
- M is the log intensity ratio (or difference between log intensities)

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

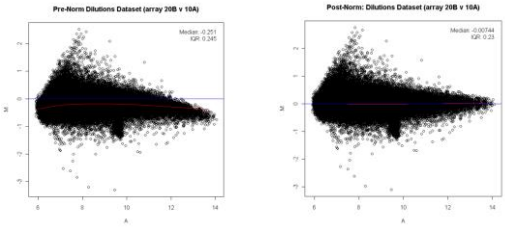
- A is the average log intensity for a dot in the plot.

$$A = \frac{1}{2}log_2(RG) = \frac{1}{2}(log_2(R) + log_2(G))$$

- The MA plot plots M on the y-axis and A on the x-axis and gives a quick overview of the distribution of the data.
- MA-plots produced by default compare the expression values on each array in the dataset with a synthetic array created using probeset-wise median expression values.
- The MA-plots have loess lines in red and the M=0 horizontal axis in blue.
- clearly aberrant loess line on these MA-plots often are indicative of potential quality problems.
- microarrays typically show a bias whereby higher A results in higher |M|, i.e. the brighter the spot the more likely an observed difference between sample and control
- The median and IQR values appearing on each plot quantify the center and vertical spread of the M values. These statistics can be turned off by supplying the show.statistics=FALSE argument to MAplot.

MA Plot

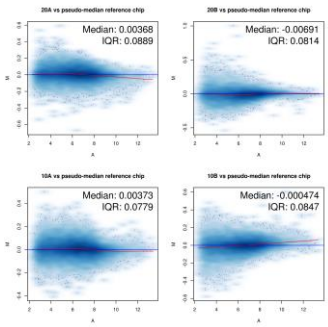
MA plot is used to visualize intensity-dependent ratio of raw microarray data



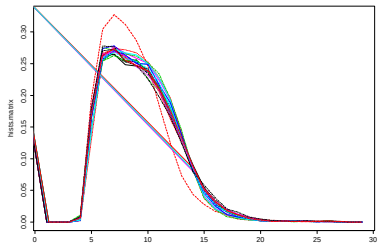
M is the log intensity ratio: $M = log_2(R/G) = log_2(R) - log_2(G)$

A is the average log intensity: $A = \frac{1}{2}log_2(RG) = \frac{1}{2}(log_2(R) + log_2(G))$

MA Plot



After RMA Normalization



After RMA normalization the overlapping histograms of the intensities for the 20 samples are now on the same scale.

(Probe) Summarization

- Once the probe-level PM (and possibly MM) values have been background-corrected and normalized, they need to be summarized into expression measures, so that the result is a single expression measure per probeset.
- Generates a single number that represents our best guess at the expression level of the targeted gene.
- The available summarization methods can be obtained using the function call:

```
> express.summary.stat.methods()

[1] "avgdiff" "liwong" "mas"
[4] "medianpolish" "playerout" "pdnn"
```


(Probe) Summarization in RMA

- RMA summarization used is motivated by the assumption that observed log-transformed PM values follow a linear additive model containing:
  - a probe affinity effect
  - a gene specific effect (the expression level)
  - and an error term.
- For RMA, the probe affinity effects are assumed to sum to zero, and the gene effect (expression level) is estimated using median polishing.
- Median polishing is a robust model fitting technique, that protects against outlier probes.
- The MM values are thrown away. RMA contends that there are too many cases where MM > PM, and hence including the MMs introduces more variability than the correction is worth.