Statistics for Human Genetics and Molecular Biology
Lecture 24: Processing Microarray Data

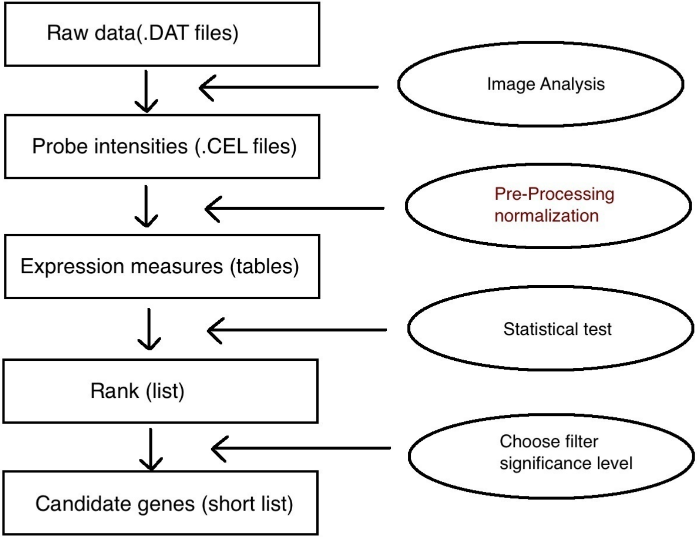
Dr. Yen-Yi Ho (yho@umn.edu)

Nov 02, 2015

Objectives of Lecture 24

- Preprocessing
    - Background Correction
    - Normalization
    - Probe Level Data Summarization

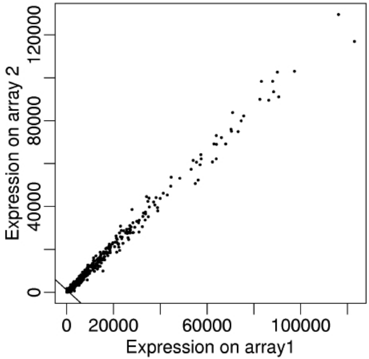
Analysis Flow Chart: Preprocessing



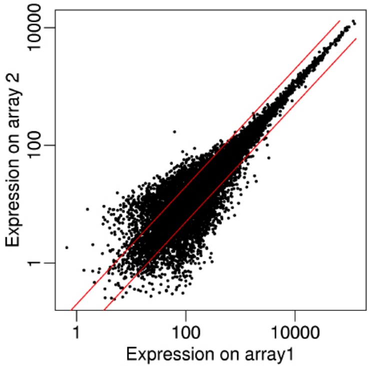
Why log

- Fold changes are the preferred quantification for differential gene expression. Fold changes are basically ratios.
- Ratios are not symmetric around 1. This makes it problematic to perform statistical operations with ratios. Hence we prefer logs.

Raw data from two arrays

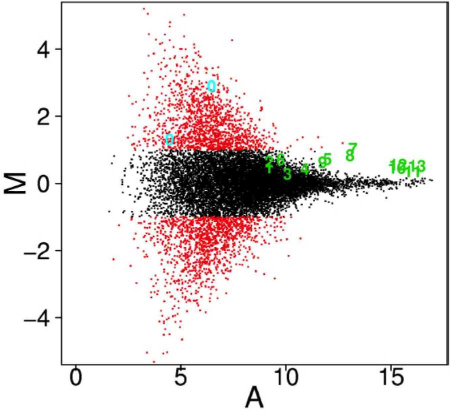


Same data in log scale



Background Noise

$$M = \log_2 I_1 - \log_2 I_2, \qquad A = (\log_2 I_1 + \log_2 I_2)/2$$



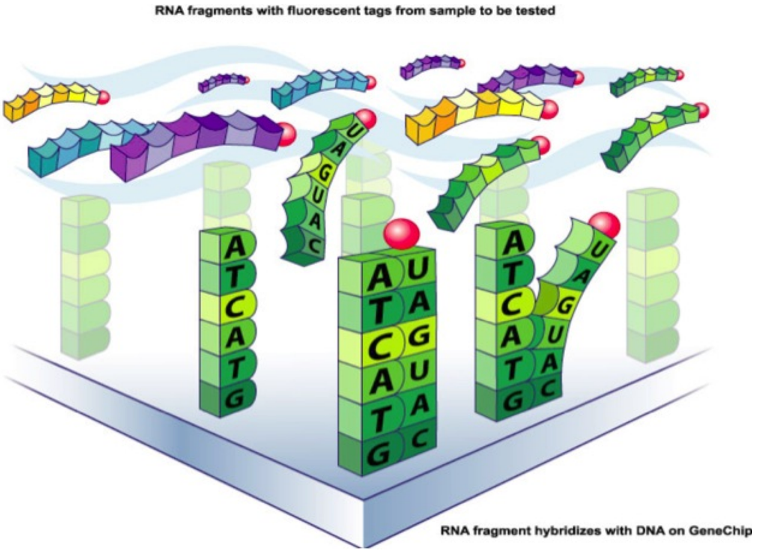
colored numbers are probes from spike-in experiment

Preprocessing: Three Steps Procedure

BioConductor breaks down the low-level processing of Affymetrix data into three 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: Adjust for Non-Specific Binding
- Normalization
- Probe Level Data Summarization

Affymetrix GeneChip®


RNA fragments with fluorescent tags from sample to be tested
UAGUAC
ATCATG
ATCATG
UAGUAC
ATCATG
UAGUAC
RNA fragment hybridizes with DNA on GeneChip

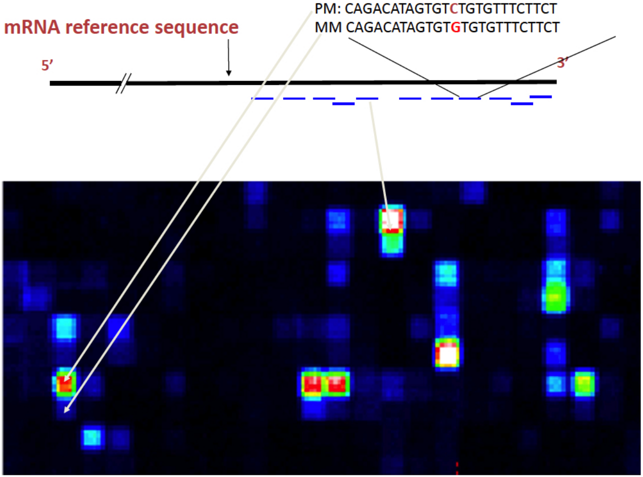
source: Affymatrix

Background Adjustment

Purpose

- Correct for background noise and processing effects
- Adjust for cross-hybridization, i.e. binding of non-specific DNA.
- Adjust expression measures so that they are linearly related to concentration

Affymetrix: PM versus MM



PM: Perfect Match
MM: Mis-Match

PM - MM problems

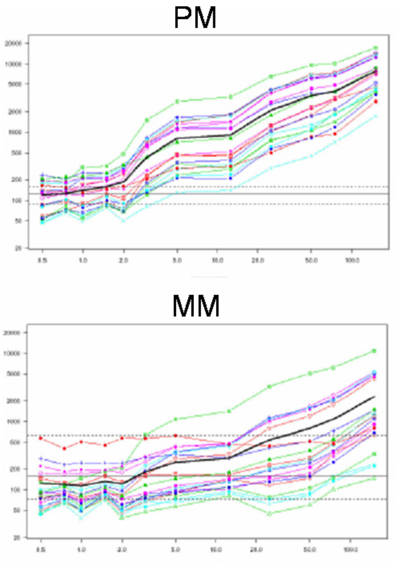
Assuming

$$\begin{aligned} PM &= \alpha + \beta \\ MM &= \alpha \end{aligned}$$

- 20% $\sim$ 30% probe-pairs have MM $>$ PM!
- MMs are PMs for some genes.
- MM may be detecting signal as well as non-specific binding
- E=PM -MM increases the variance of E

Note: Everything is on log-scale from now

Spike-In Data



$$
\begin{aligned}
PM &= \alpha_1 + \beta \\
MM &= \alpha_2
\end{aligned}
$$

MM measures signal too!

Robust Multi-Array Average (RMA)

Assume

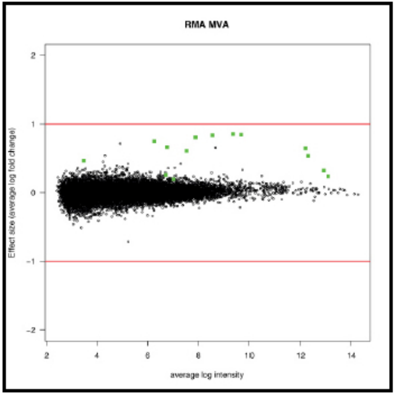
$$\begin{aligned} PM &= \alpha + \beta \\ \alpha &\sim \mathit{Normal}(\mu, \sigma^2) \\ \beta &\sim \mathit{Exponential}(\lambda) \end{aligned}$$

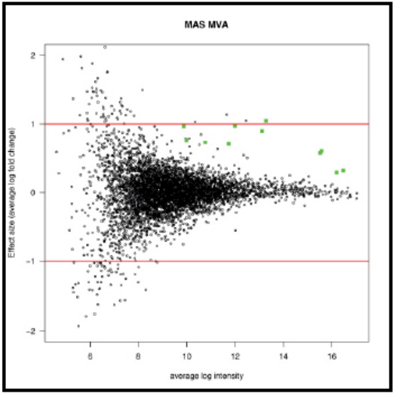
Use all the probe intensities on the array to estimate (μ, σ^2, λ)

$$E[\beta|PM] = PM - \mu - \lambda\sigma^2 + \sigma\{1/\sqrt{2\pi}e^{-1/2(PM/\sigma)^2/\Phi(PM/\sigma)}\}$$

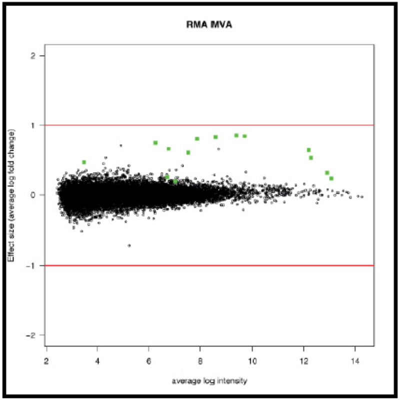
No need for MM!

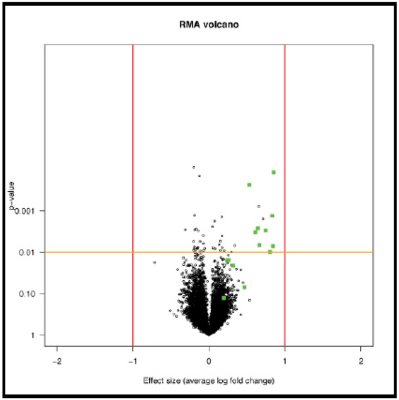
RMA versus MAS 5.0





Volcano plot





Summary

- Take logs: probe effect is additive on log scale
- Background correction reduces noise from non-specific binding
- RMA improves precision and power to detect differentially expressed genes

Exercise: Homework 7 (1) (2) (3)

1. Download CEL files from GSE18088 at gene expression omnibus (http://www.ncbi.nlm.nih.gov/geo/)
2. Pre-process the data in the study.
Hint:
library(oligo)
library(siggenes)
library(limma)
library(pd.hg.u133.plus.2)
library(hgu133plus2.db)
library(hgu133a.db)
exprs