Introduction to Bioinformatics

Giri Narasimhan

ECS 254; Phone: x3748

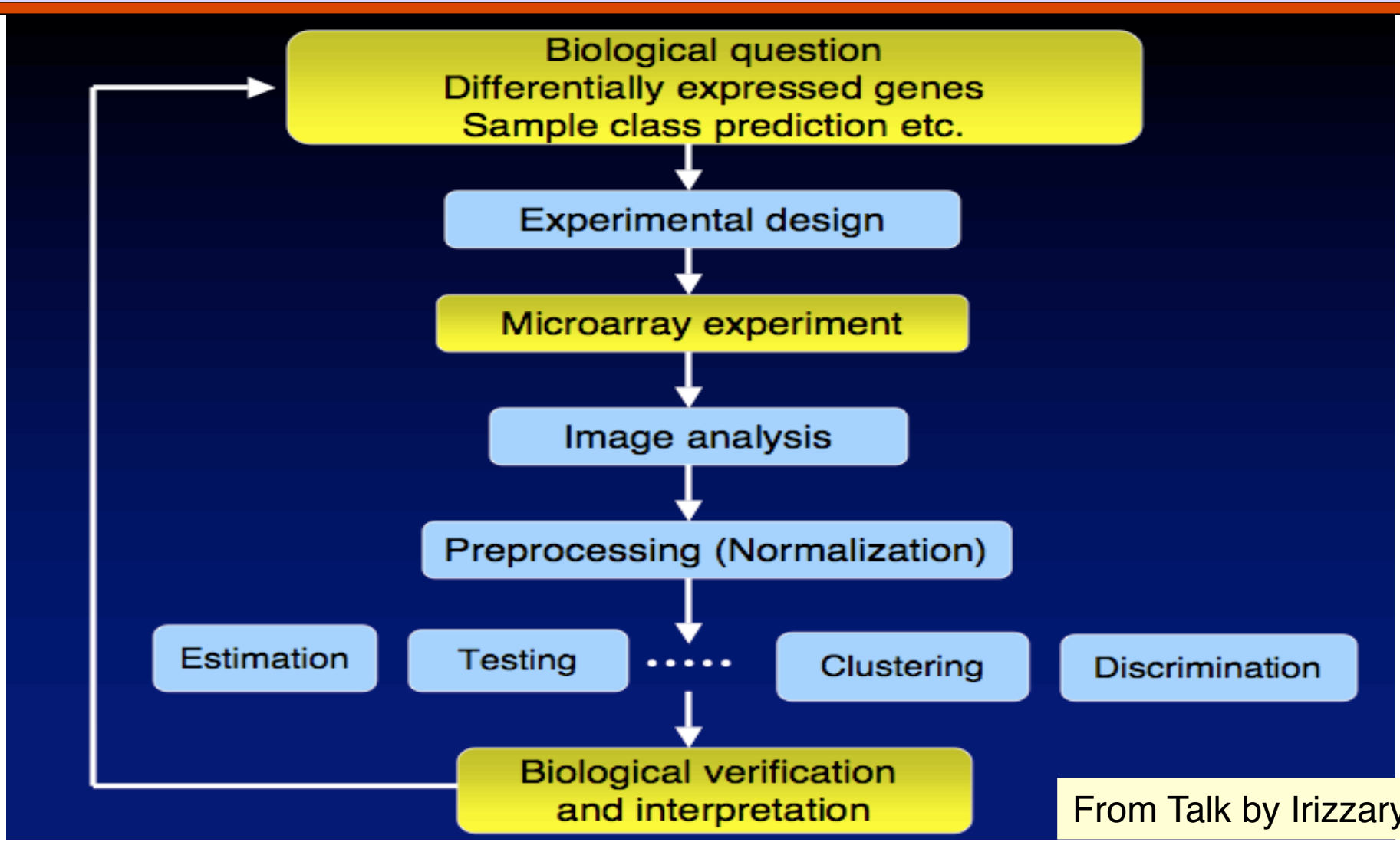
giri@cs.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS11.html

Reading

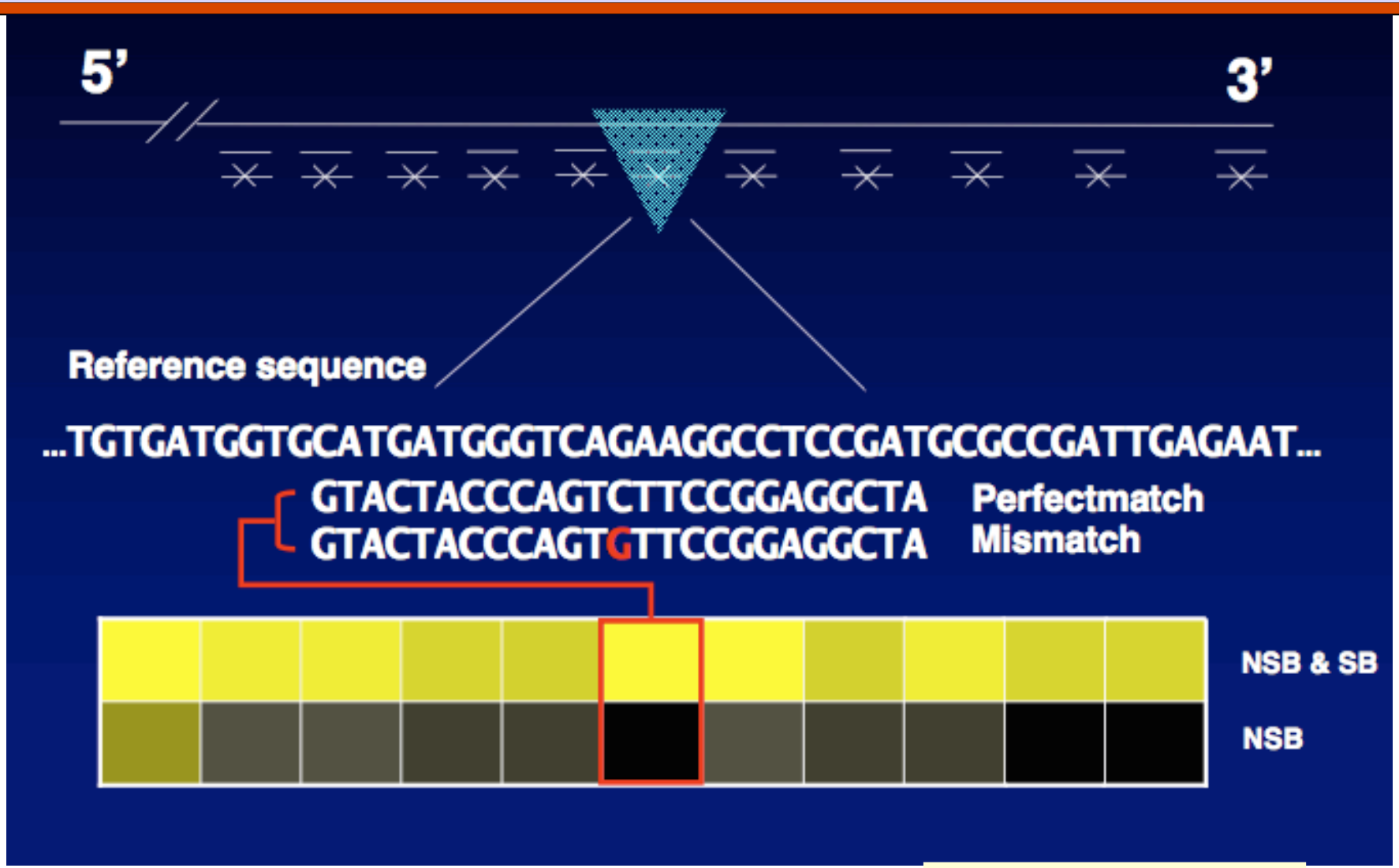
- The following slides come from a series of talks by Rafael Irizzary from Johns Hopkins
- Much of the material can be found in detail in the following papers from [http://www.biostat.jhsph.edu/~ririzarr/papers/]
  - Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP (2003) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics. Vol. 4, Number 2: 249-264.
  - Bolstad, B.M., Irizarry RA, Astrand, M, and Speed, TP (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics. 19(2):185-193.

Inference Process



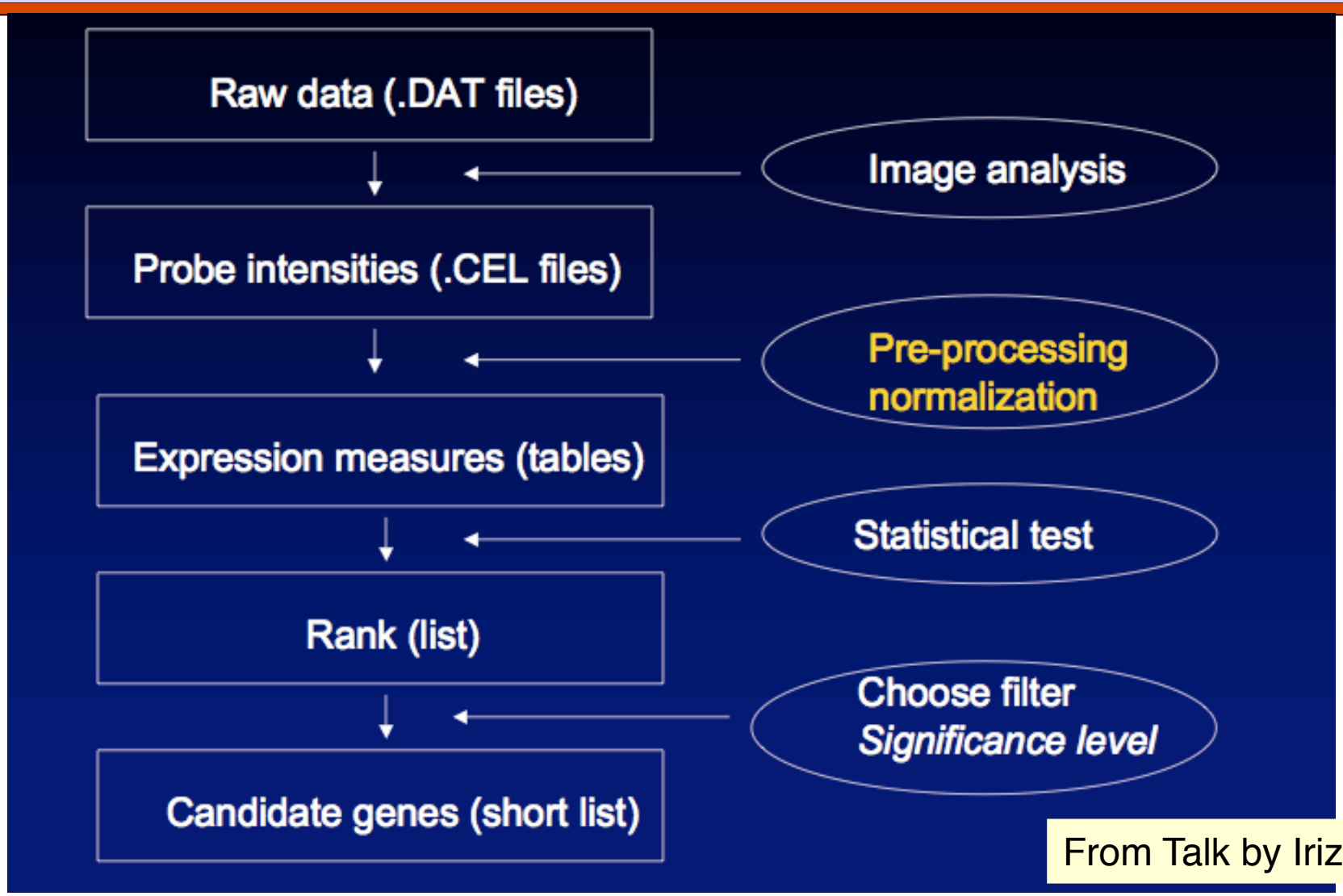
From Talk by Irizzary

Affymetrix Genechip Design



From Talk by Irizzary

Workflow: Analyzing Affy data



From Talk by Irizzary

Affy Files

- **DAT** file: image file, about 10 million pixels, 30-50 MB
- **CEL** file: cell intensity file with probe level PM and MM values
- **CDF** file: chip description file describing which probes go in which probe sets and the location of probe-pair sets (genes, gene fragments, ESTs)

From Talk by Irizzary

Image analysis & Background Correction

- Each probe cell: 10 X 10 pixels
- Gridding estimates location of probe cell centers
- Signal is computed by
  - Ignoring outer 36 pixels leaving a 8 X 8 pixel area
  - Taking the 75 percentile of the signal from the 8 X 8 pixel area
- Background signal is computed as the average of the lowest 2% probe cell values, which is then subtracted from the individual signals

From Talk by Irizzary

Standard Normalization Procedure

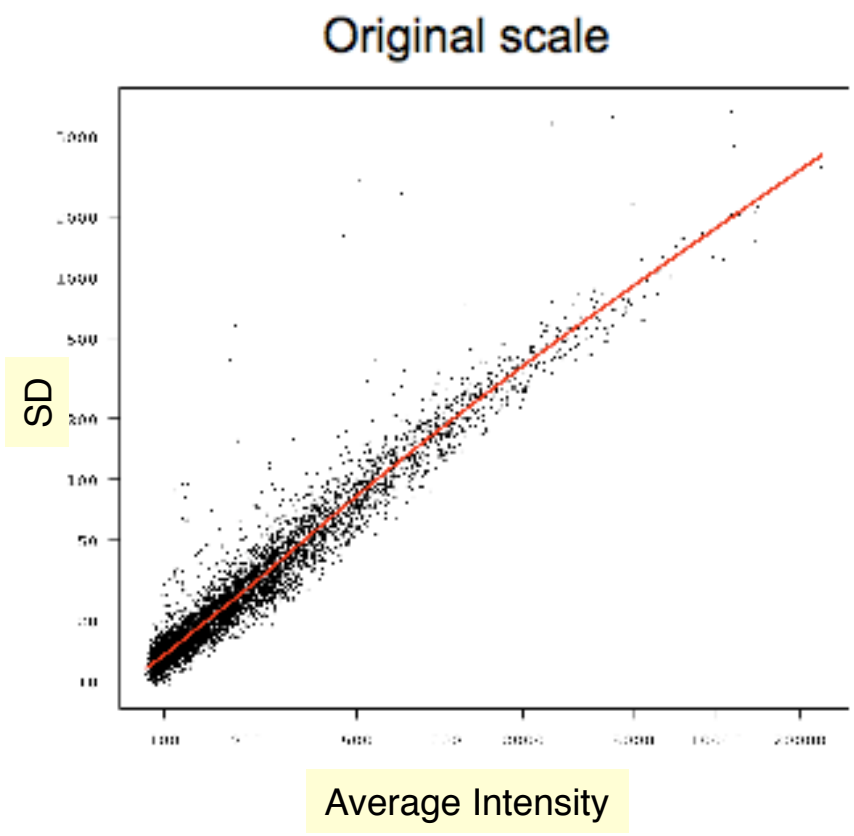
- Log-transform the data
- Ensure that the average intensity and the standard deviation are the same across all arrays.
- This requires the choice of a baseline array, which may or may not be obvious.

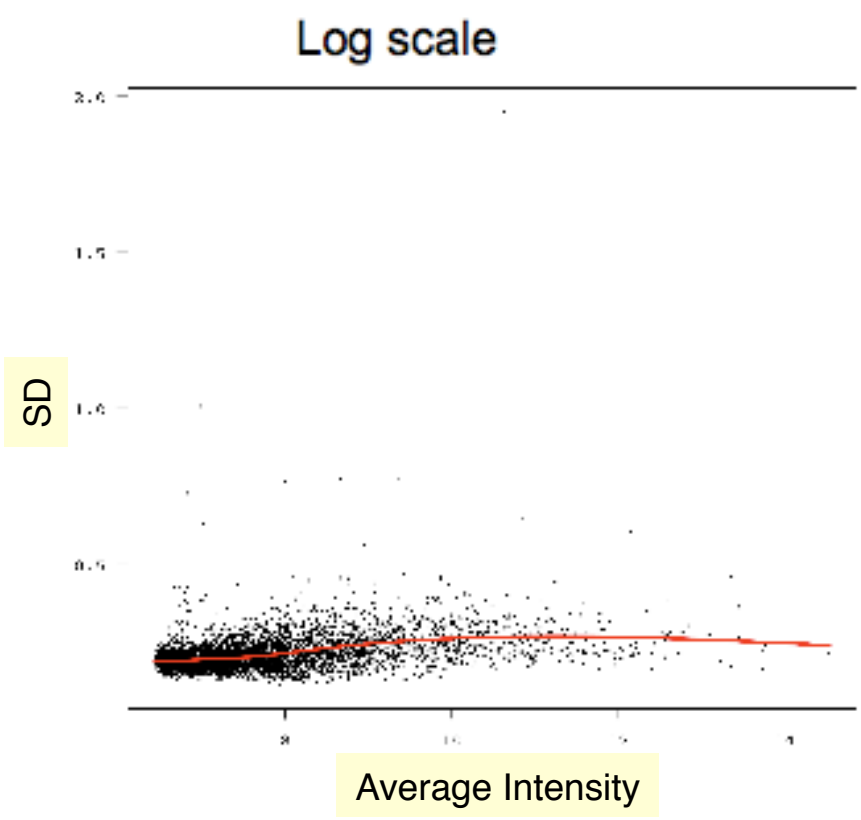
Analyzing Affy data

- MAS 4.0
  - Works with PM-MM
  - Negative values result very often
  - Very noisy for low expressed genes
  - Averages without log-transformation
- dChip [Li & Wong, PNAS **98**(1):31-36]
  - Accounts for probe effect
  - Uses non-linear normalization
  - Multi-chip analysis reveals outliers
- MAS 5.0
  - Improves on problems with MAS 4.0

From Talk by Irizzary

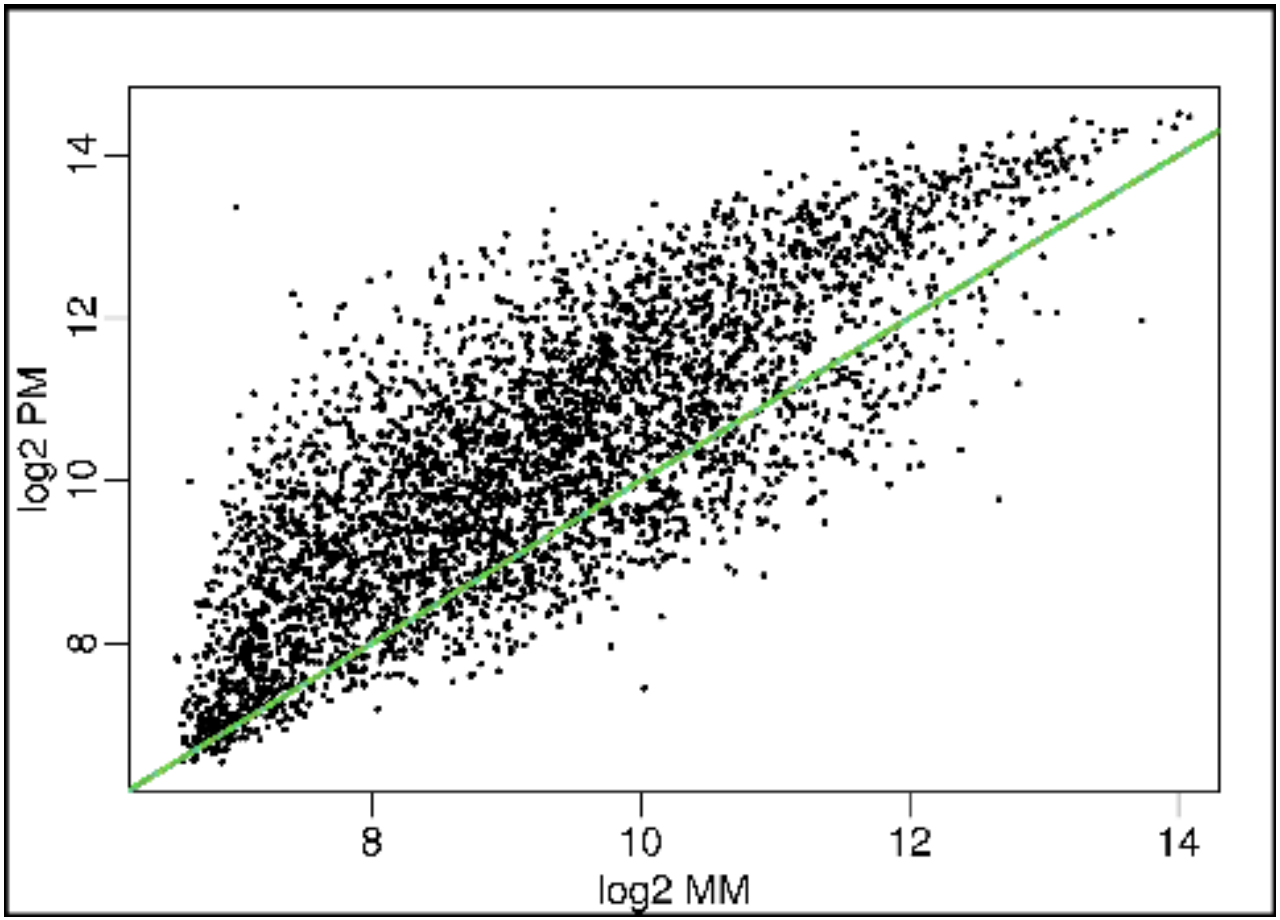
Why you use log-transforms?





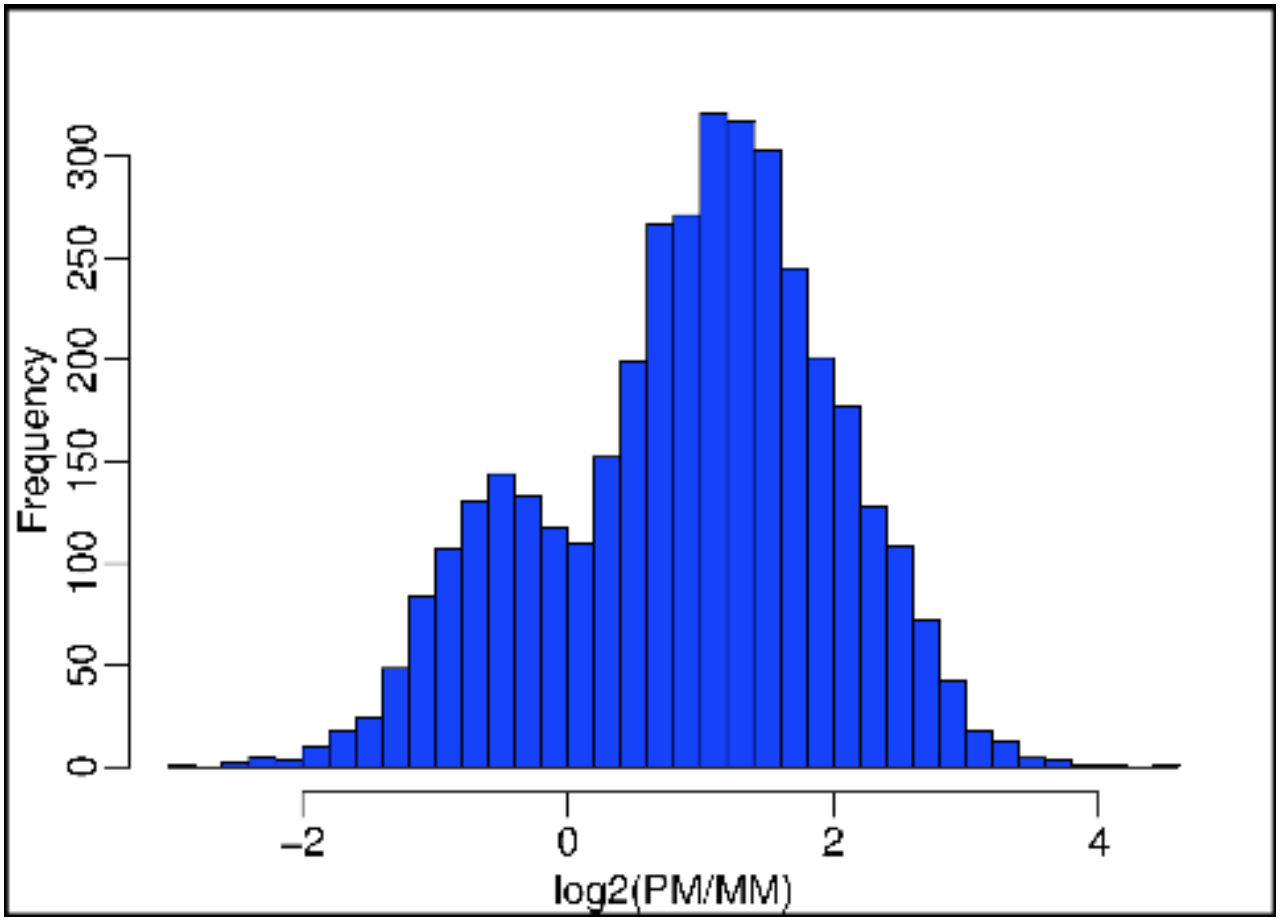
From Talk by Irizzary

Problem with using (transformed) PM-MM



From Talk by Irizzary

Bimodality for large expression values



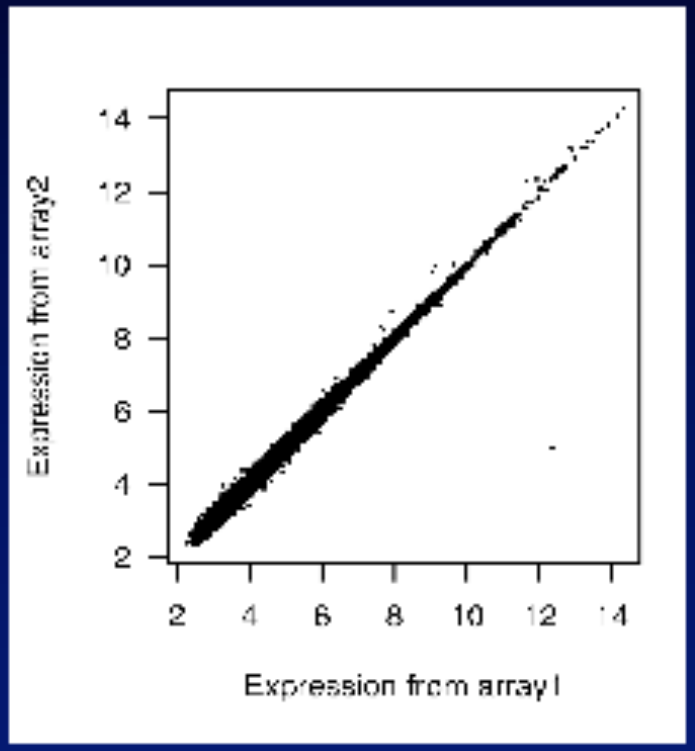
From Talk by Irizzary

MAS 5.0

- **MAS 5.0** is Affymetrix software for microarray data analysis.
- Ad hoc background procedure used
- For summarization, they use:
  - **Signal = TukeyBiweight{log($PM_j - MM_j^*$)}**
  - Tukey Biweight: $B(x) = (1 - (x/c)^2)^2$, if $x<c$
    $= 0$ otherwise
- Ad hoc scale normalization used

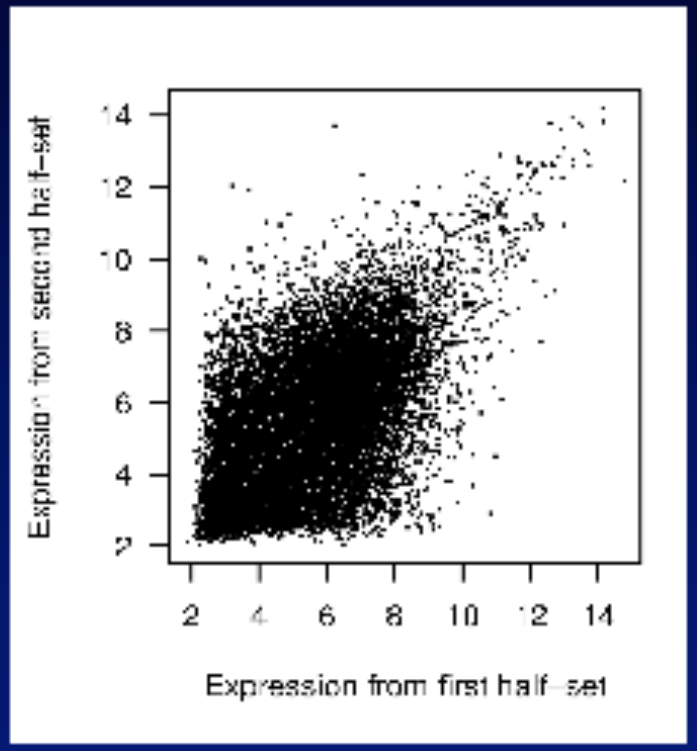
From Talk by Irizzary & PhD thesis by Astrand

2 replicate arrays



Expression from corresponding probes are highly correlated

Correlation is higher than 0.99

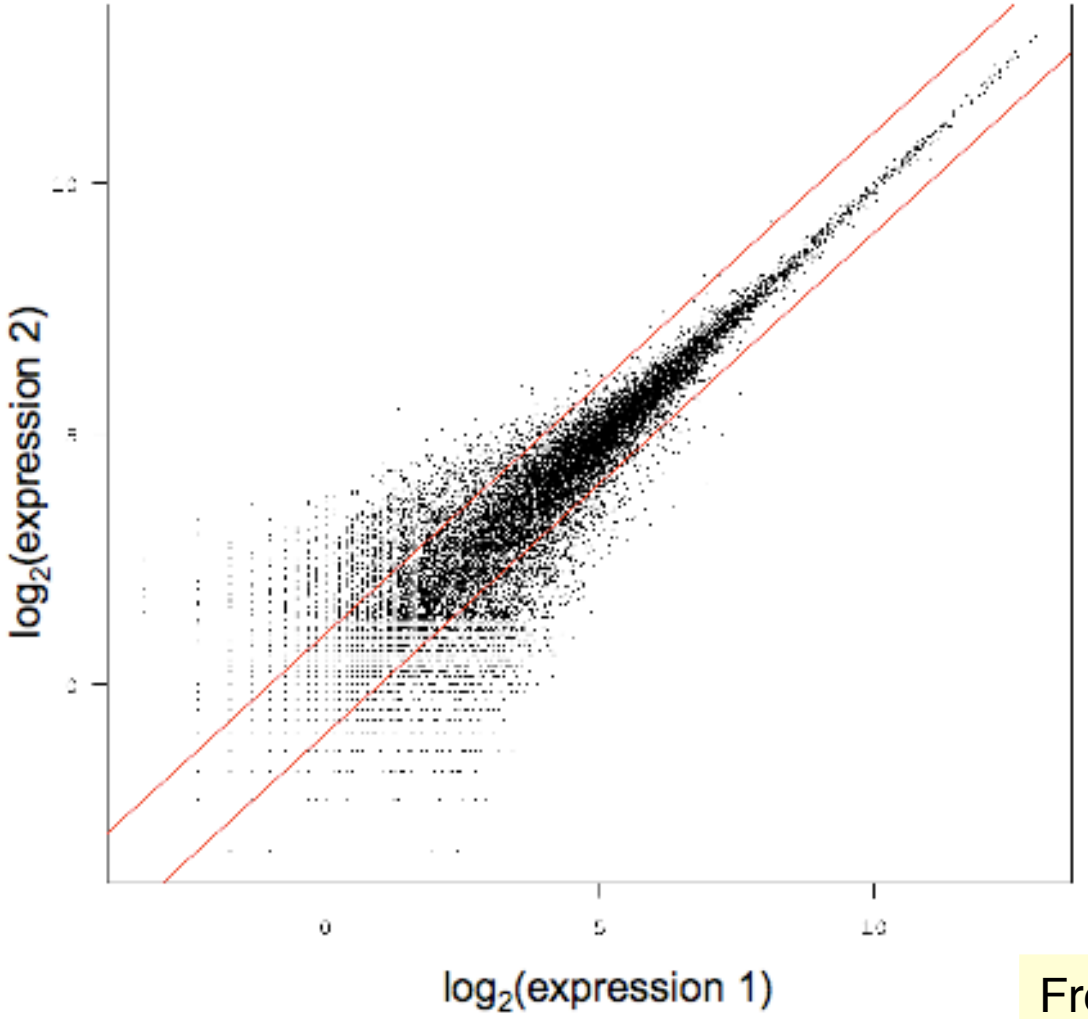


Expression not correlated when probes randomly partitioned

Correlation drops to 0.55

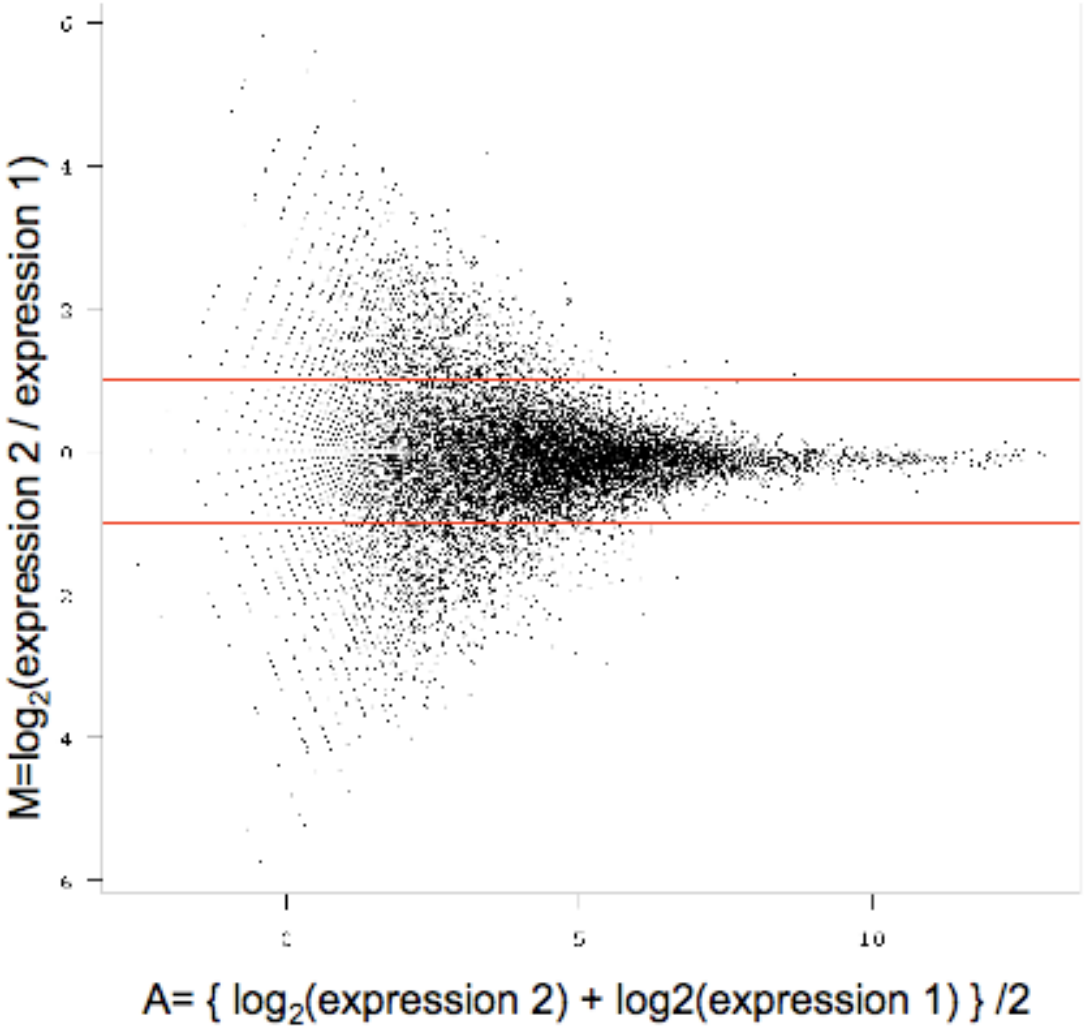
From Talk by Irizzary

We have to deal with **variations**!



From Talk by Irizzary

MvA Plots



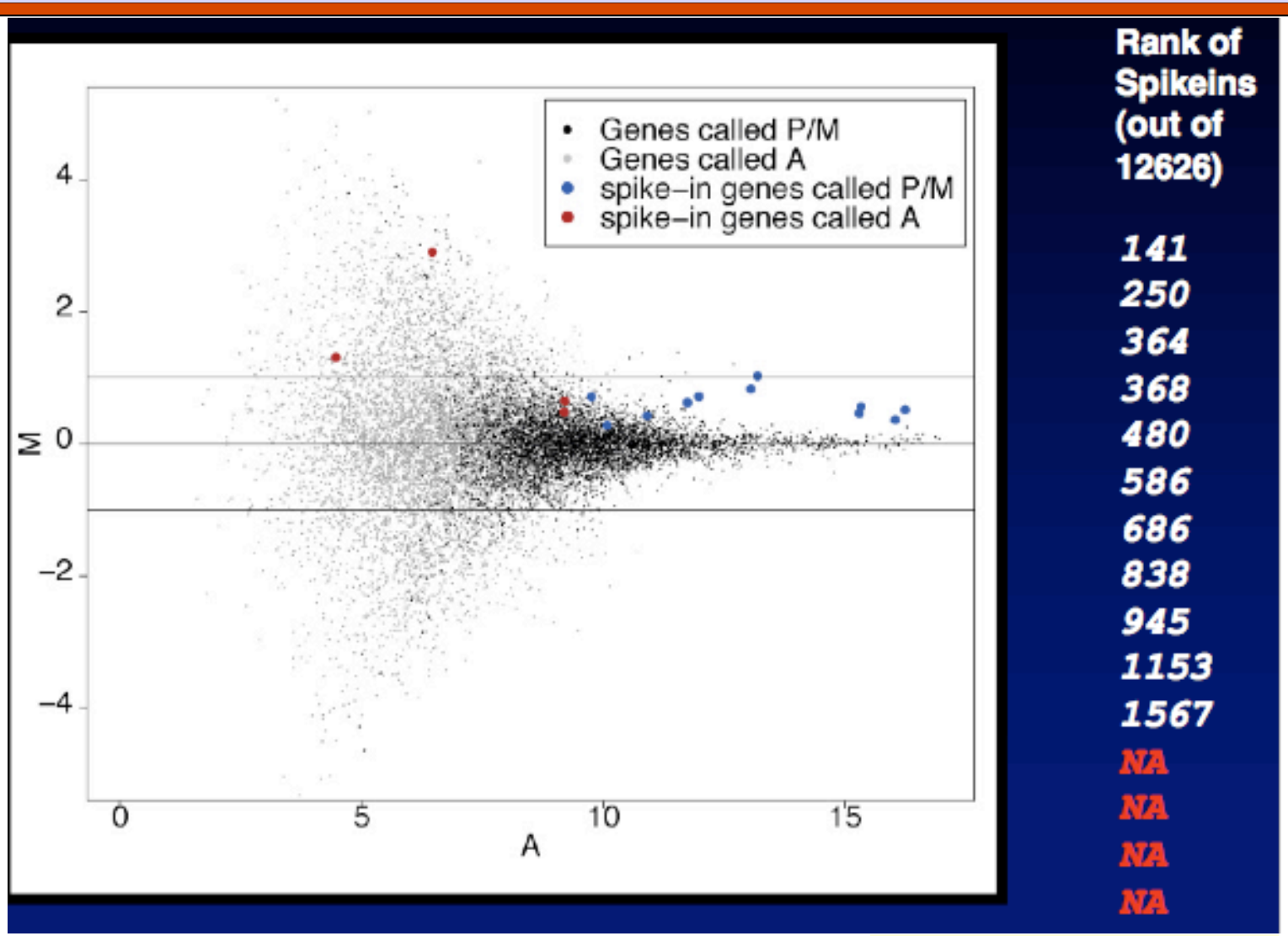
From Talk by Irizzary

Spike-in Experiment

- Replicate RNA samples were hybridized to various arrays
- Some probe sets were spiked in at different concentrations across the different arrays
- Goal was to see if these spiked probe sets "stood out" as differentially expressed

From Talk by Irizzary

Analyzing Spike-in data with MAS 5.0



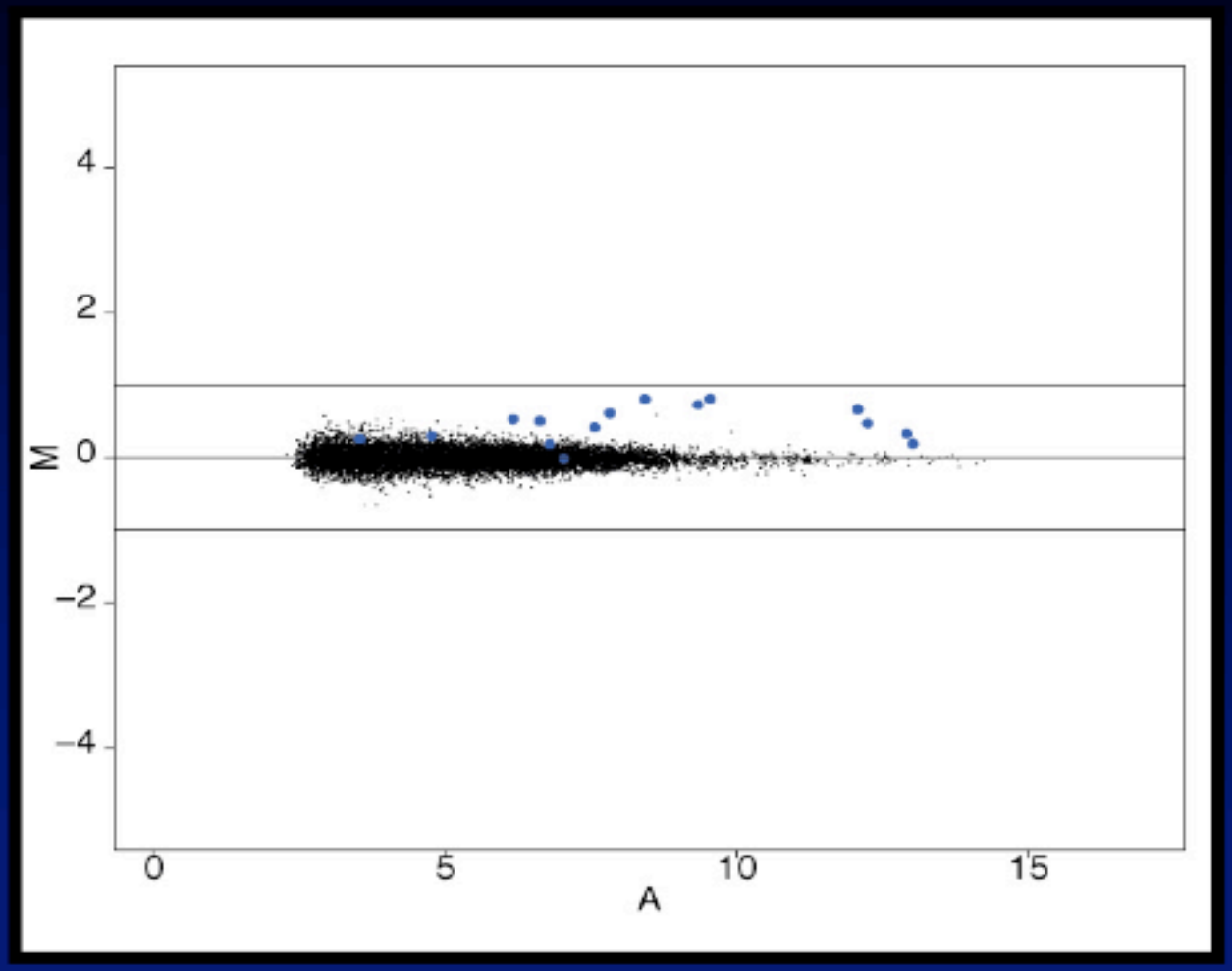
From Talk by Irizzary

Robust Multiarray normalization (RMA)

- Background correction separately for each array
  - Find E{Sig | Sig+Bgd = PM}
  - Bgd is normal and Sig is exponential
- Uses quantile normalization to achieve "identical empirical distributions of intensities" on all arrays
- Summarization: Performed separately for each probe set by fitting probe level additive model
- Uses median polish algorithm to robustly estimate expression on a specific chip
- Also see GCRMA [Wu, Irizzary et al., 2004]

From Talk by Irizzary & PhD thesis by Astrand

Analyzing Spike-in data with RMA



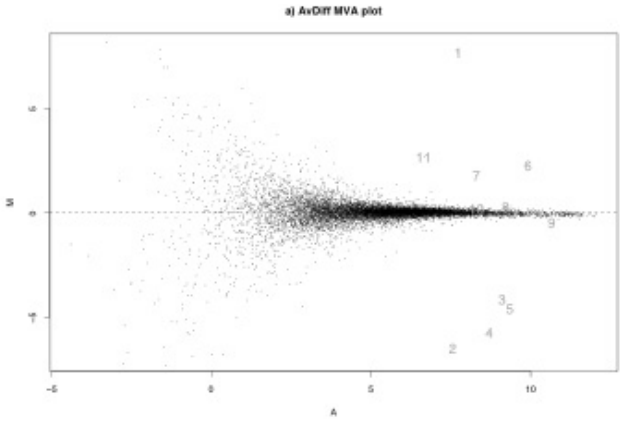
Rank of Spikeins (out of 12626)

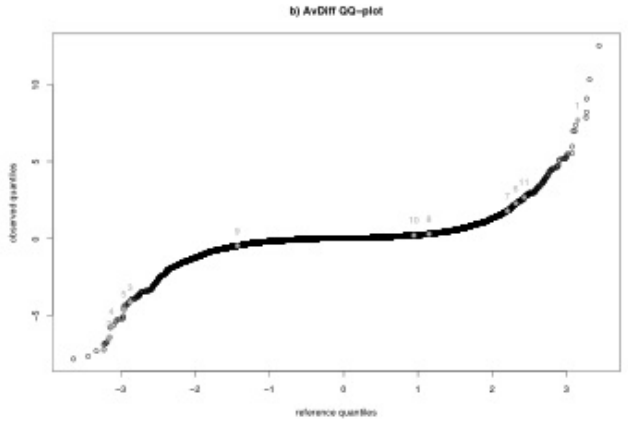
1
2
3
4
7
11
15
21
35
122
1182
230
450
1380
11700

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

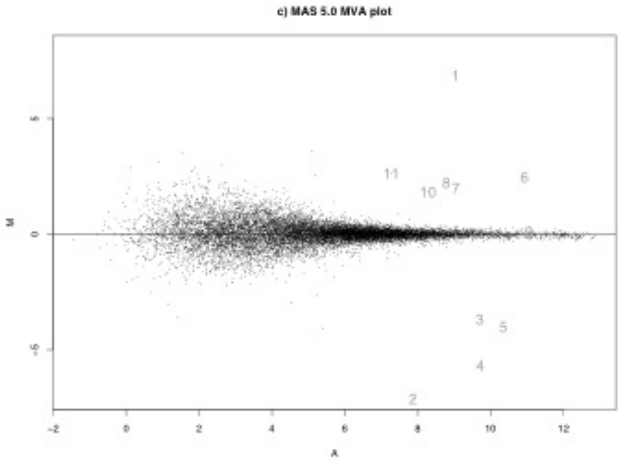
From Talk by Irizzary

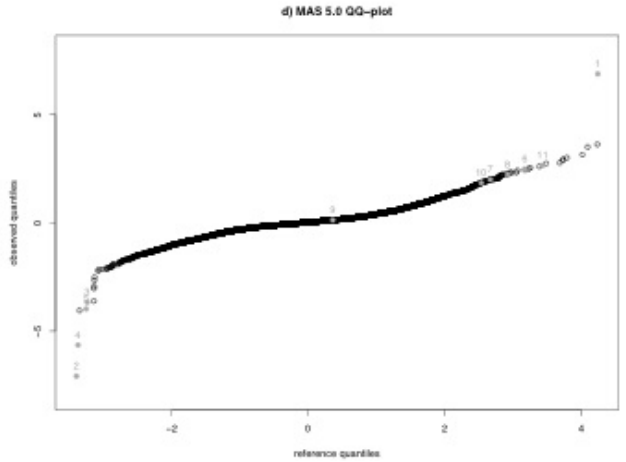
MvA and q-q plots





MAS 4.0

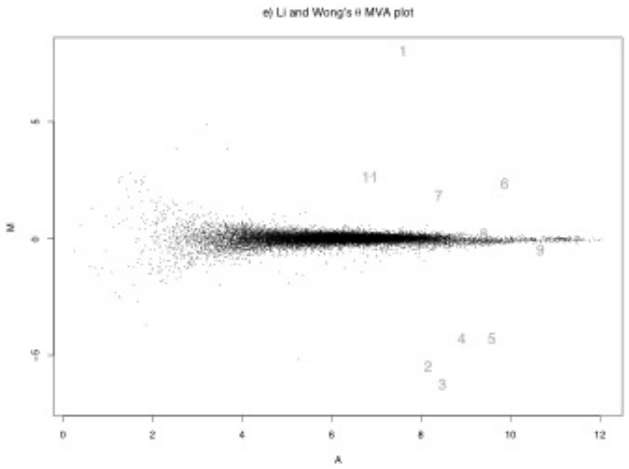


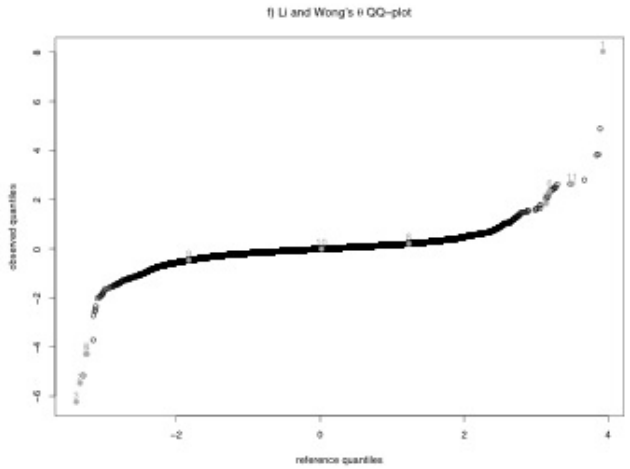


MAS 5.0

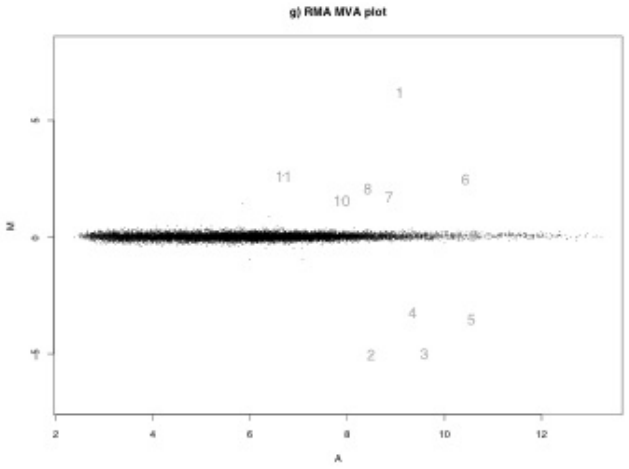
From Talk by Irizzary

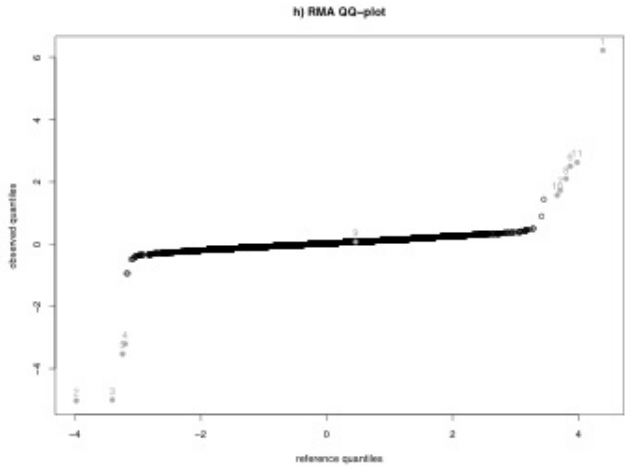
MvA and q-q Plots





MBEI





RMA

From Talk by Irizzary

Before and after quantile normalization

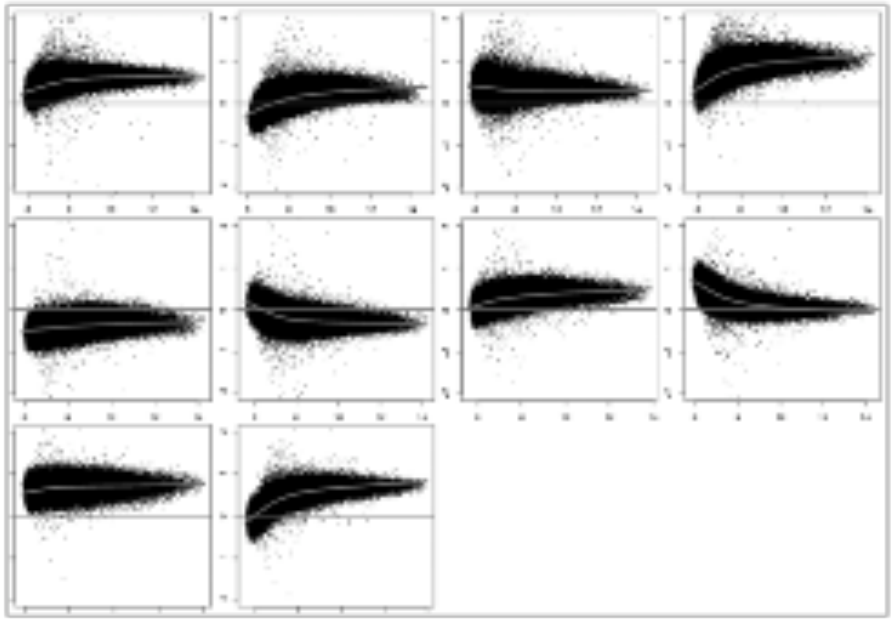


Fig. 2. 10 pairwise M versus A plots using liver (at concentration 10) dilution series data for unadjusted data.

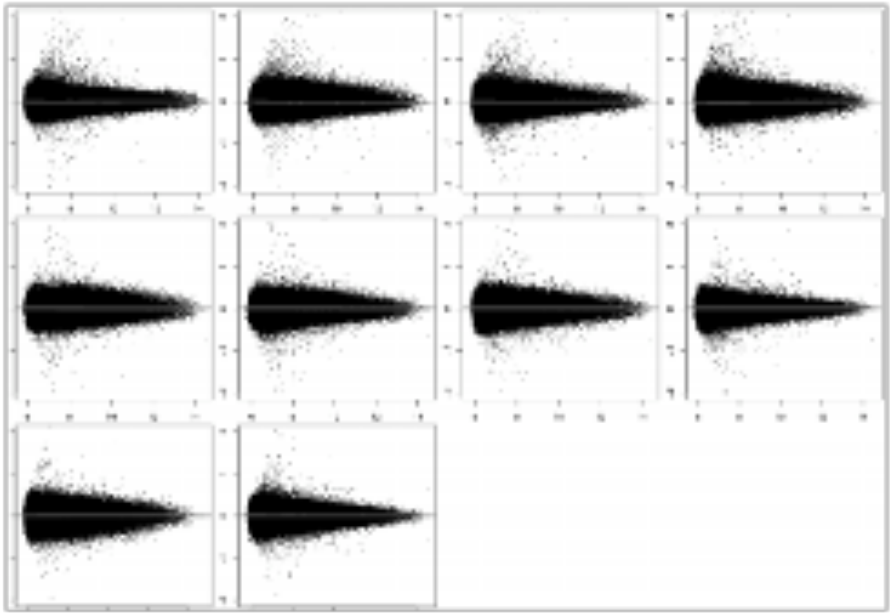


Fig. 3. 10 pairwise M versus A plots using liver (at concentration 10) dilution series data after quantile normalization.

From Talk by Irizzary

Bioconductor

- **Bioconductor** is an **open source** and open development software project for the analysis of biomedical and genomic data.
- World-wide project started in 2001
- **R** and the **R package system** are used to design and distribute software
- Commercial version of Bioconductor software called **ArrayAnalyzer**

From Talk by Irizzary

R: A Statistical Programming Language

- Try the tutorial at: [http://www.cyclismo.org/tutorial/R/]
- Also at: [http://www.math.ilstu.edu/dhkim/Rstuff/Rtutor.html]

Installing a package from Bioconductor

- Let's consider **LIMMA**: Linear Models for Microarray Data. It is a software package for the analysis of gene expression microarray data, especially the use of linear models for analyzing designed experiments and the assessment of differential expression. The package includes pre-processing capabilities for two-color spotted arrays. The differential expression methods apply to all array platforms and treat Affymetrix, single channel and two channel experiments in a unified way.
- Here's how you install and load it:
  - Here is an installation script
    ```
    > source("http://www.bioconductor.org/biocLite.R")
    > biocLite("limma")
    > biocLite("statmod")
    ```
  - If you want to install some other package (say "affy"), then you type:
    ```
    > biocLite("affy")
    ```

Analyzing Swirl Data (Agilent)

- Section 8.1 of LIMMA User's Guide on Swirl data set (**http://pbil.univ-lyon1.fr/library/limma/doc/usersguide.html**)
- Follow Sec 8.3 as homework. Note that the data for the experiment in Sec 8.3 is not from the address given there, but from: **http://cybert.microarray.ics.uci.edu/tutorial/Affy%20Data/**
- More comments on microarray analysis (http://discover.nci.nih.gov/microarrayAnalysis/Microarray.Home.jsp)