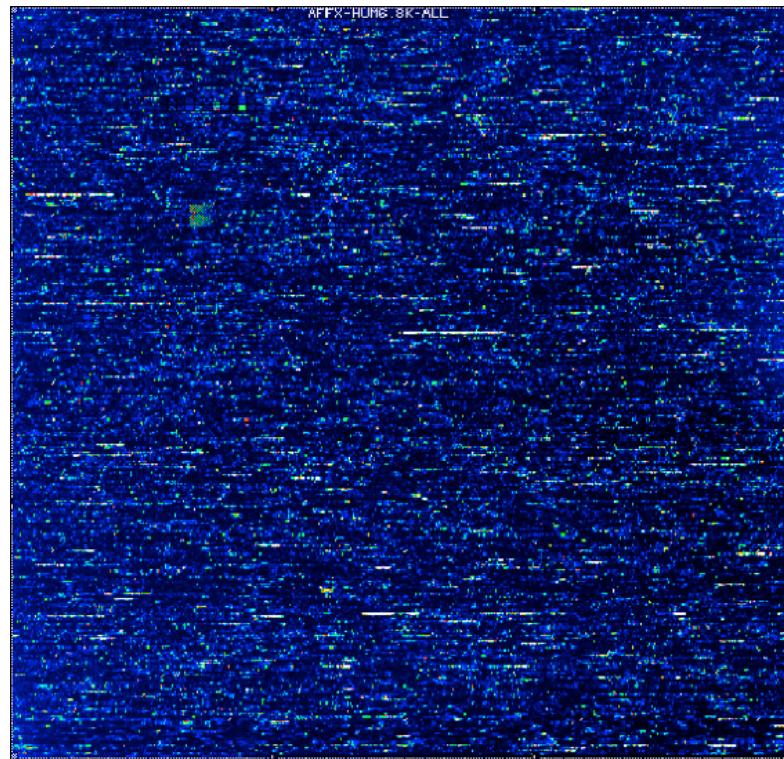


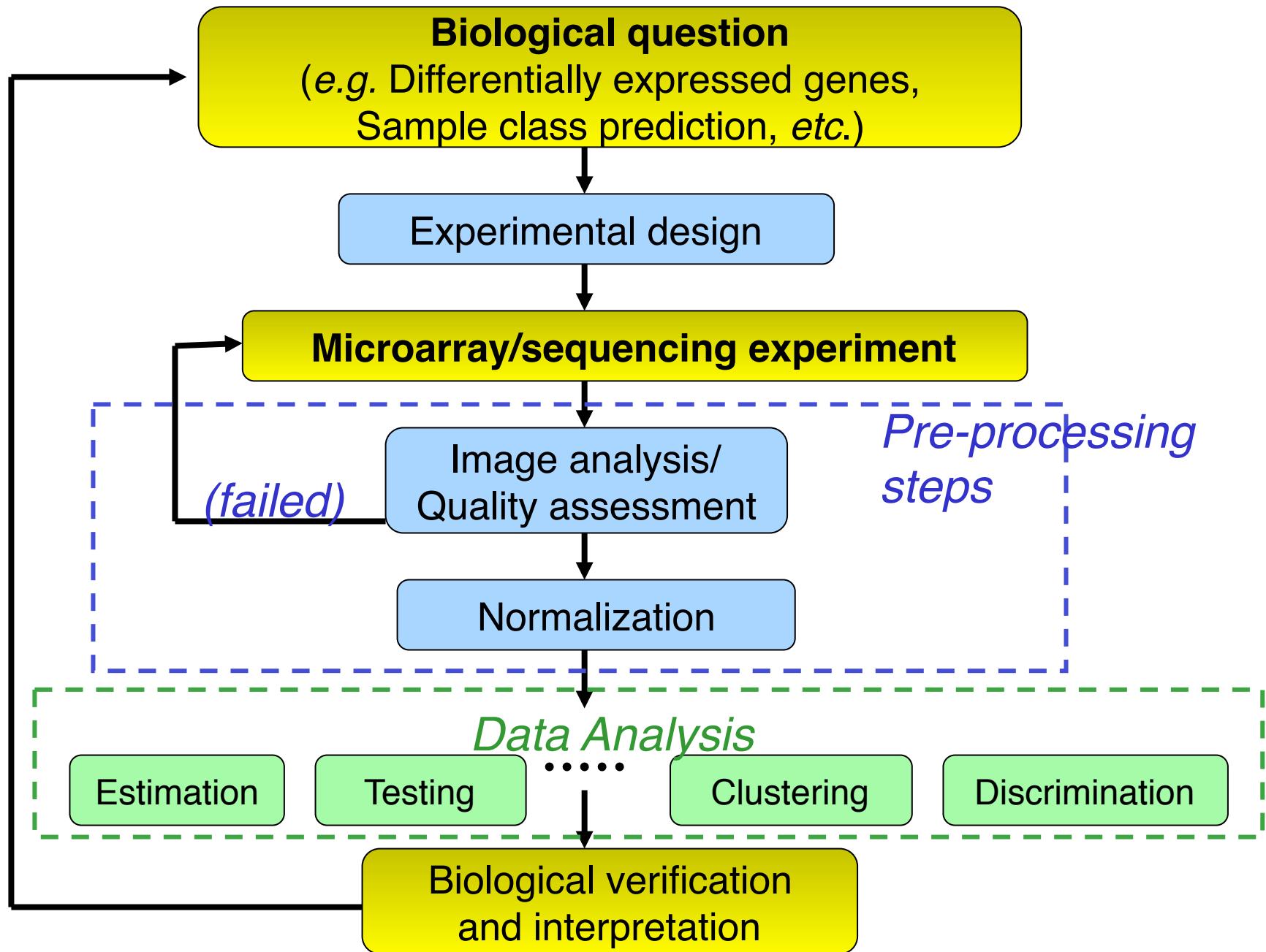
Statistics for Genomic Data Analysis

*Affymetrix signal quantification;
Bayesian estimation and IDE*



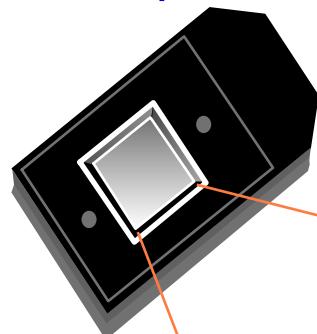
<http://moodle.epfl.ch/course/view.php?id=15271>





Affymetrix GeneChip Probe Arrays

GeneChip Probe Array



1.28cm

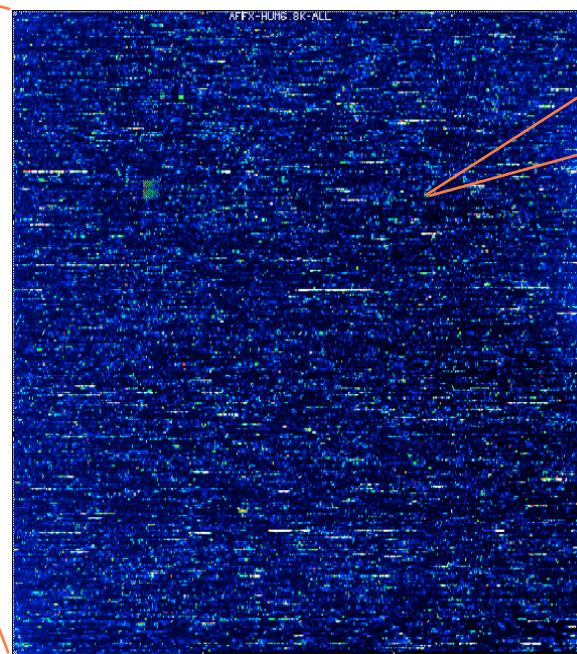
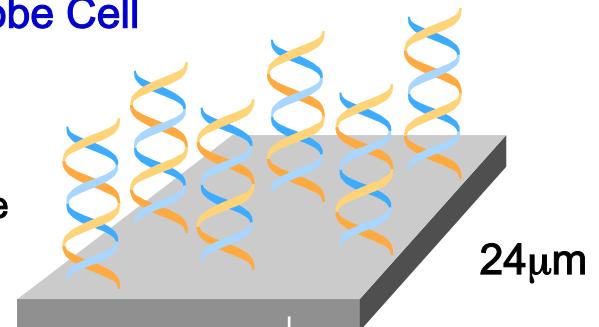


Image of Hybridized Probe Array

Hybridized Probe Cell

Single stranded,
labeled RNA target
Oligonucleotide probe



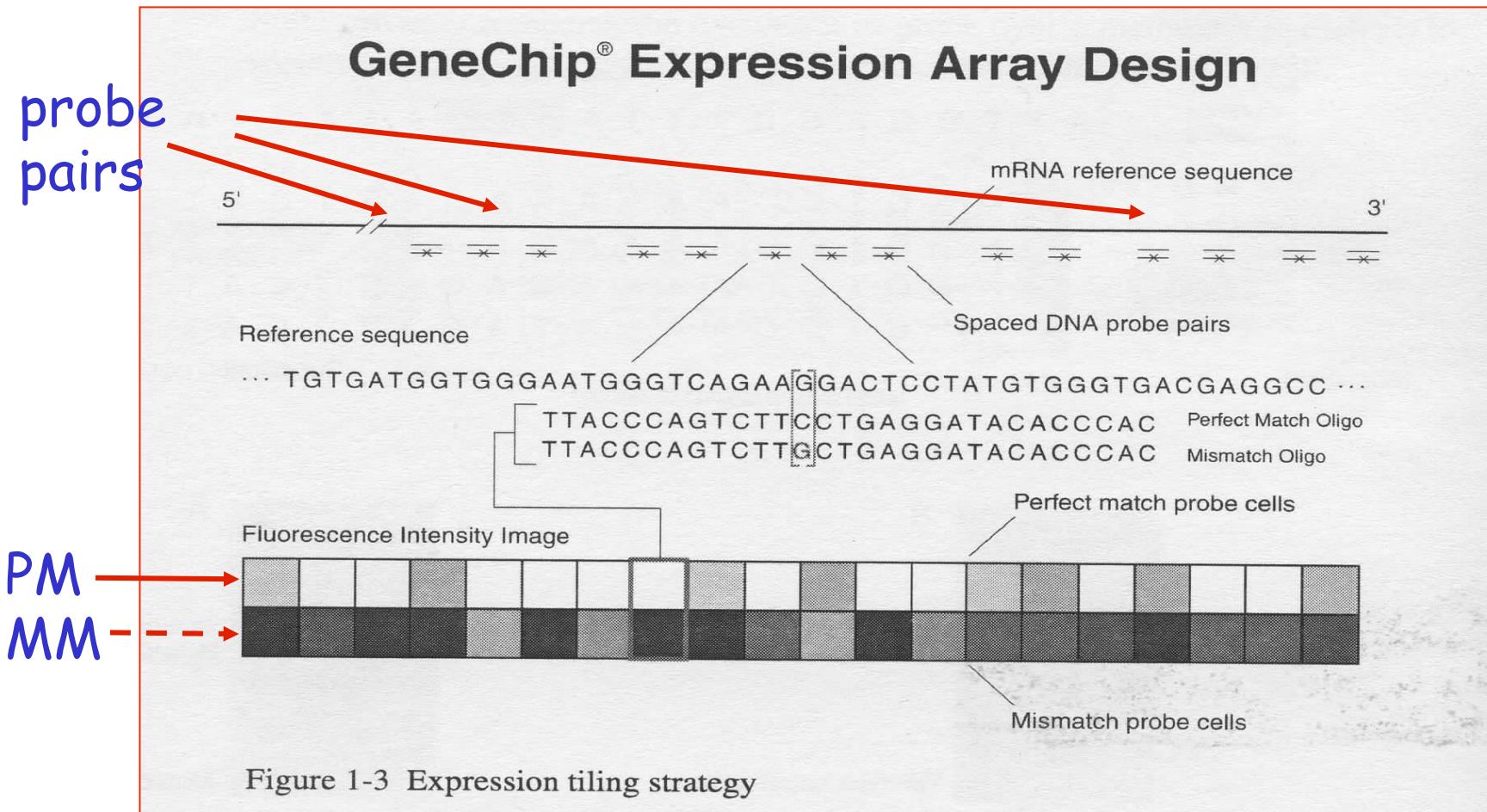
Millions of copies of a specific
oligonucleotide probe

>200,000 different
complementary probes

Compliments of D. Gerhold



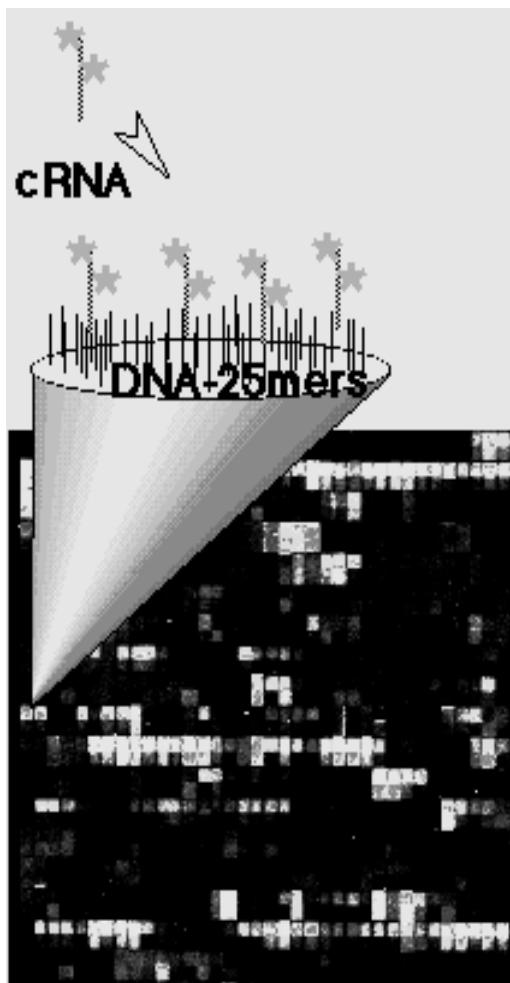
Array design



probe set = collection of probe pairs;
There are tens of thousands of probe sets per chip



Image analysis



- About 100 pixels per probe cell
- These intensities are combined to form one number representing expression for the probe cell oligo



Artifacts in microarrays

- We are interested in finding true *biologically meaningful differences* between sample types
- Due to other sources of systematic variation, there are also usually *artifactual differences*
- Sources of artifacts include:
 - batch effects
 - hybridization artifacts



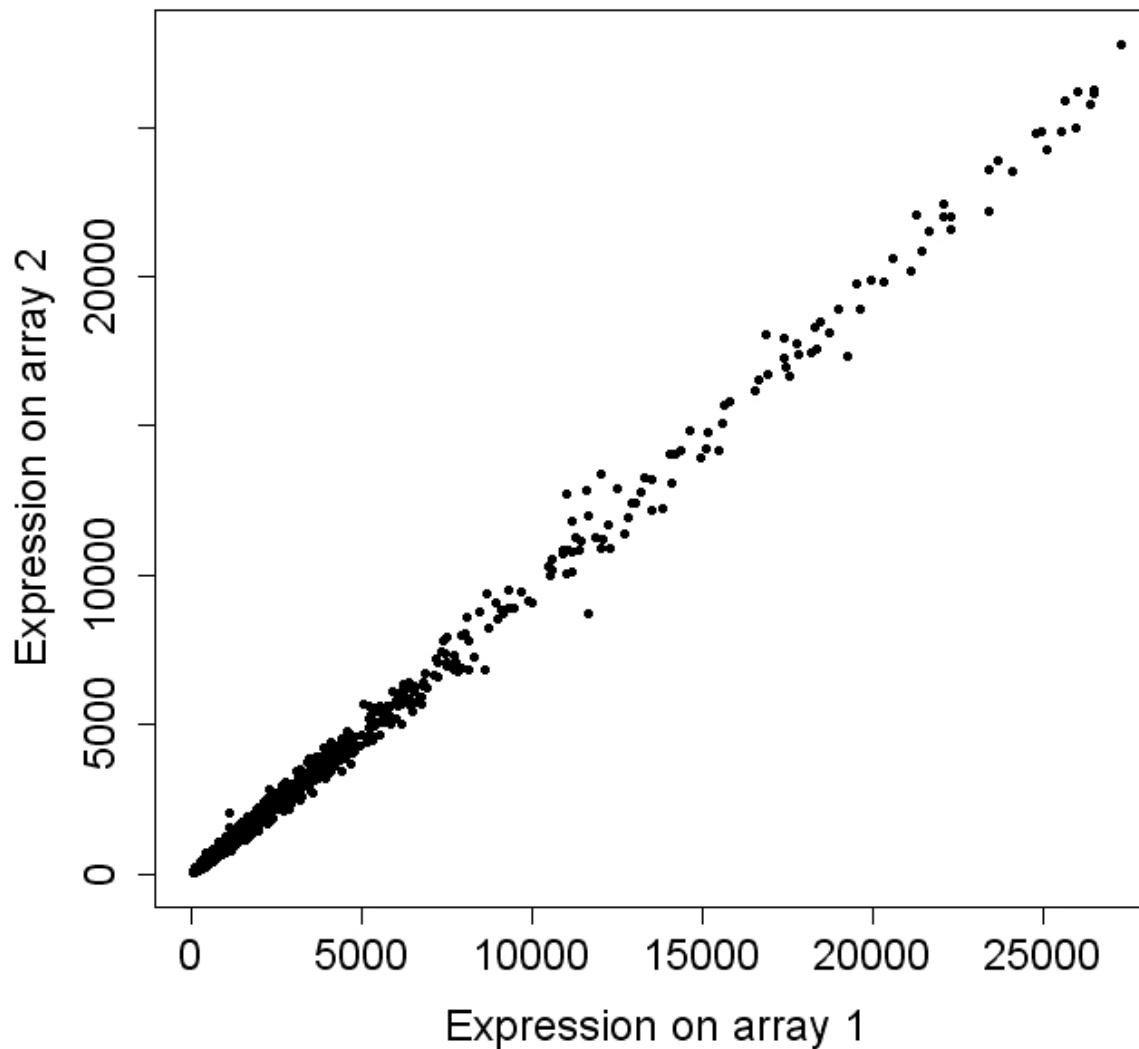
Looking for artifacts

- Exploratory data analysis (EDA) is an important component of microarray data preprocessing
- EDA involves identifying data artifacts
- We will use several types of plots for data visualization, primarily
 - *scatterplots*
 - *boxplots*
 - *spatial plots/pseudo-images*



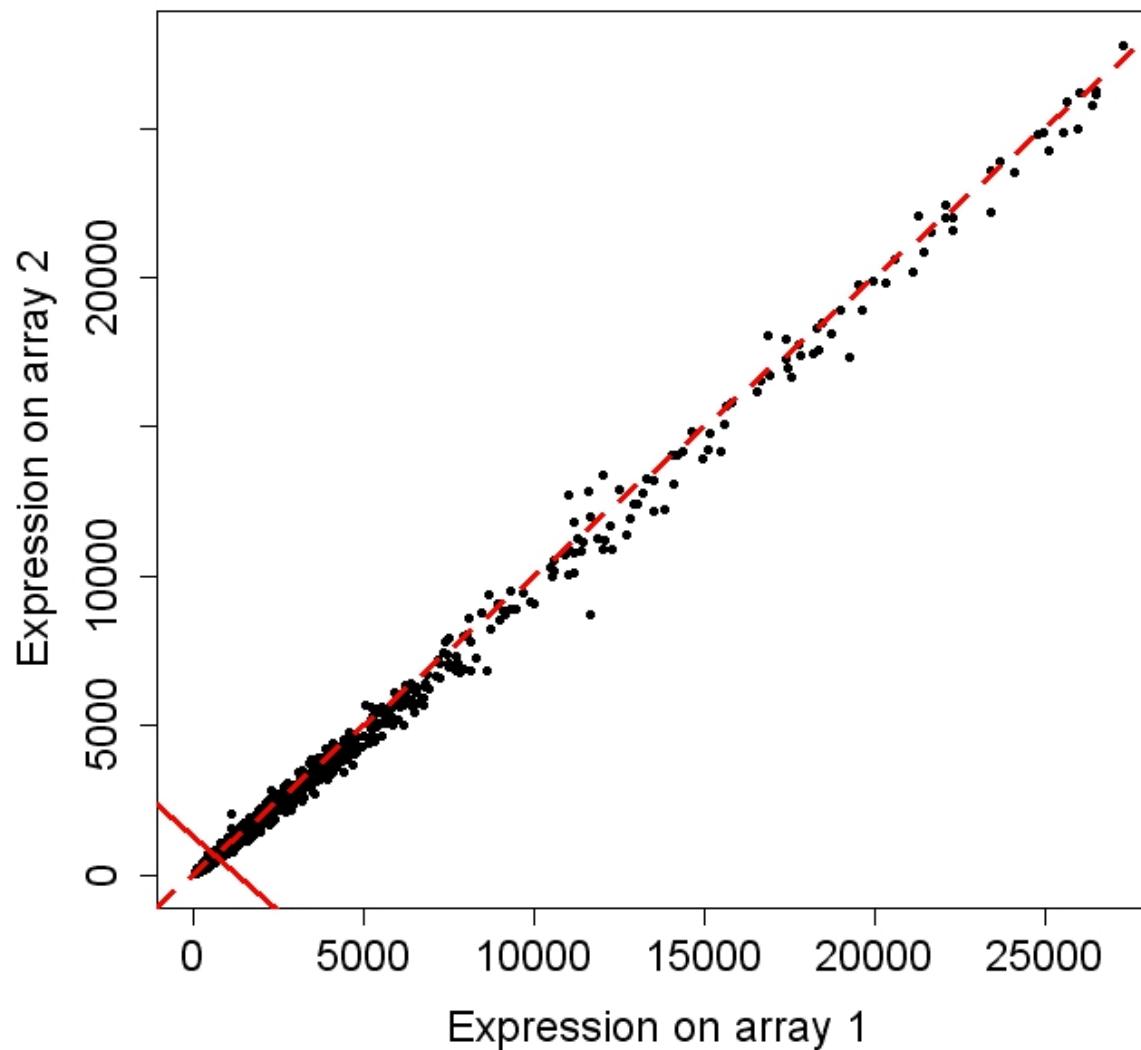
Scatterplots

Expression data from 2 arrays



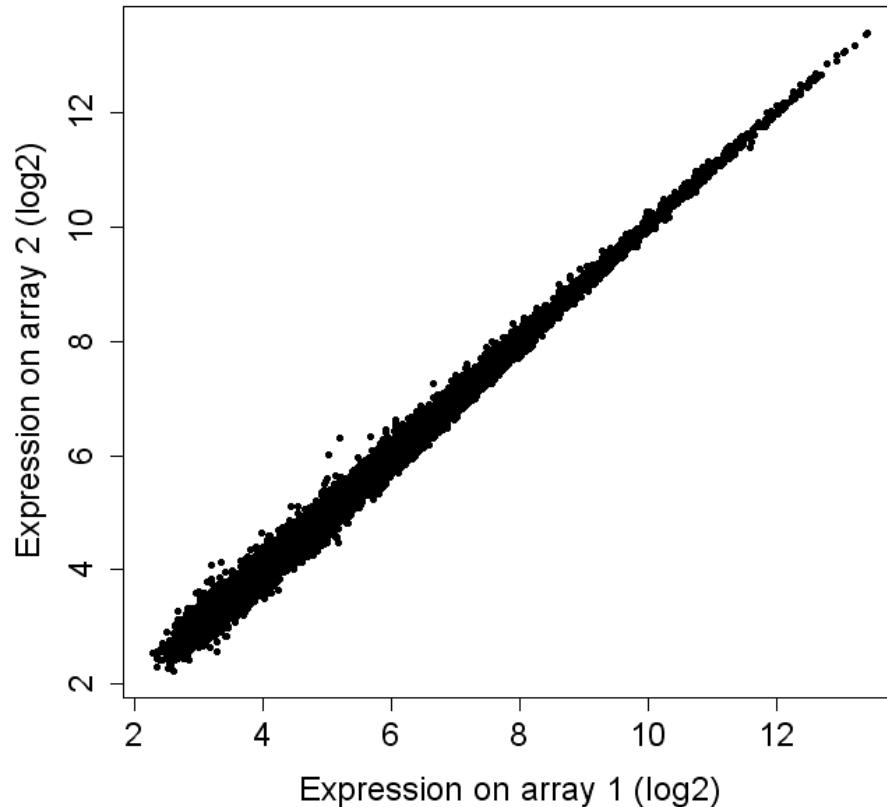
Where are the points?

Expression data from 2 arrays

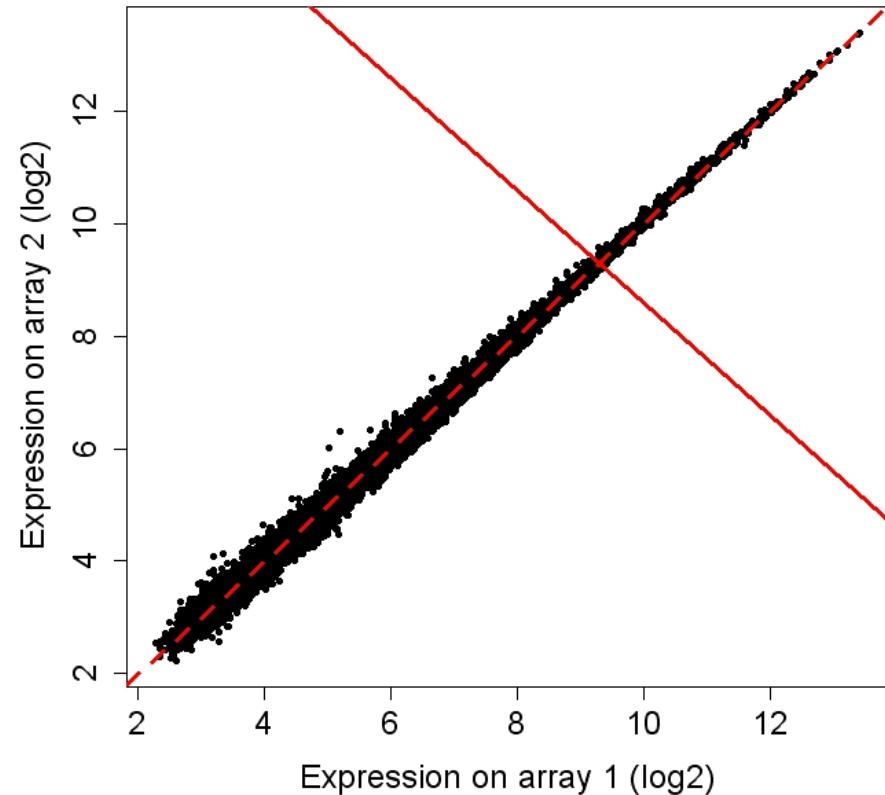


Take logs...

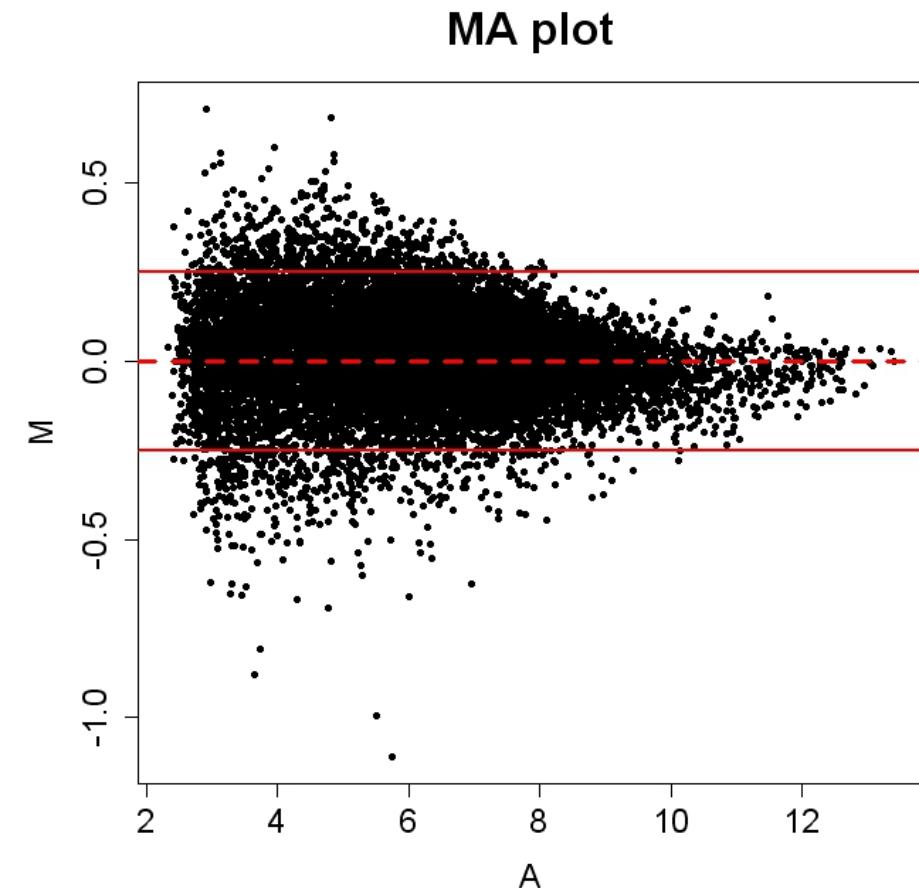
log2 Expression data from 2 arrays



log2 Expression data from 2 arrays



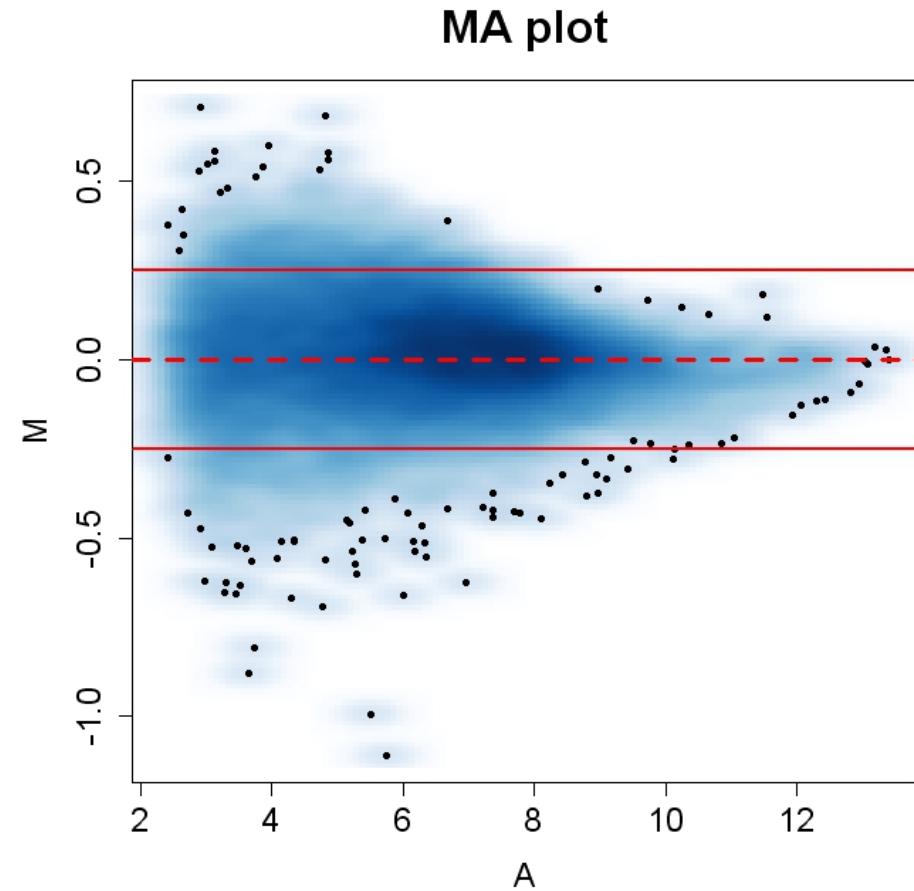
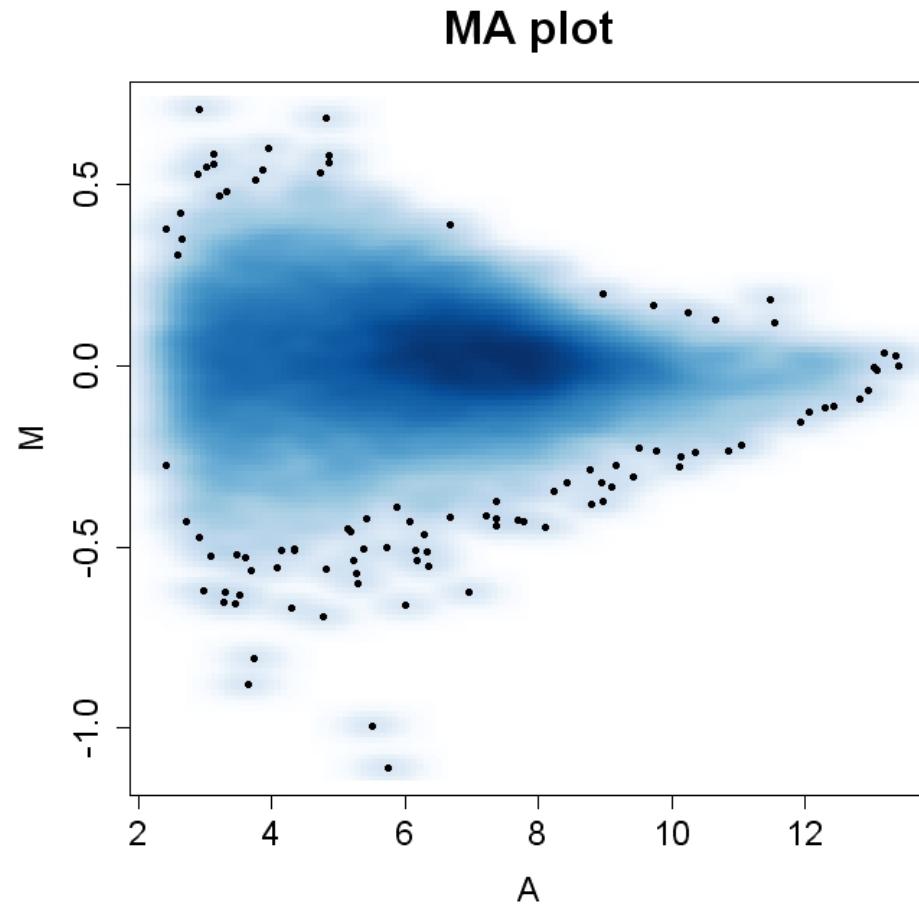
... and rotate



- $M = \text{'minus'} = \log_2(\text{expression 2}) - \log_2(\text{expression 1})$
- $A = \text{'average'} = [\log_2(\text{expression 1}) + \log_2(\text{expression 2})]/2$

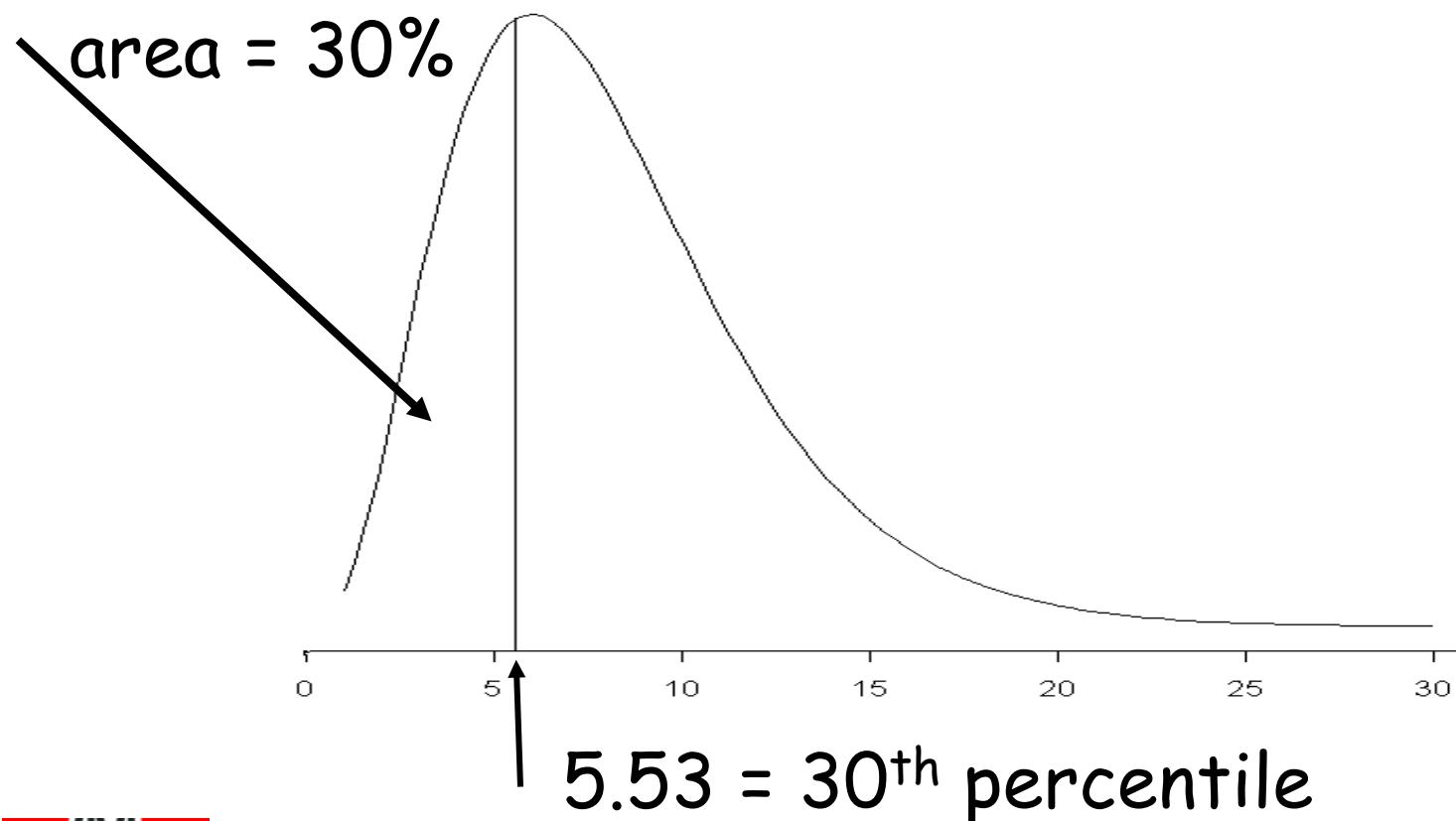


smoothScatter



Quantiles

- The p^{th} *quantile* is the number that has the proportion p of the data values smaller than it

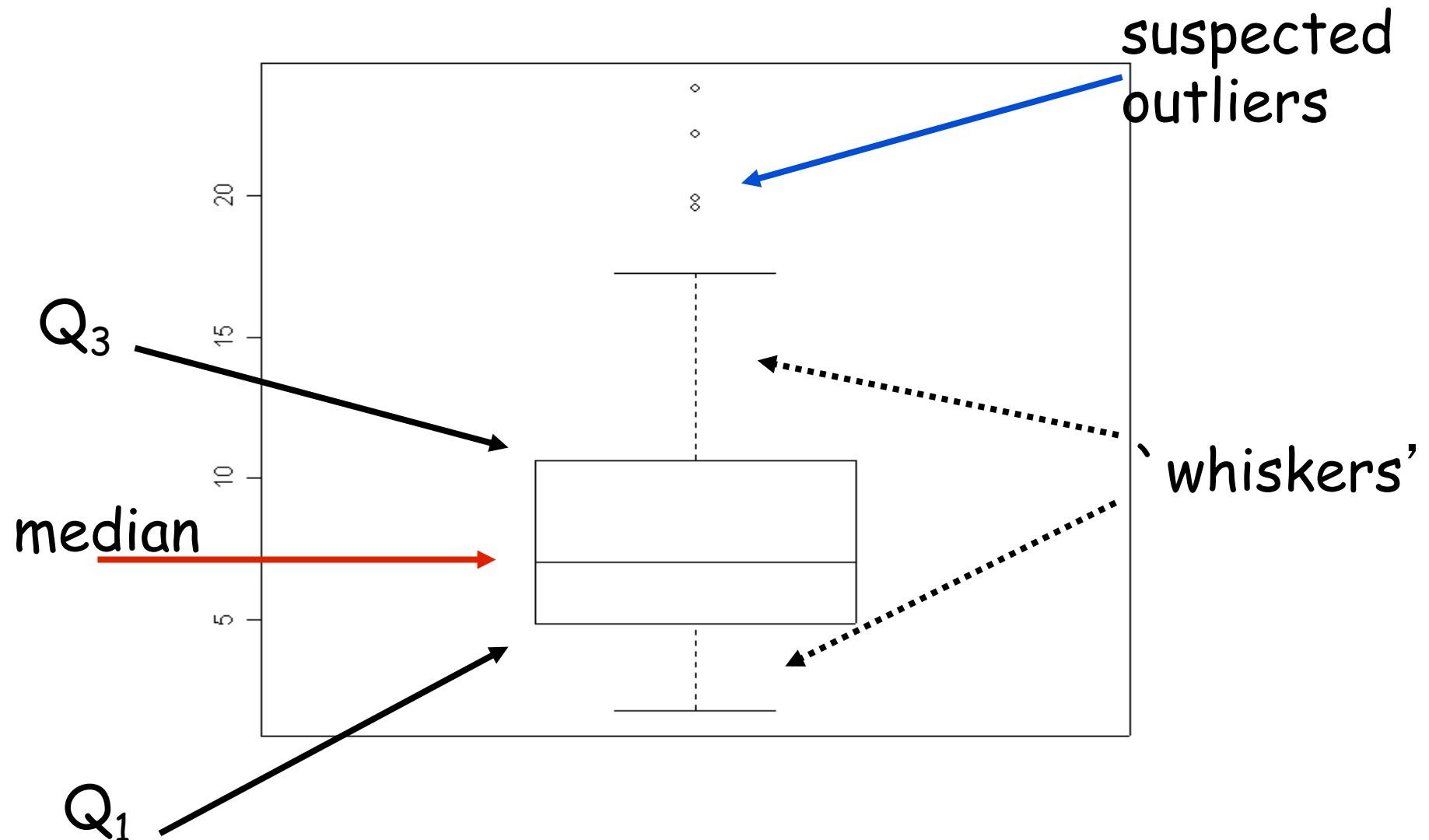


Five-number summary and boxplot

- The 25th (Q_1), 50th (median), and 75th (Q_3) percentiles divide the data into 4 equal parts; these special percentiles are called *quartiles*
- An overall summary of the distribution of variable values is given by the five values:
Min, Q_1 , Median, Q_3 , and Max
- A *boxplot* provides a visual summary of this five-number summary



Sample boxplot



Measuring expression

- *Summarize* fluorescence intensities from ~11-20 PM,MM pairs (probe level data) into *one number* for each probe set ('gene')
- Call this number a *measure of expression* (*ME*)



A few expression measures

- *MAS 5.0/GCOS* - older Affymetrix
- *PLIER* - (Hubbell, newer Affymetrix)
- *Model Based Expression Index* (MBEI)
 - Li-Wong method, implemented in **dChip** (windows executable)
- *Robust Multichip Analysis* (RMA)
 - Irizarry *et al.*, Bolstad *et al.*;
implemented in R package **affy**
 - gcrma (Wu *et al.*)
- *VSN* (Huber *et al.*, Rocke)



RMA

- Use only PM, ignore MM (variant: gcrma)
- Background correct PM on raw intensity scale
- Quantile Normalization of $\log_2(\text{PM}-\text{BG})$
- Assume additive model (on \log_2 scale):

$$\log_2 \text{normalized}(\text{PM}_{ij} - \text{BG}) = a_i + b_j + e_{ij}$$

- Estimate chip effects (log gene expression) a_i and probe effects b_j using a robust method
 - Median polish - quick
 - robust linear model - yields quality diagnostics



Why ignore MM values?

- The MM values have information about both signal and noise
- Using it without adding more noise is challenging and is a topic of current research (gcrma)
- It should be possible to improve the BG correction using MM, without having the noise level increase greatly

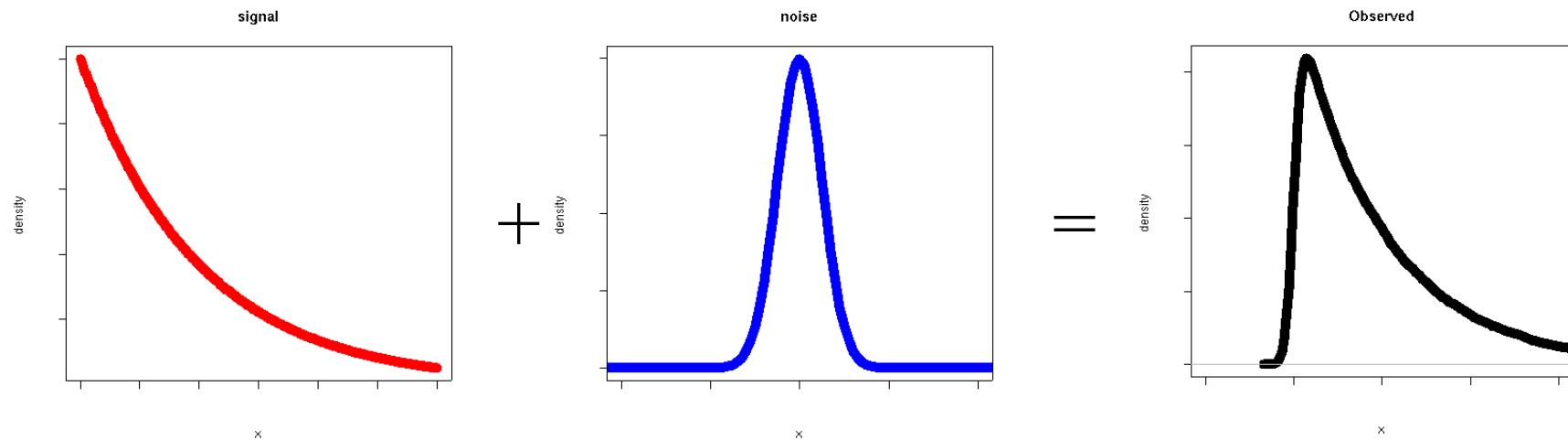


Background model

- Model observed PM intensity S as the sum of a *signal* X and *background* Y , $S=X+Y$, where
 - X is exponential (α)
 - Y is Normal (μ, σ^2)
 - X, Y independent random variables
- BG adjusted values are then $E(X|S=s)$



Background model pictorially

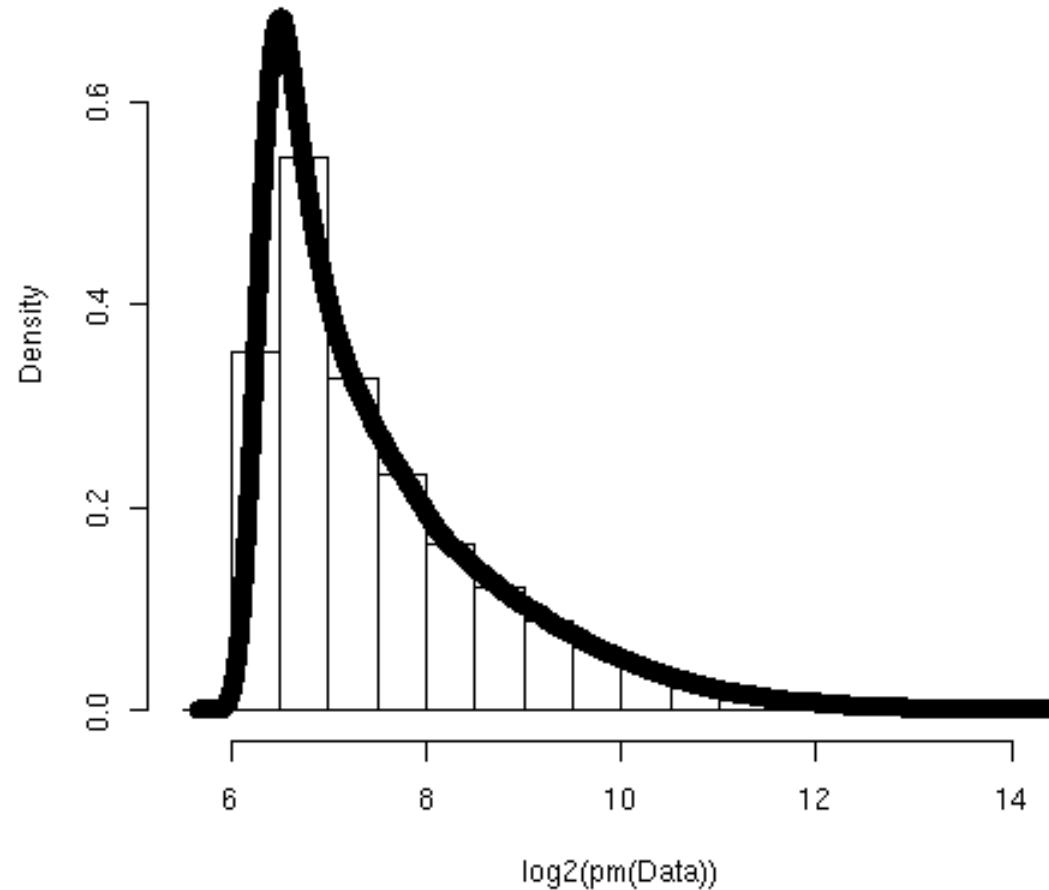


Signal + Noise = Observed



PM data on \log_2 scale

histogram of $\log(\text{PM})$ with fitted model

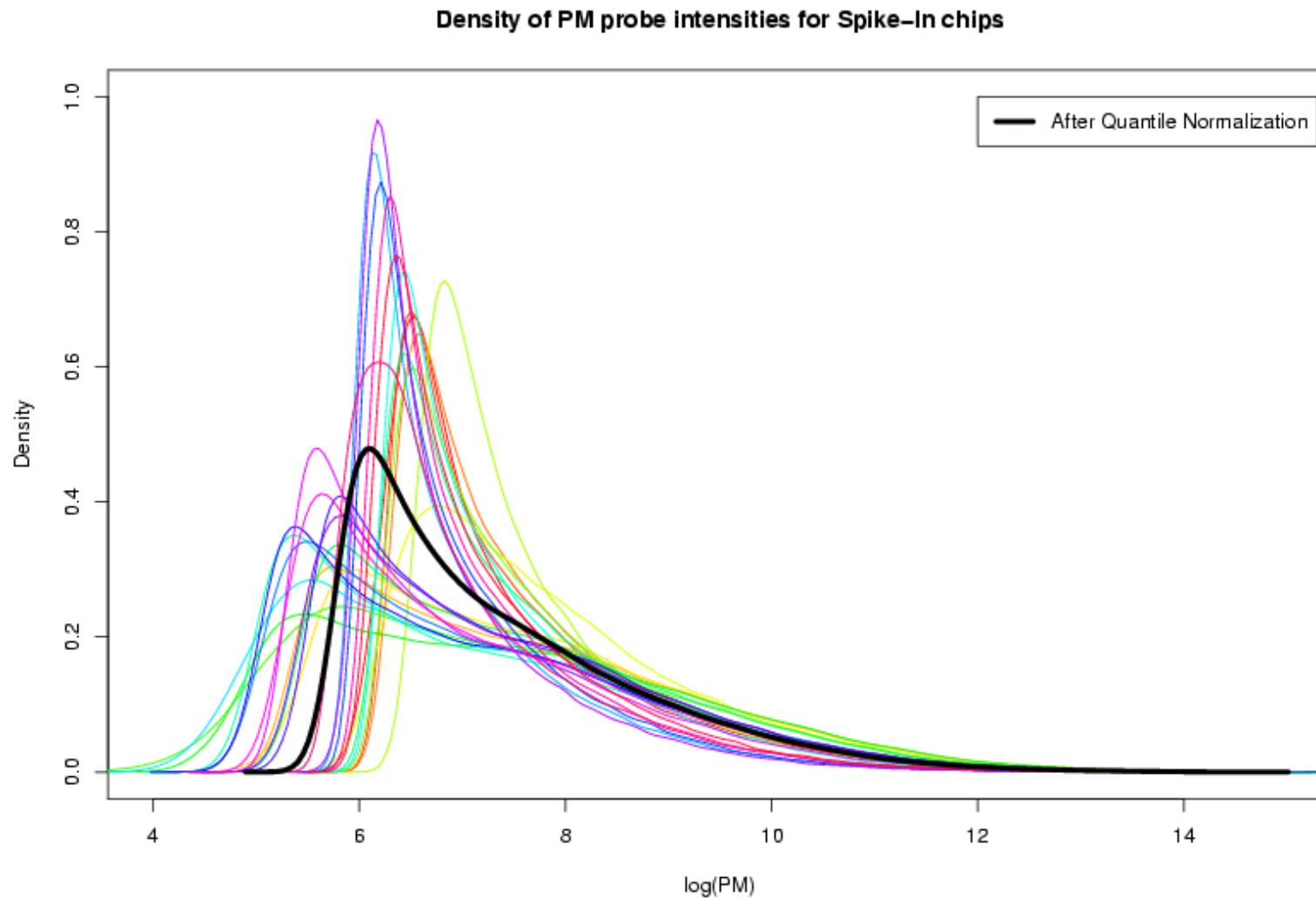


Quantile normalization

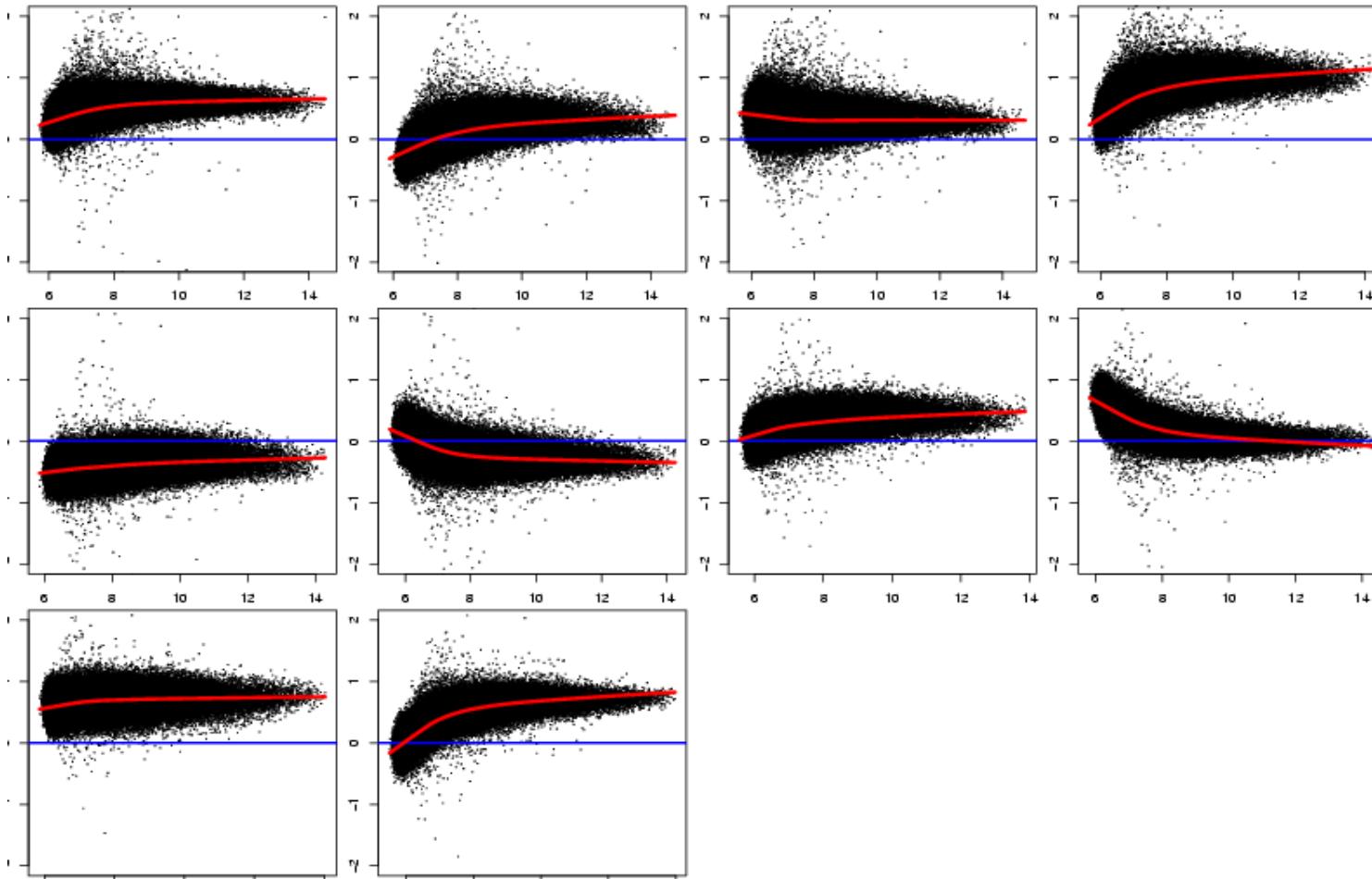
- The purpose of *normalization* is to remove artifactual differences between arrays (e.g., differences in total intensity)
- Quantile normalization makes the distribution of probe intensities *the same for every chip*
- The normalization distribution is chosen by *averaging each quantile across chips*
- After normalization, *variability of expression measures across chips reduced*
- (this results in a normalization that is probably overly conservative)



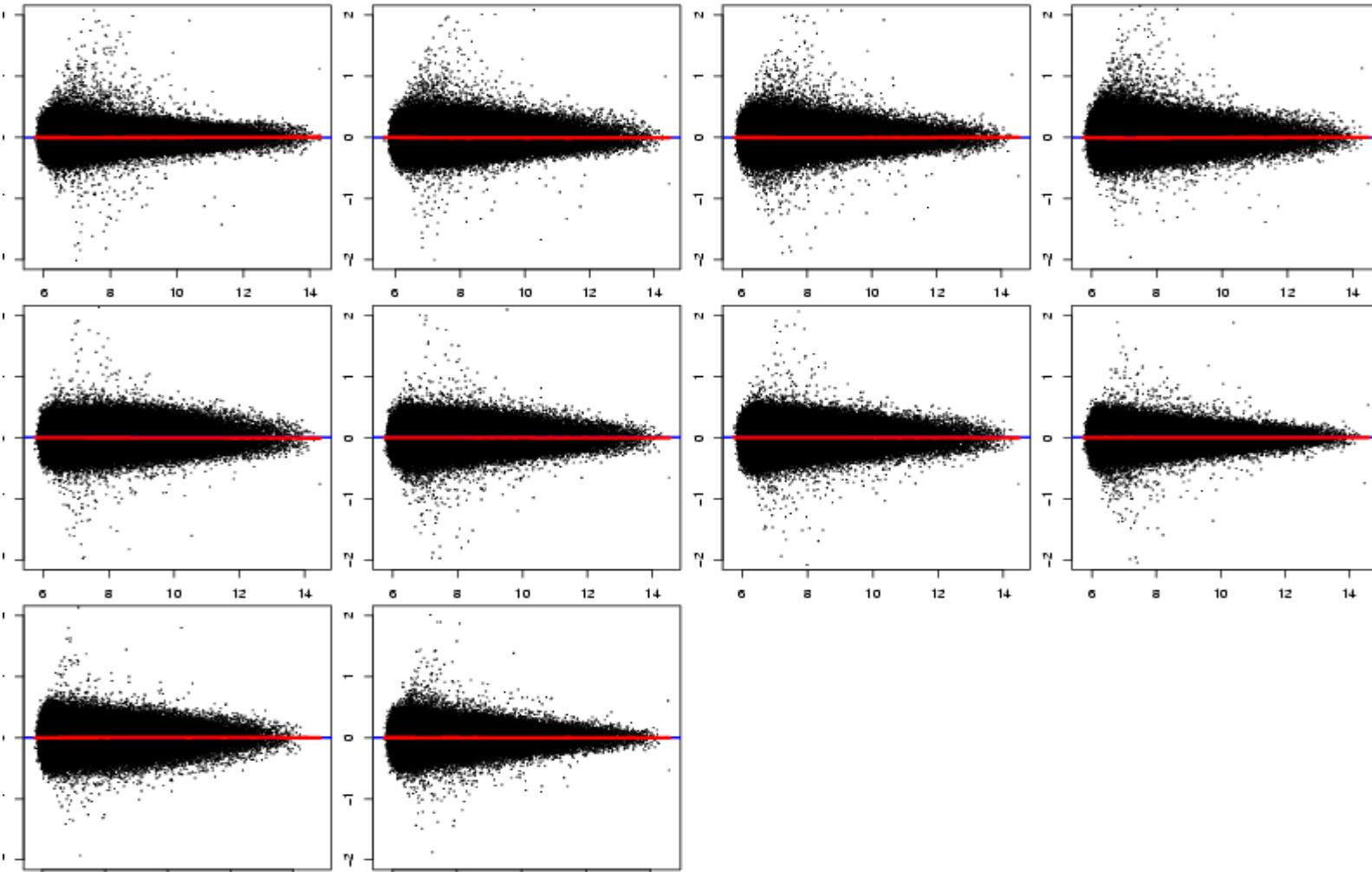
Quantile normalization: pictorially



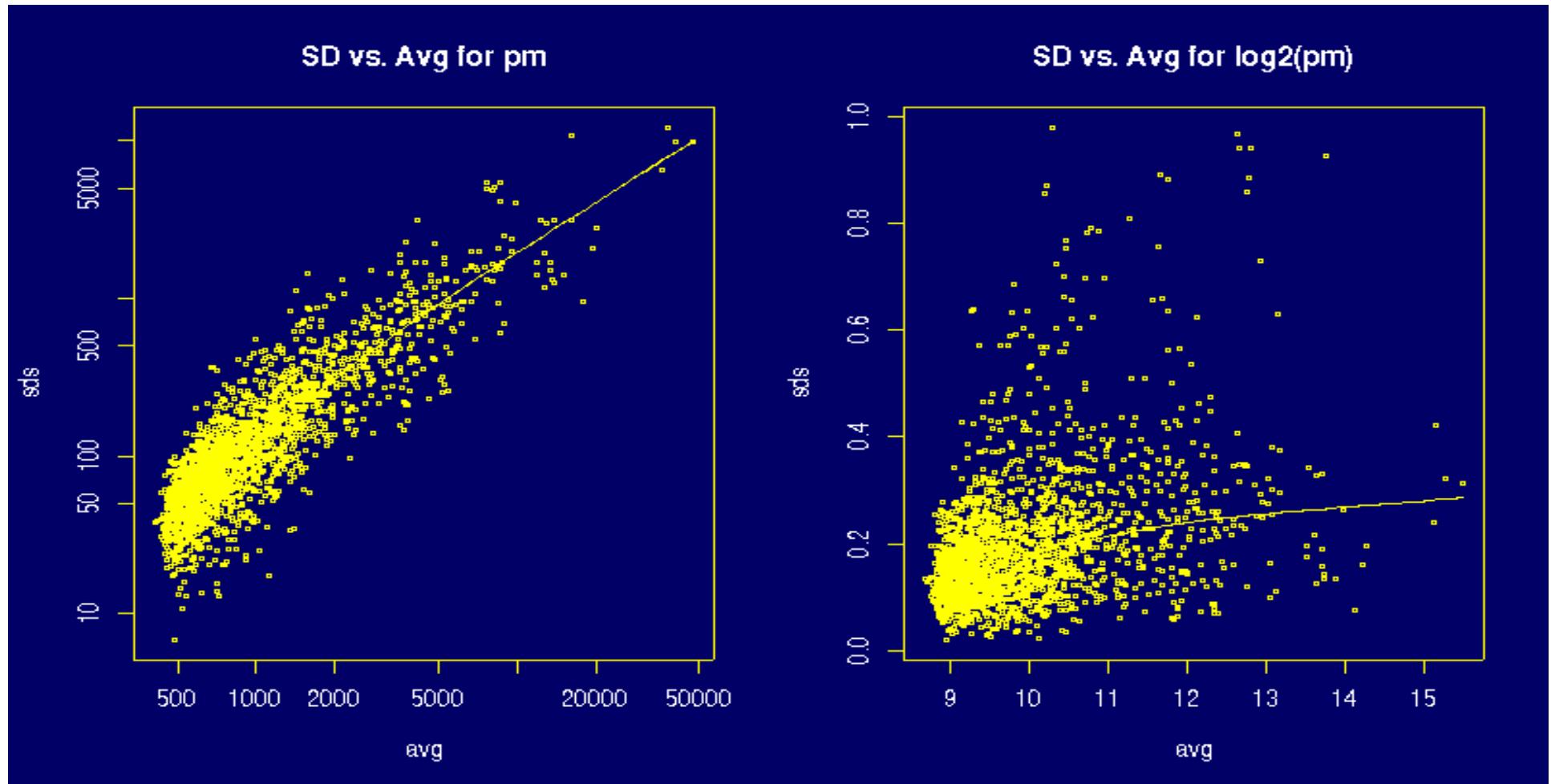
MA plots of chip pairs: before norm



MA plots of chip pairs: quantile norm



Why take \log_2

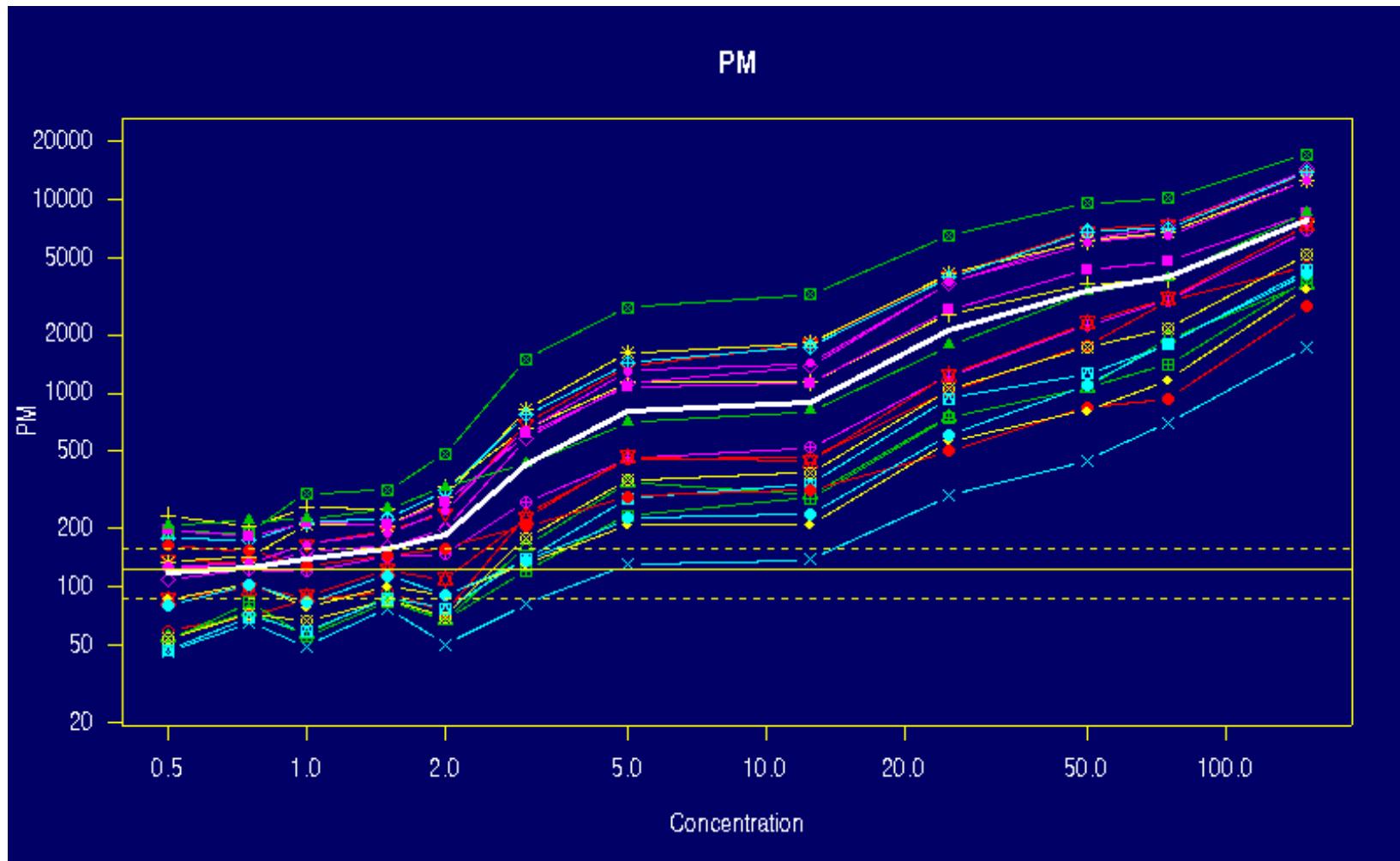


Why $\log_2 n(\text{PM} - * \text{BG}) = \text{chip effect} + \text{probe effect}$

- Spike in data set A: 11 control cRNAs spiked in (added in known amounts), all at the same concentration, which varies across 12 chips
- The example on the next slide is typical of the set of 11



Probe level data exhibiting parallel behavior on the log scale



Why Robust Multi-chip Analysis

- *Why multi-chip?*
 - To put each chip's values in the *context* of a set of similar values
 - helps even if not done robustly
- *Why robust?*
 - robust summaries improve over standard ones by *down-weighting outliers*



Robust Multi-chip Analysis

- Base analysis on the linear model embodying the parallel behavior:

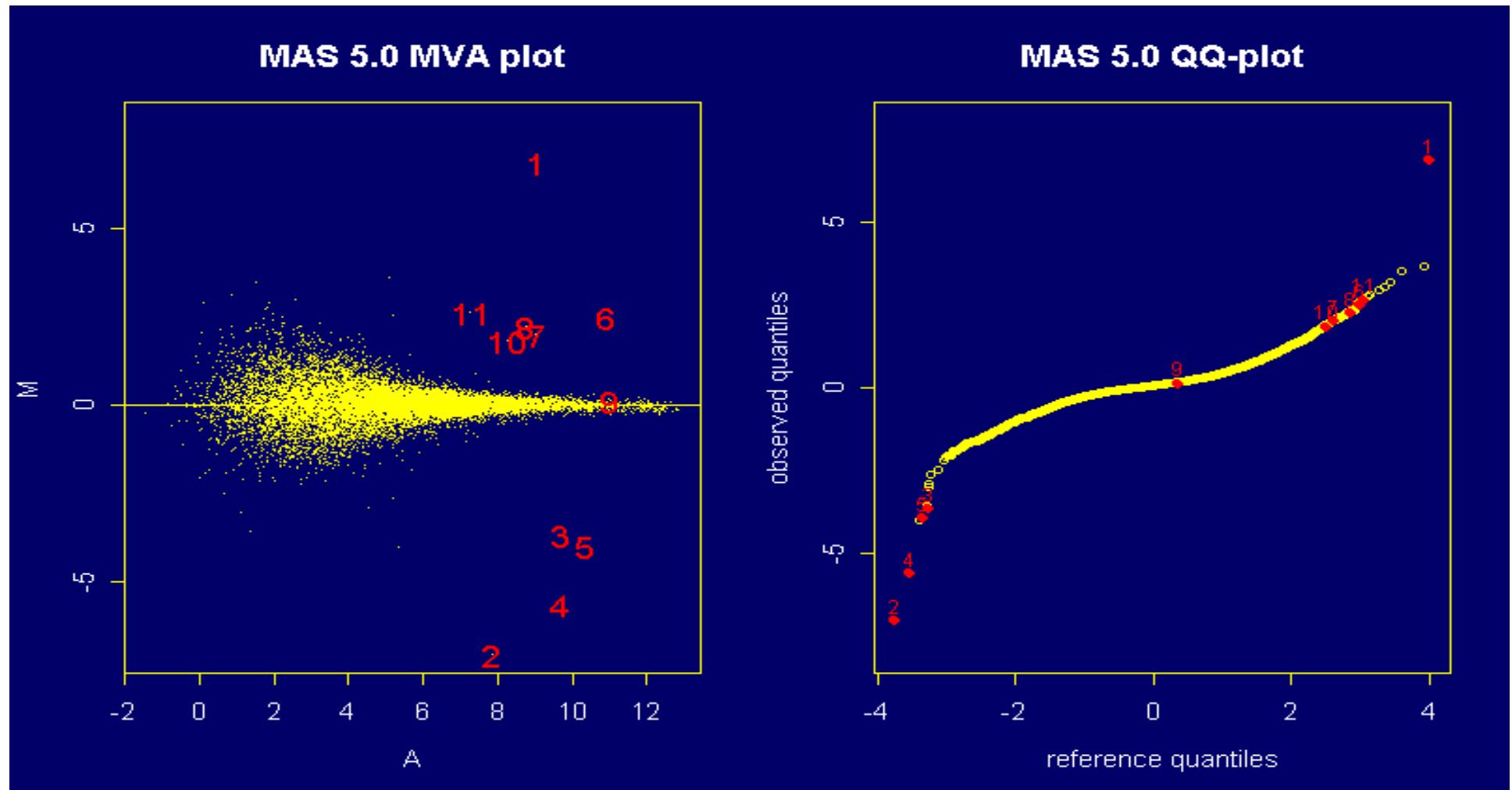
$$\log_2 n(\text{PM}_{ij} - * \text{BG}) = \text{chip effect}_i + \text{probe effect}_j + \varepsilon_{ij}$$

where i labels chips and j labels probes ('genes')

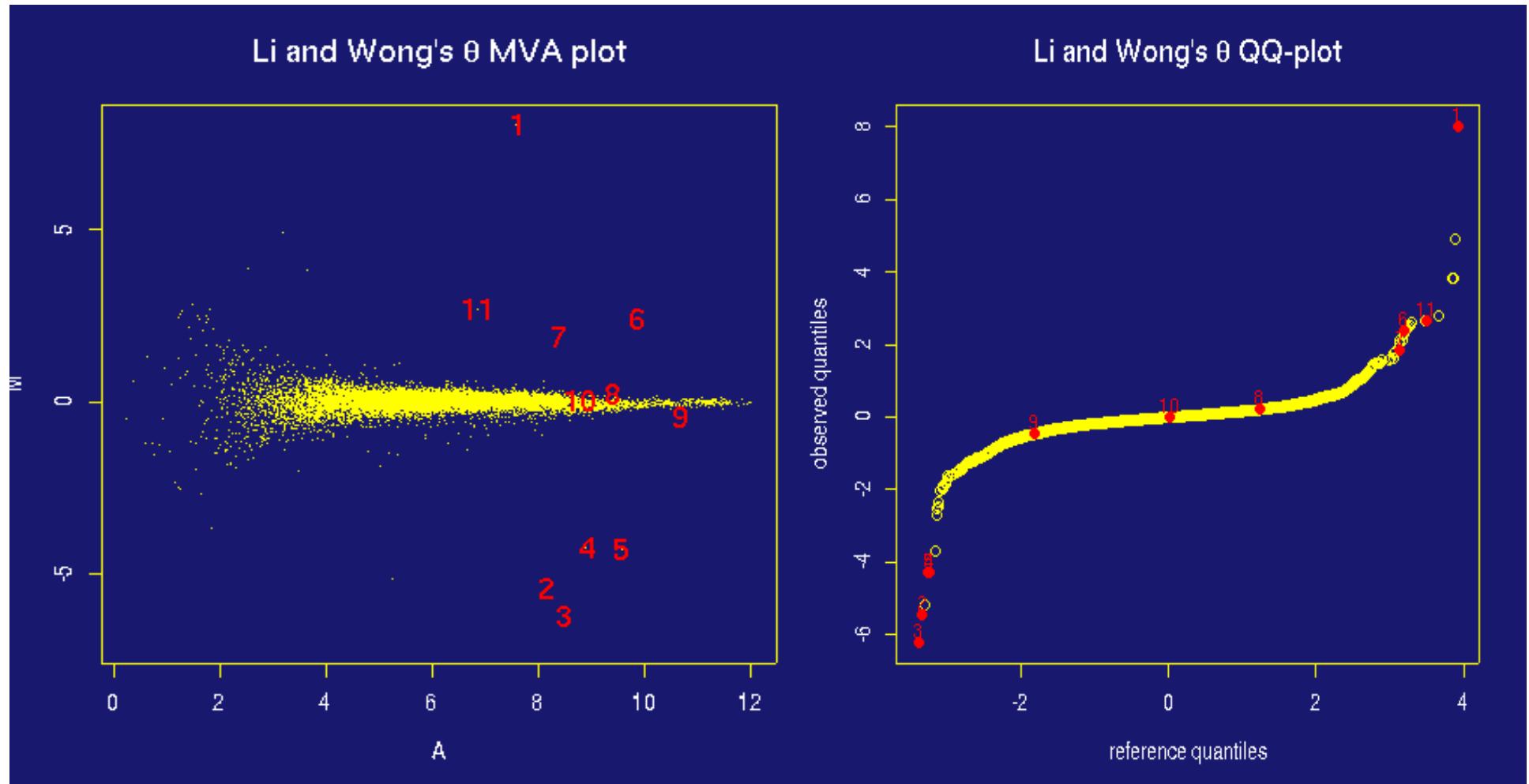
- RMA implementation estimates parameters using *median polish* (it's faster than IRLS)



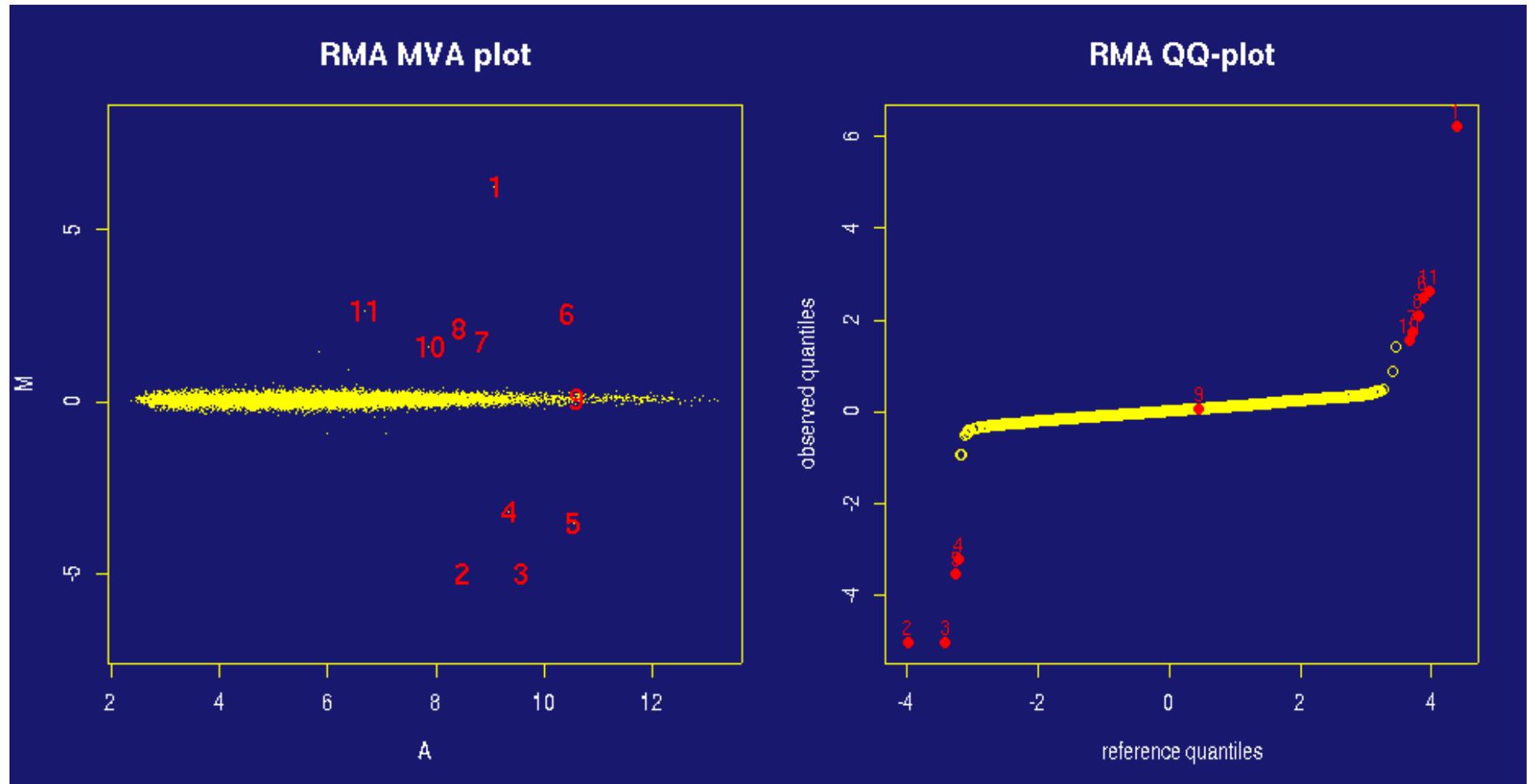
Differential expression: MAS 5.0



Differential expression: Li-Wong



Differential expression: RMA



Bias-variance tradeoff

- MAS 5.0 has less *bias* (for estimating fold change) in comparison with RMA and dChip
- The problem is that it pays a very large price in *extra variability* for this low bias
 - $MSE = \text{bias}^2 + \text{variance}$
- but ... 0 + large > small + small
- Overall, a little bias but greatly reduced variance seems better
- (There is also much more evidence)



Conclusions of Irizarry *et al.*

- Studied a number of ME on specially designed experiments (spike-in, dilution series)
- Use normalized $\log_2(\text{PM-BG})$
- Using global background improves on use of probe-specific MM* (but...gcrma)
- Gene Logic spike-in and dilution study show technology works well
- *RMA was arguably the best summary in terms of bias, variance and model fit*





Affycomp III

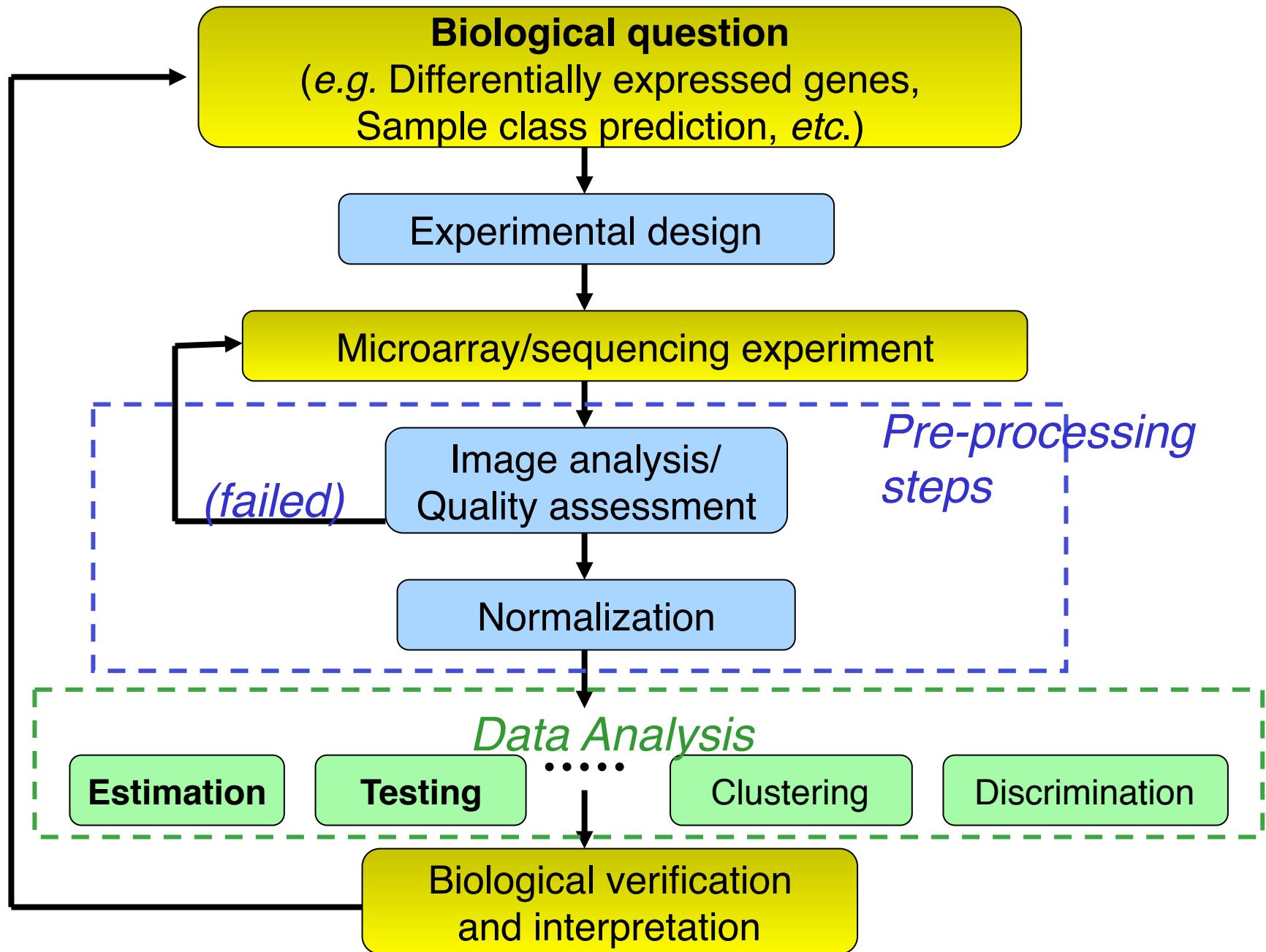
A Benchmark for Affymetrix GeneChip Expression Measures

- [The advent of Affycomp III](#)
 - [Background](#)
 - [Data and instructions](#)
 - [Submission form](#)
- [New assessments](#) (of *SPIKE-IN HGU95* and *HGU133* studies)
- [Entry comparison / downloads](#)
- Old assessments
 - [Affycomp II](#) (of *SPIKE-IN HGU95* and *HGU133*)
 - [old entry comparison tool](#)
 - [original assessments](#) (of *DILUTION* and *HGU95*)
- [Papers](#)
 - *A Benchmark for Affymetrix GeneChip Expression Measures*, *Bioinformatics*, Vol 20, No 3, 2004, 323-331
 - *Comparison of Affymetrix GeneChip Expression Measures*, *Bioinformatics*, Vol 1, No 1, 2005, 1-7
- [The affycomp R package](#)
 - [FAQ](#) (in prep)
 - [old FAQ](#)
 - [Contact us](#)



(BREAK)





cDNA gene expression data

Data on G genes for n samples:
mRNA samples

		sample1	sample2	sample3	sample4	sample5	...
Genes	1	0.46	0.30	0.80	1.51	0.90	...
	2	-0.10	0.49	0.24	0.06	0.46	...
	3	0.15	0.74	0.04	0.10	0.20	...
	4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	5	-0.06	1.06	1.35	1.09	-1.09	...

Gene expression level of gene i in mRNA sample j

2-color (e.g. cDNA) = M = normalized $\log_2(\text{Red}/\text{Green})$

1-color (e.g. Affy) = RMA



Identifying Differentially Expressed Genes (IDE)

- **Goal:** Identify genes associated with covariate or response of interest
- **Examples:**
 - Qualitative covariates or factors: treatment, cell type, tumor class
 - Quantitative covariate: dose, time
 - Responses: survival, cholesterol level
 - Any combination of these!



Informal methods

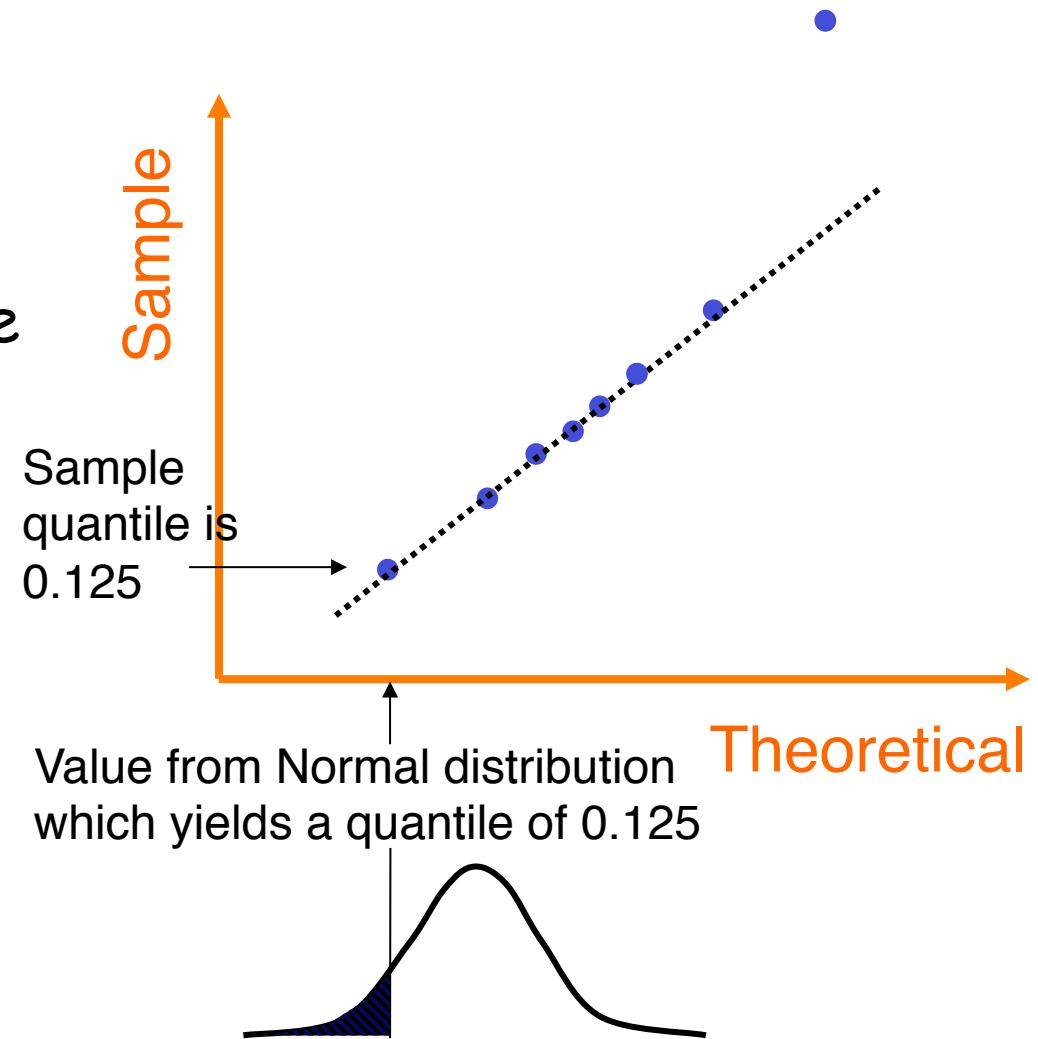
- If *no replication* (i.e. only have a single array for each condition), not many options!
- Common methods include:
 - (log) Fold change exceeding some threshold, e.g. more than 2 (or less than -2)
 - Graphical assessment, e.g. QQ plot
- Threshold for DE is pretty arbitrary



QQ-Plots

Used to assess whether a sample follows a particular (e.g. normal) distribution (or to compare two samples)

A method for looking for outliers when data are mostly normal

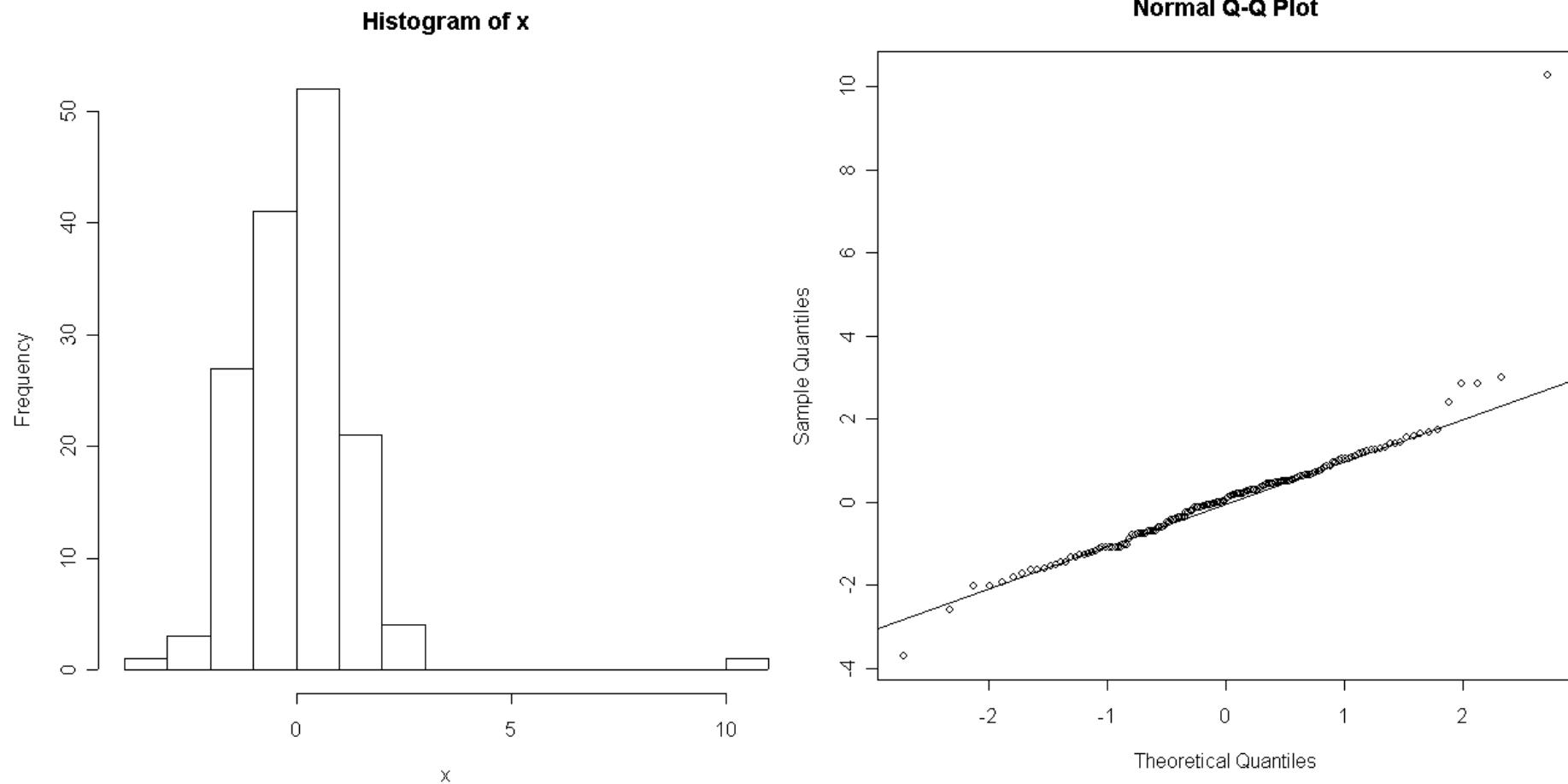


Typical deviations from straight line patterns

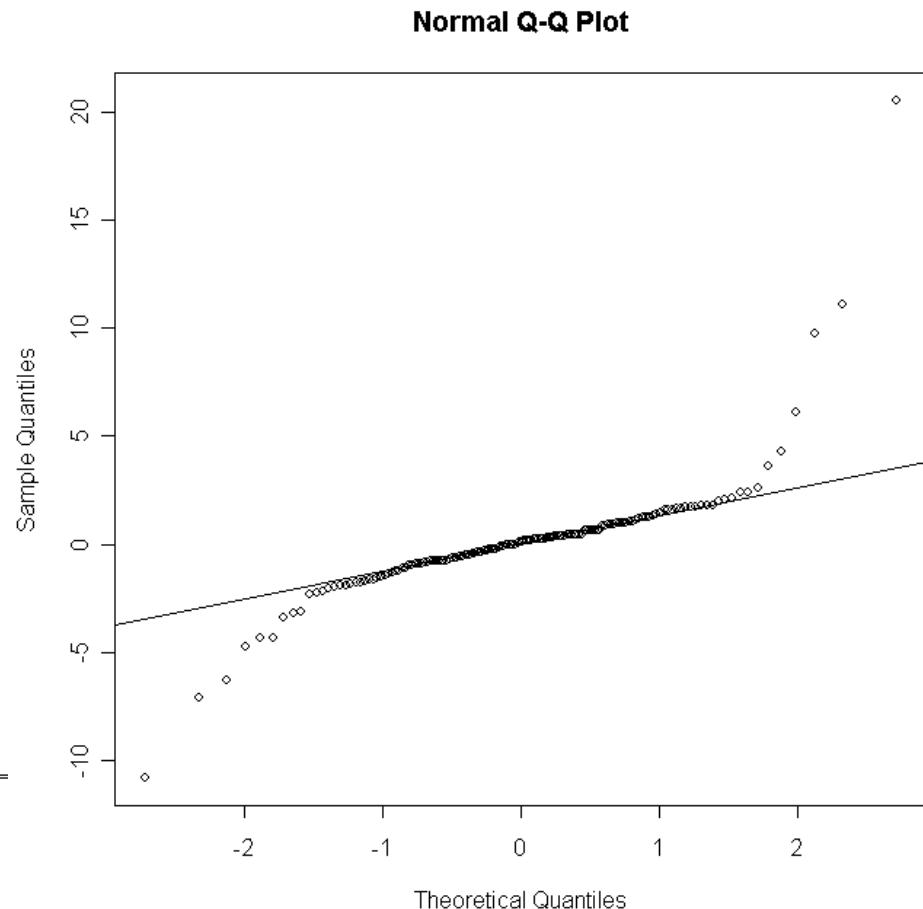
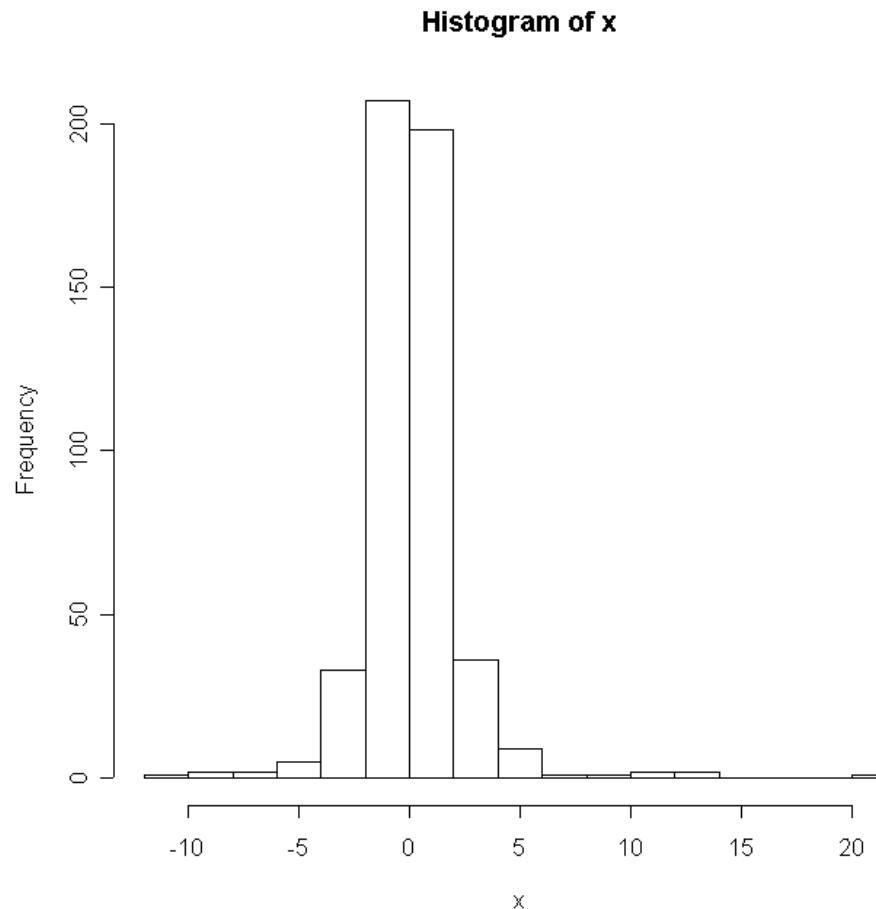
- Outliers
- Curvature at both ends (long or short tails)
- Convex/concave curvature (asymmetry)
- Horizontal segments, plateaus, gaps



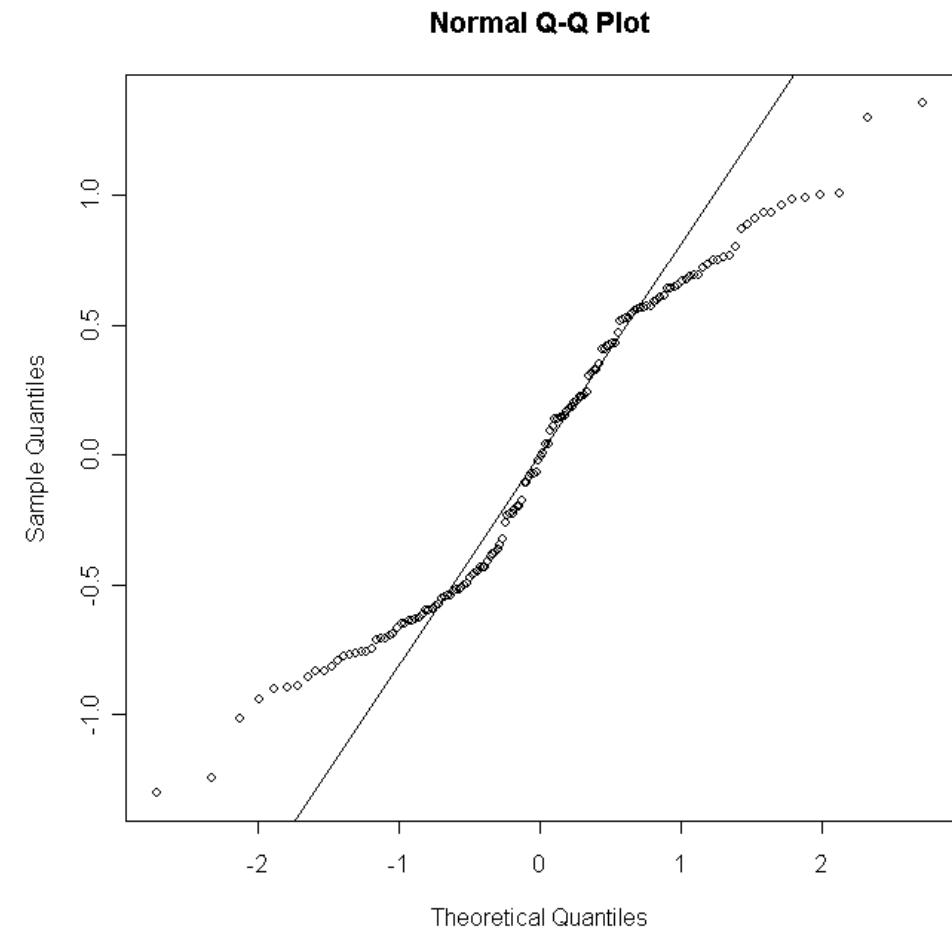
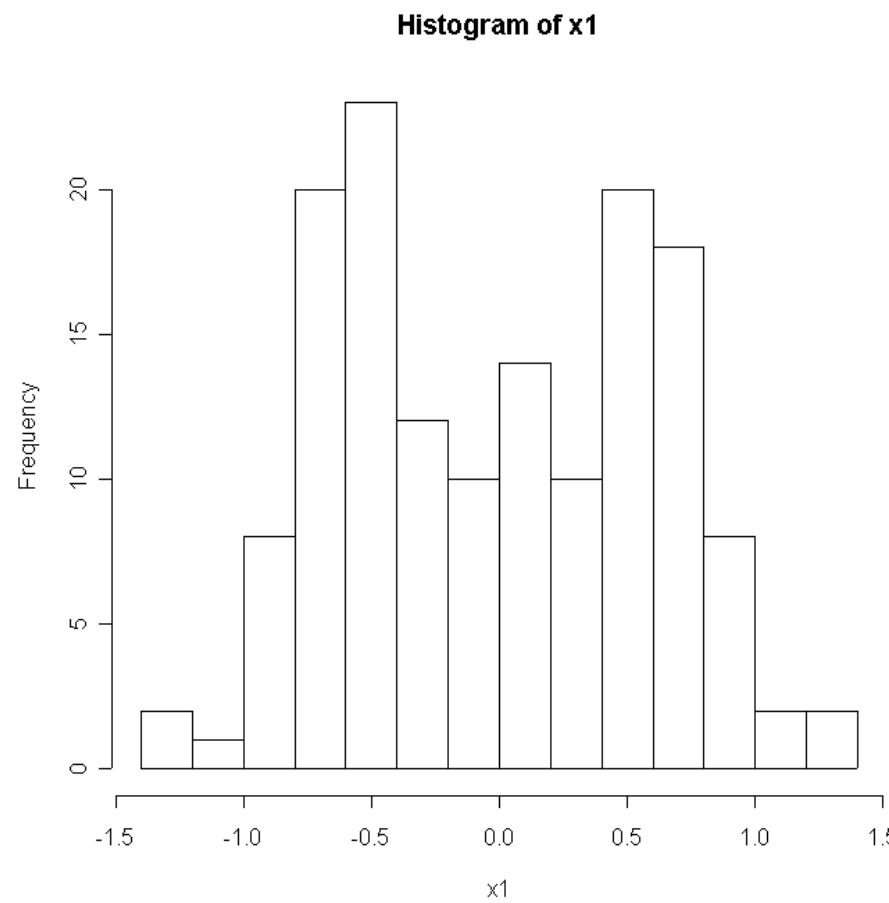
Outliers



Long Tails

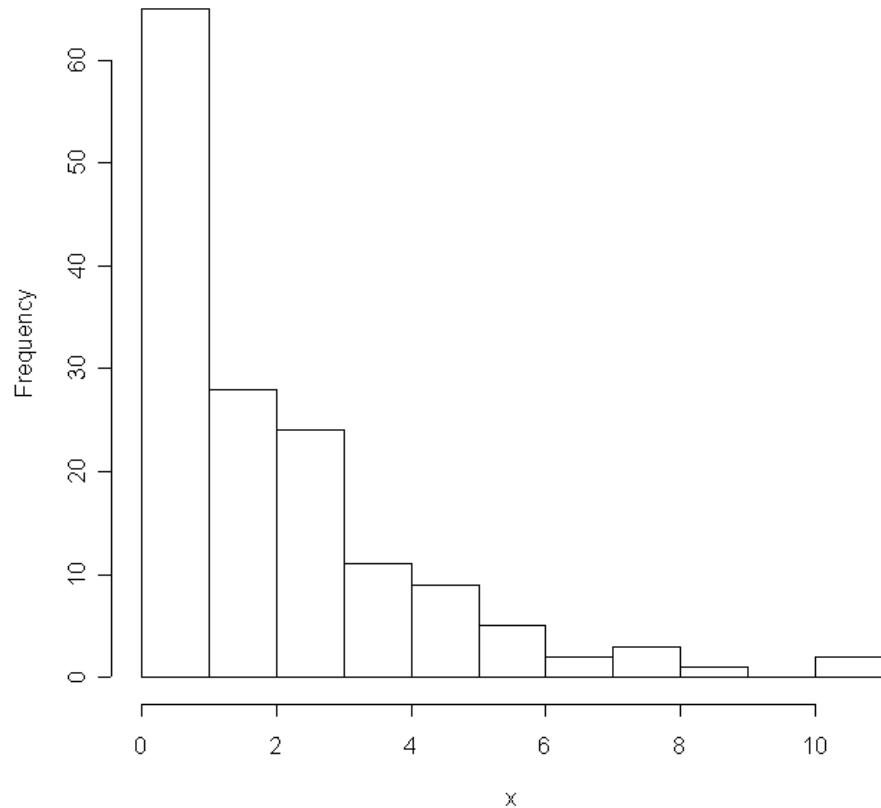


Short Tails

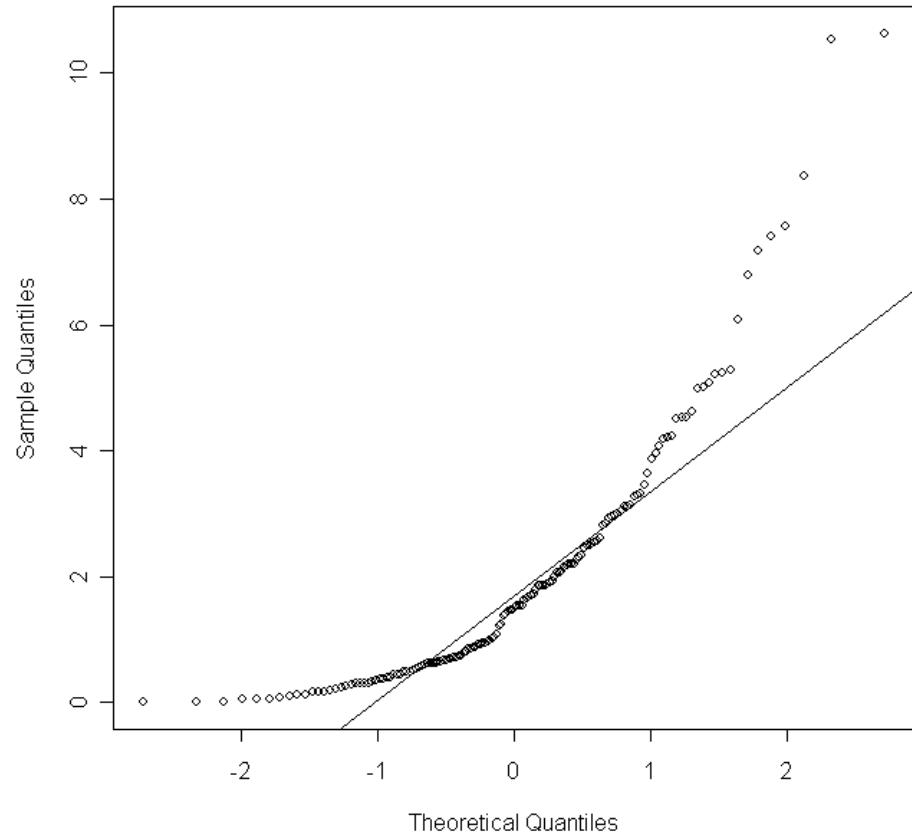


Asymmetry

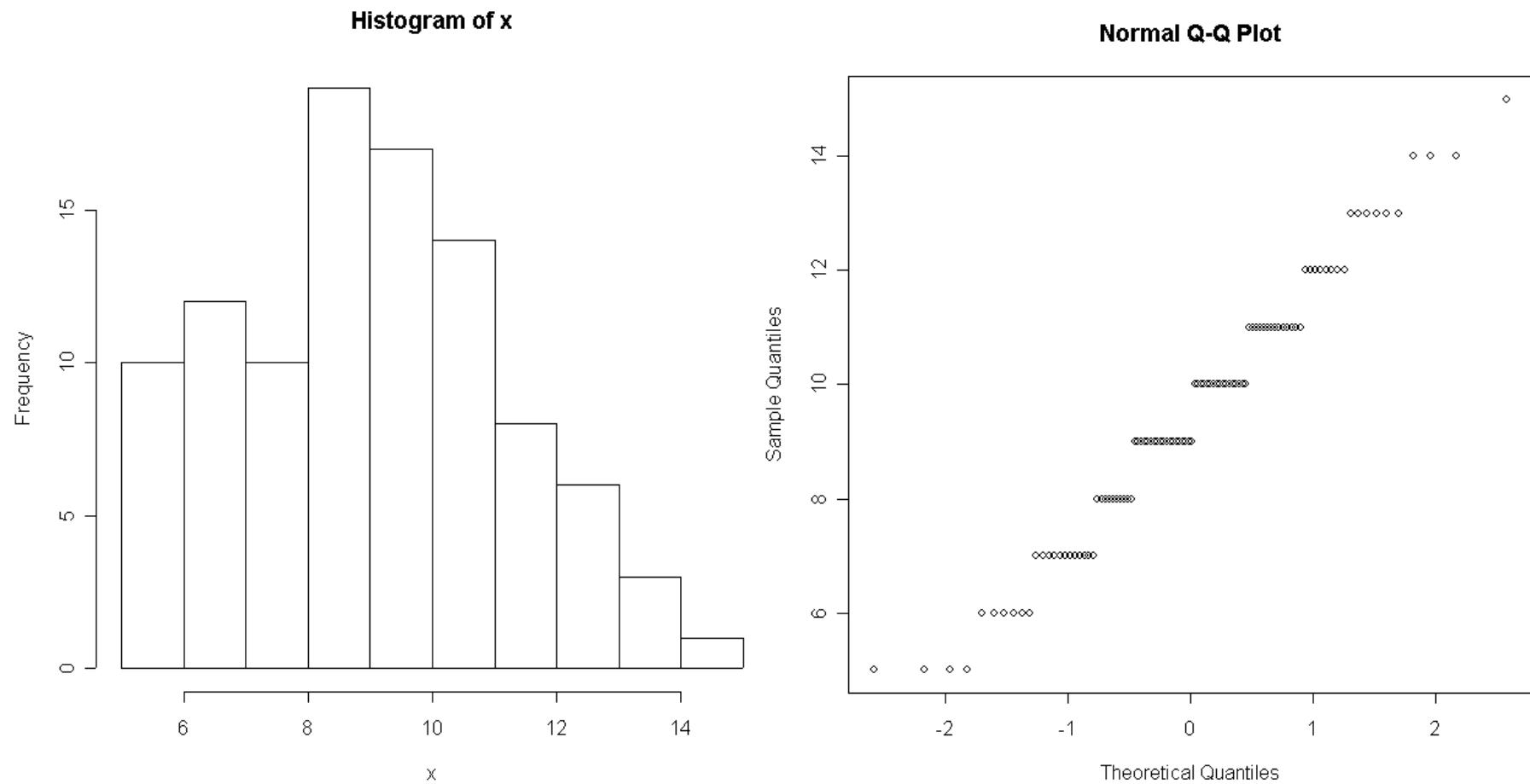
Histogram of x



Normal Q-Q Plot

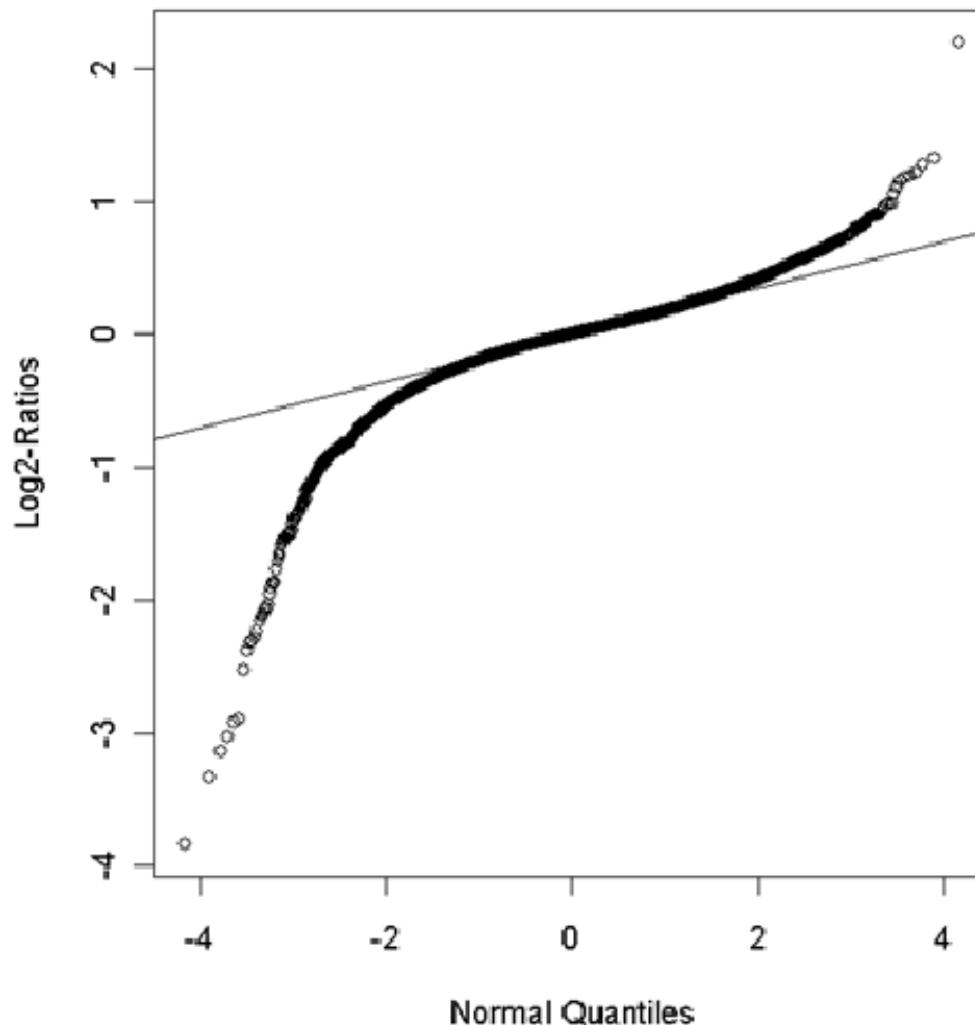


Plateaus/Gaps



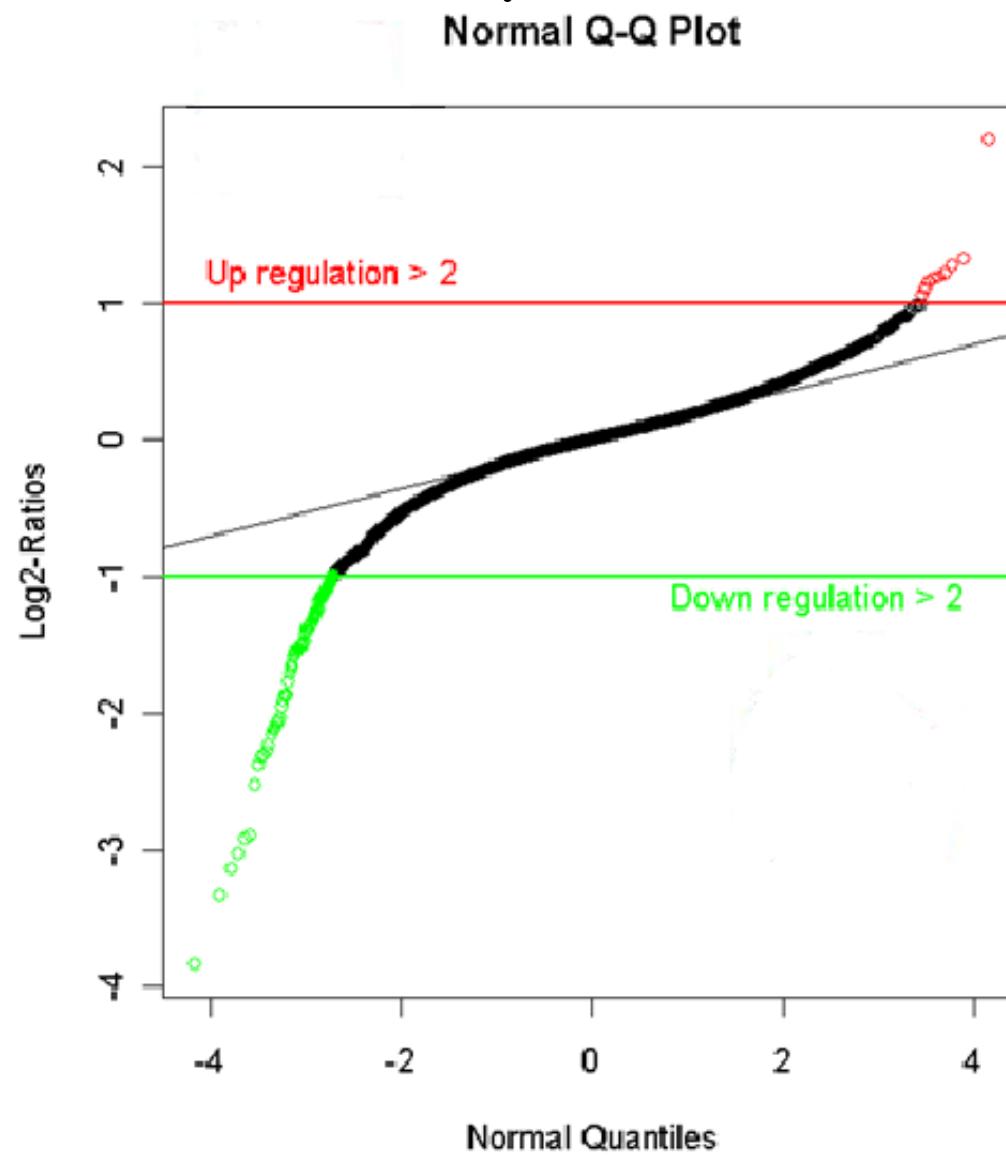
QQ Plot

Normal Q-Q Plot



DE in a QQ plot

In this case,
the two
conditions are
the same - i.e.
NO genes are
truly DE!



Replicated experiments

- Have *replicates* for each condition
- Then can use statistical methods
- *Summarize* difference of averages for each gene by
 - $M = \text{average (Treatment)} - \text{average (Control)}$
 - $s = \text{SE}(M \text{ values})$
- *Rank* genes in order of strength of evidence in favor of DE
- How might we do this??



Which genes are DE?

- Difficult to judge significance
 - massive *multiple testing* problem
 - genes *dependent*
 - don't know null distribution of M
- Strategy
 - aim to *rank* genes
 - assume most genes are not DE (depending on type of experiment and array)
 - find genes *separated* from the majority



Ranking criteria

- Genes $i = 1, \dots, p$
- $M_i = \log_2$ fold change for gene i
 - *Problem*: genes with *large variability* likely to be selected, even if not truly DE
- Take variability into account: use $t_i = M_i / (s_i / \sqrt{n})$
 - *Problem*: genes with extremely small variances make very large t
 - Genes with small fold-change might not be biologically interesting
 - When the number of replicates is small, the smallest s_i are *likely to be underestimates* (too few degrees of freedom)



Shrinkage estimators

- Idea: *borrow information* across genes
- Here, we ‘shrink’ the t_i towards zero by modifying the s_i in some way (get s_i^*)
- mod $t_i = t_i^* = M_i / (s_i^*/\sqrt{n})$

$$t_i \longleftrightarrow t_i^* \longleftrightarrow M_i$$

- Many ways to get s_i^*
- We will use the version implemented in the BioConductor package **limma**

Moderated t -statistics (Smyth)

- Using empirical Bayesian approach to estimate:
- Overall variability estimate s_0^2
- Per-gene variability estimate s_g^2
- Shrinkage variability: $\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$
- *Contrast* estimator $\hat{\beta}_g$ (*difference in means* between two groups)
- Moderated t -statistics: $\tilde{t}_g = \frac{\hat{\beta}_g}{\sqrt{\tilde{s}_g v_g}}$
- (v_g = Factor in covariance matrix of linear model estimate)



Linear modeling

- We will cover this in greater detail in a few weeks when we look at experimental design
- For now, it will be ok to follow along the example in the **limma** user manual
- (This will be part of the TP next Monday)
- Details about mod-t statistics in Smyth's paper:

<http://www.statsci.org/smyth/pubs/ebayes.pdf>



An empirical Bayes (EB) story

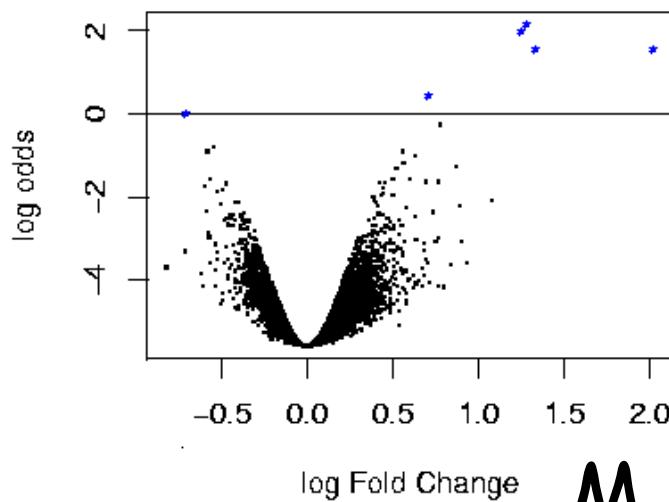
- M_{ij} (fold change) $\sim N(\mu_i, \sigma_i^2)$
- Proportion p of genes have $\mu_i \neq 0$ (i.e. are DE)
- Normal prior on nonzero μ_i
- Inverse-gamma prior on σ_i^2
- The priors on μ_i and σ_i^2 involve *hyperparameters* (parameters for the priors of the parameters)
- In EB estimation, the hyperparameters are *estimated from the data*
- (Lönnstedt and Speed): For each gene, compute *posterior log odds* that gene is DE:

$$B = \log[P(\mu_i \neq 0) / P(\mu_i = 0)]$$

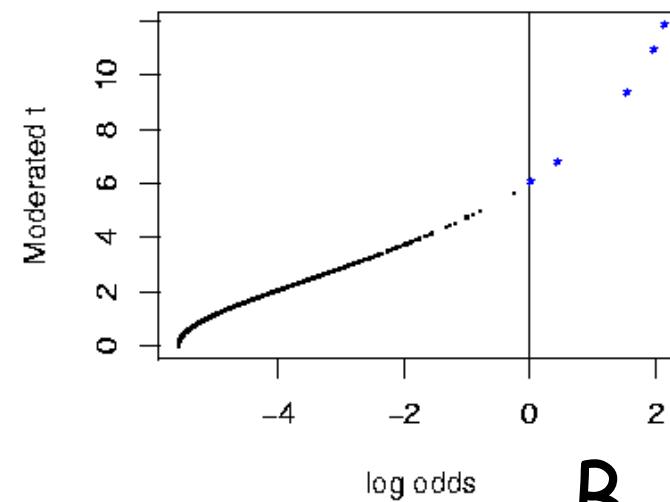


$M, B, \text{mod } t, |\bar{t}|$

B



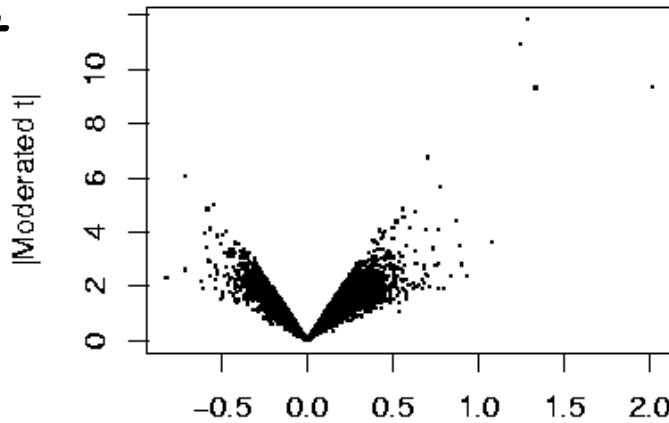
M



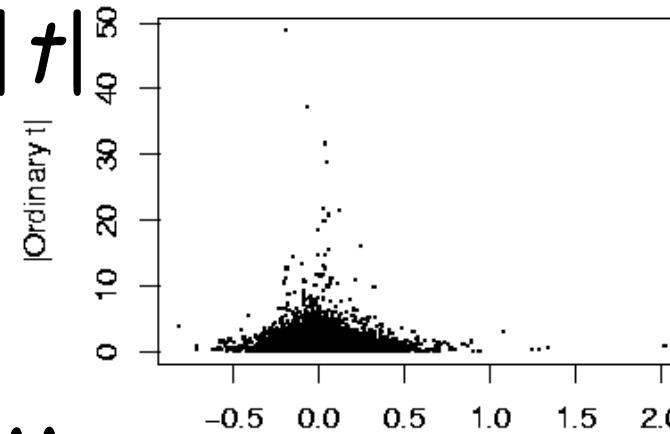
B

mod- t

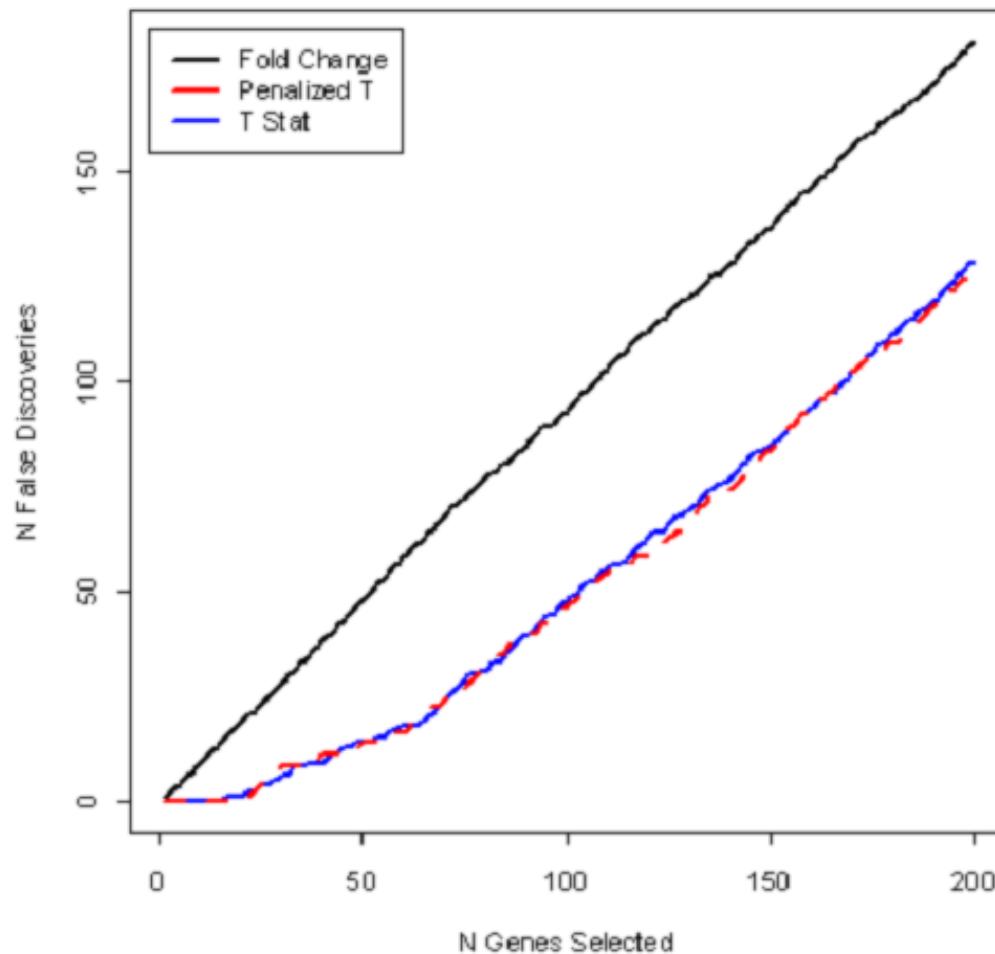
mod- t



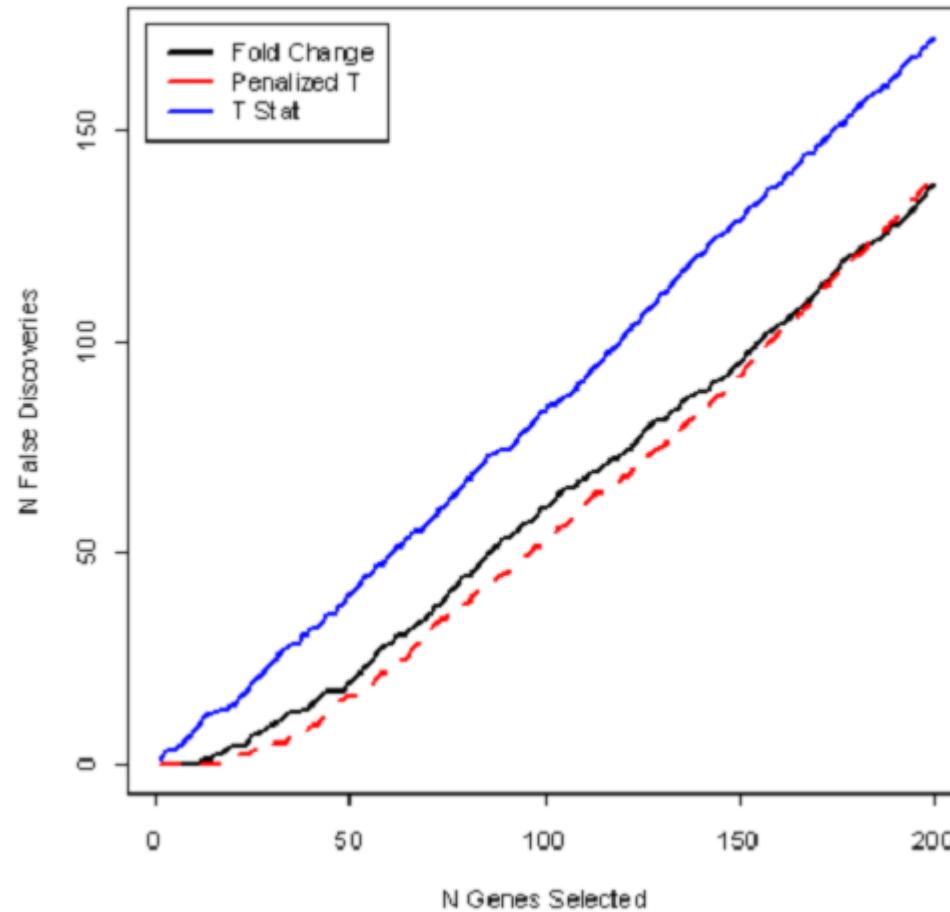
M



Simulation study: genes with different SDs



Simulation study: genes with similar SDs



Simulation study: genes with different SDs, small number of arrays

