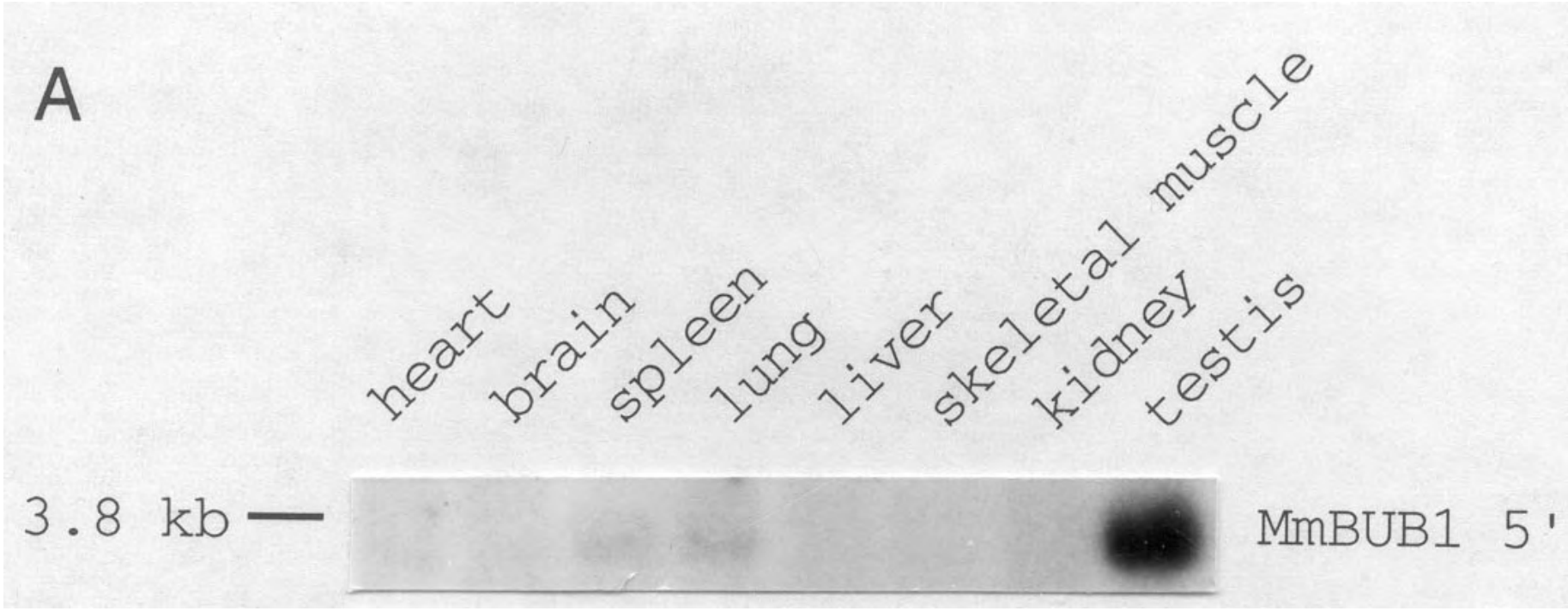
Statistics Applied To Microarray Data

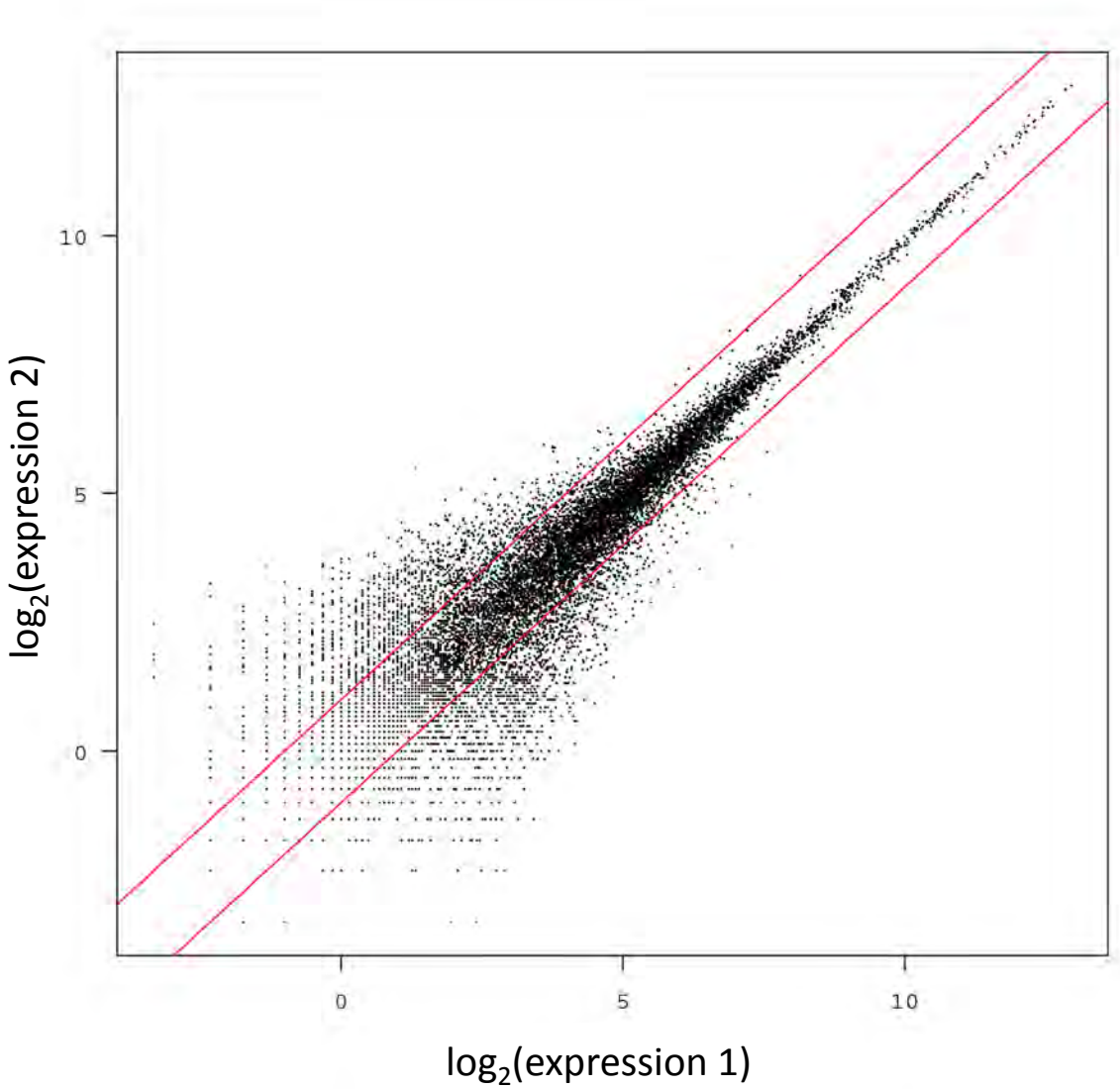
Outline

- Detecting differentially expressed genes
- Preprocessing
- Probe and batch effects

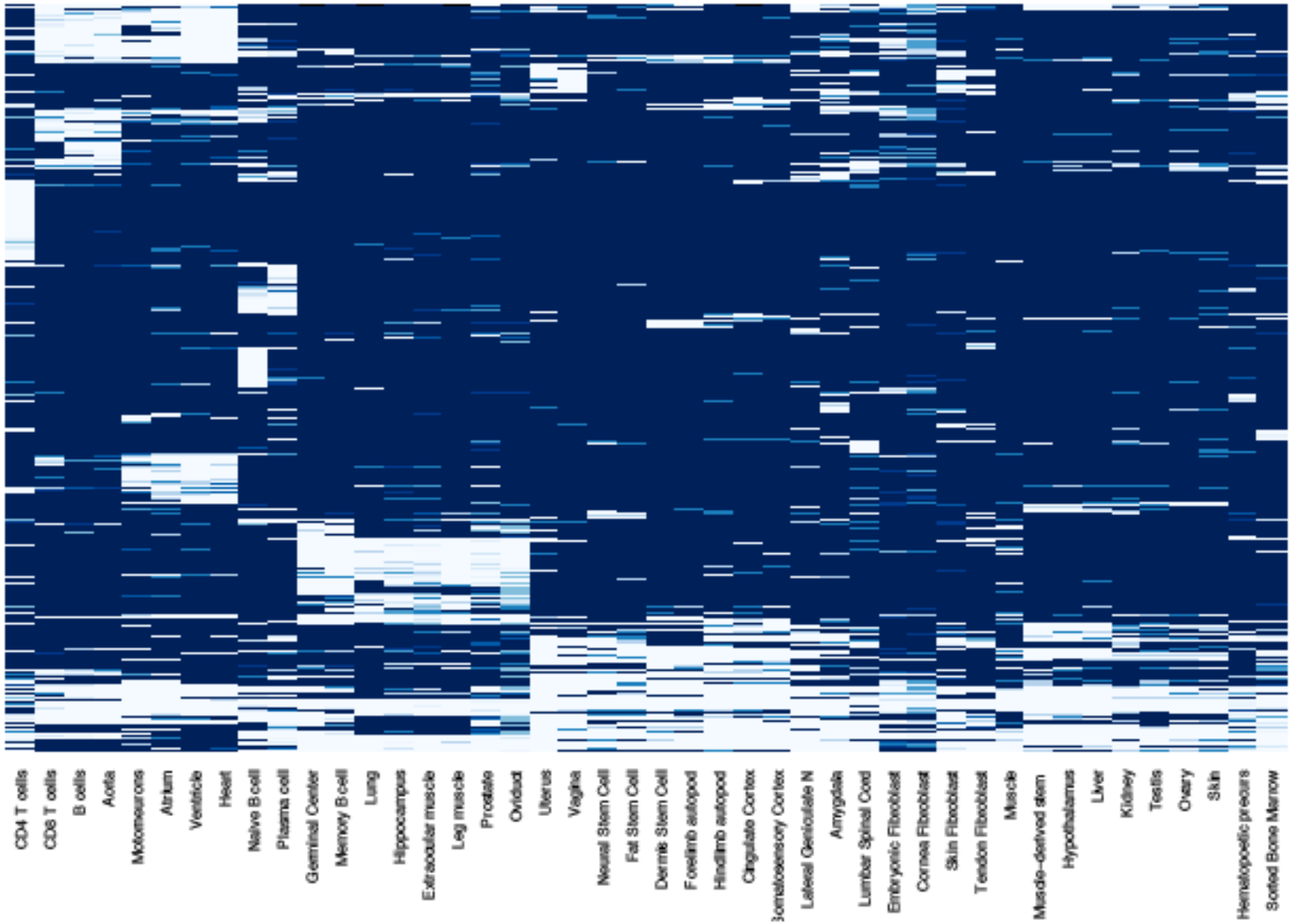
Pre-revolution Gene Expression Measurement



Post-revolution Gene Expression Measurements



Different tissues, different genes expressed

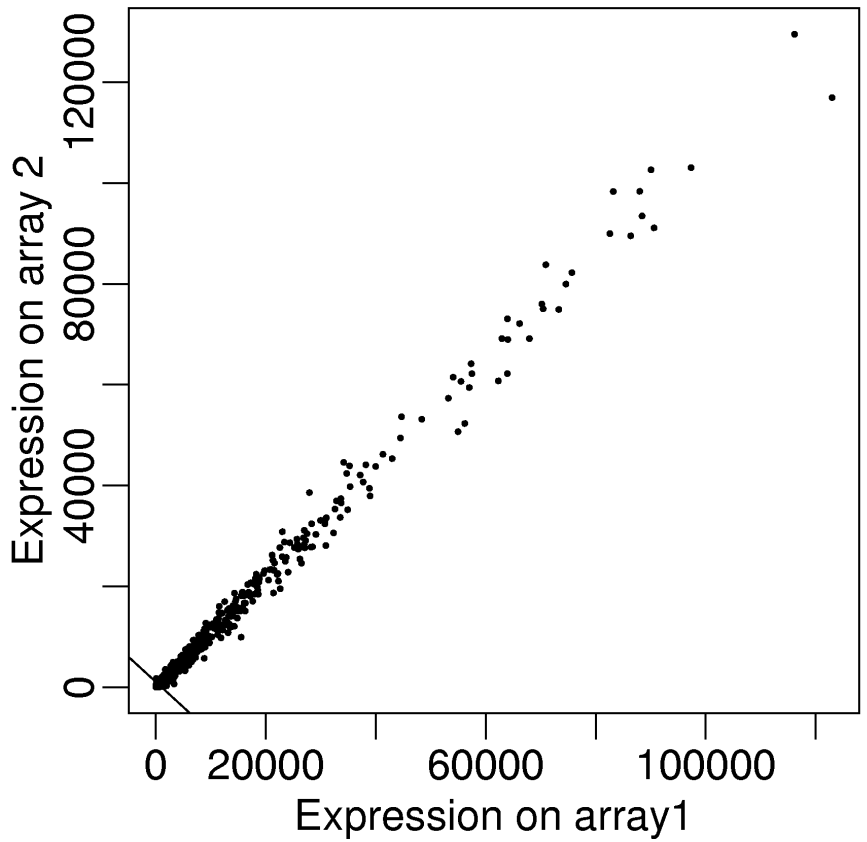


Differential Expression

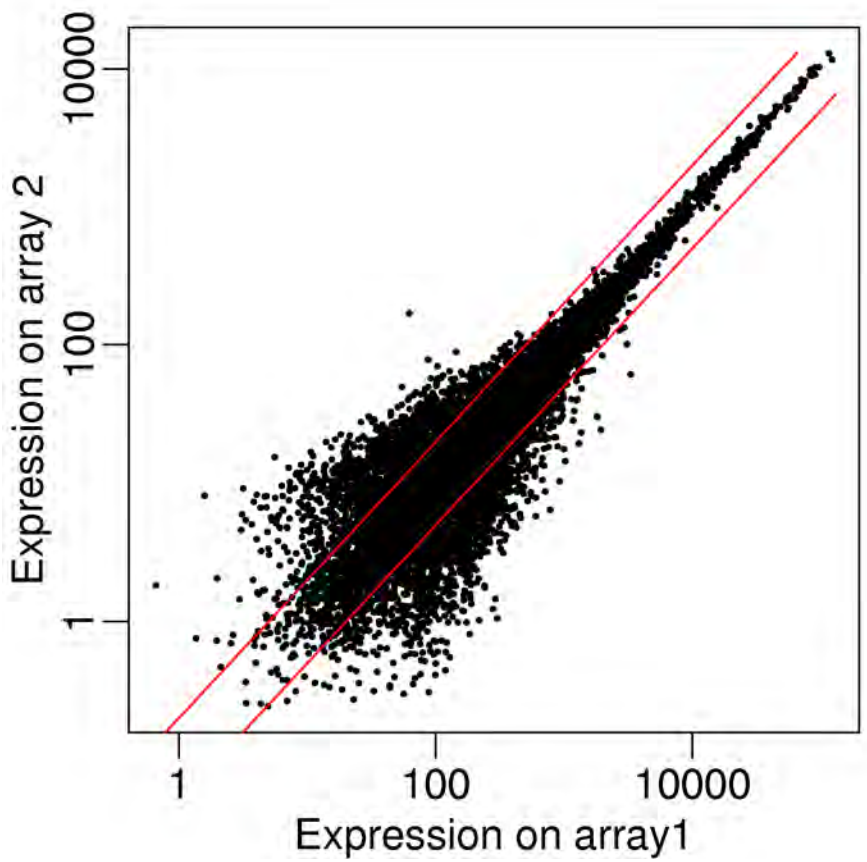
Data characteristics, useful plots, review of basic inference, and the use of empirical Bayes.

Data characteristics

Raw data from two arrays



Same data in log scale



More reasons to work in log-scale

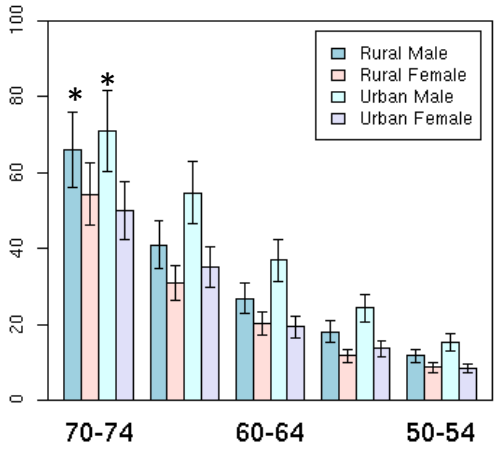
- For better of worst, fold changes are the preferred quantification of differential expression. Fold changes are basically ratios
- Biologist sometimes use the following weird notation: -2 means 1/2, -3 means 1/3, etc… Note there are no values between -1 and 1!
- Ratios are not symmetric around 1. This makes it problematic to perform statistical operations with ratios. We prefer logs

Quantifying differentially expression

Example

- Consider a case were we have observed two genes with fold changes of 2

- Is this worth reporting? Some journals require *statistical significance*. What does this mean?



Review of Statistical Inference

- Let *Y-X* be our measurement representing diferential expression for a given gene
- What is the typical null hypothesis?
- For simplicity let us assume *Y-X* follows a normal distribution
- *Y-X* may have a different distribution under the null hypothesis for different genes. What will be different?
- The standard deviation σ of *Y-X* may be different.
- We could consider the z-statistic *(Y-X) /* σ instead. What is the distribution of z-statistic? Can we compute z-statistic?
- We do not know σ!
- What is σ? Why is it not 0? How do we estimate σ?
- t-test

Sample Summaries

Observations: $X_1,\ldots,X_M \quad Y_1,\ldots,Y_N$

Averages: $\bar{X} = \frac{1}{M}\sum_{i=1}^{M} X_i \qquad \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$

SD^2 or variances:

$$s_X^2 = \frac{1}{M-1}\sum_{i=1}^{M}(X_i - \bar{X})^2 \quad s_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

The t-statistic

t - statistic:

$$\frac{\overline{Y} - \overline{X}}{\sqrt{\frac{s_Y^2}{N} + \frac{s_X^2}{M}}}$$

Properties of t-statistic

- If number of independent replicates is large what is t-statistic distribution
- Normally distributed with mean 0 and and SD of 1

- If sample size not large but observed data is normally distributed then what is t-statistic?
- t-distribution with degrees of freedom depending on sample size

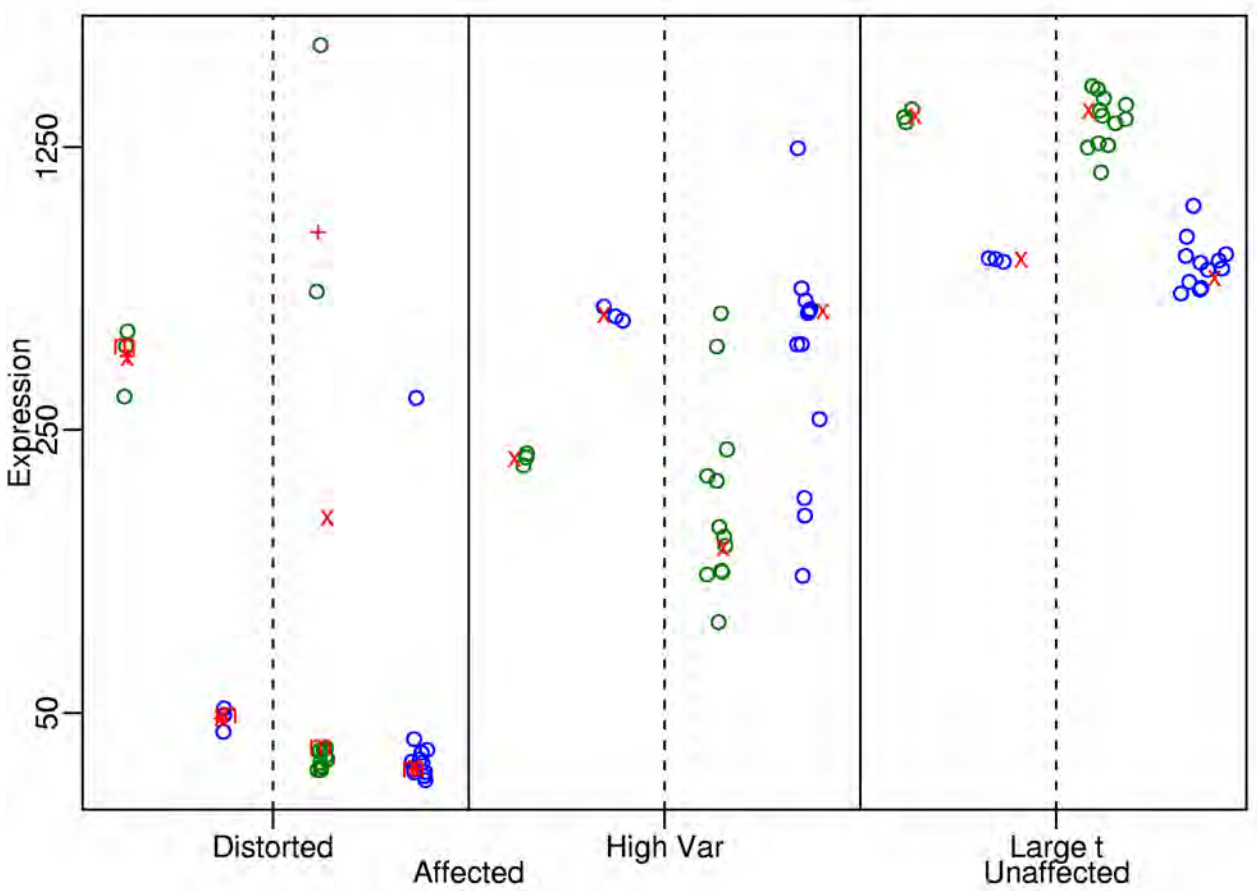
- We can then compute probability that t-statistic is as extreme or more when null hypothesis is true

- The tails of t-statistic are fat compared to z-statistic (when you know σ)
- Where does extra variability come from?

- Estimating the standard error.

- Should we bother estimating for each gene? How about assume all genes have same σ?

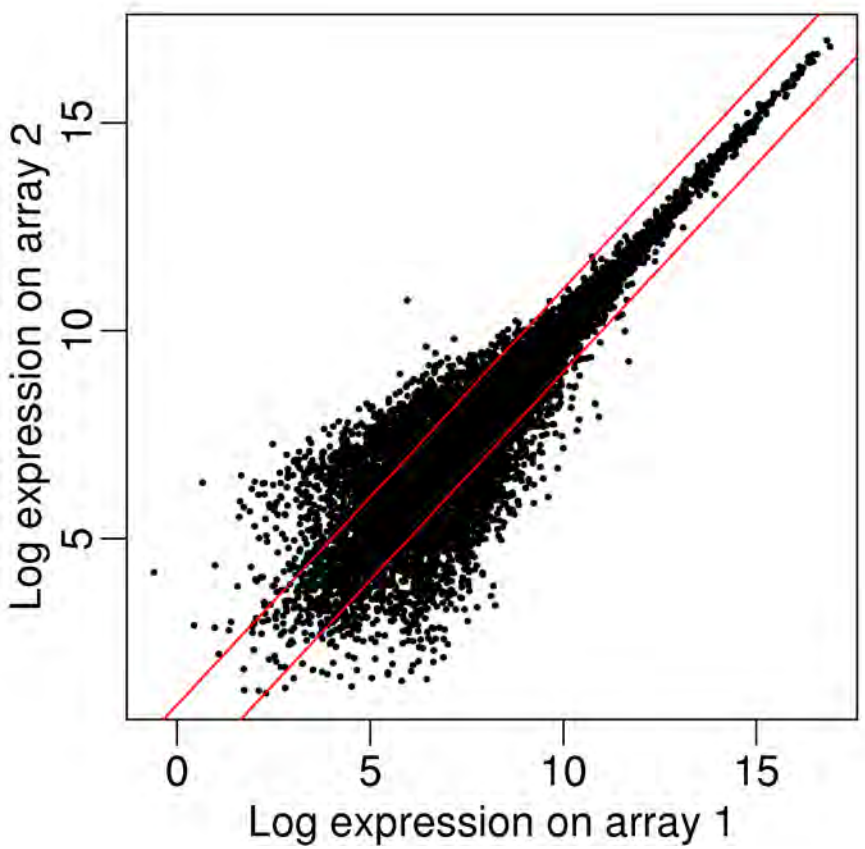
Biological versus technical replicates



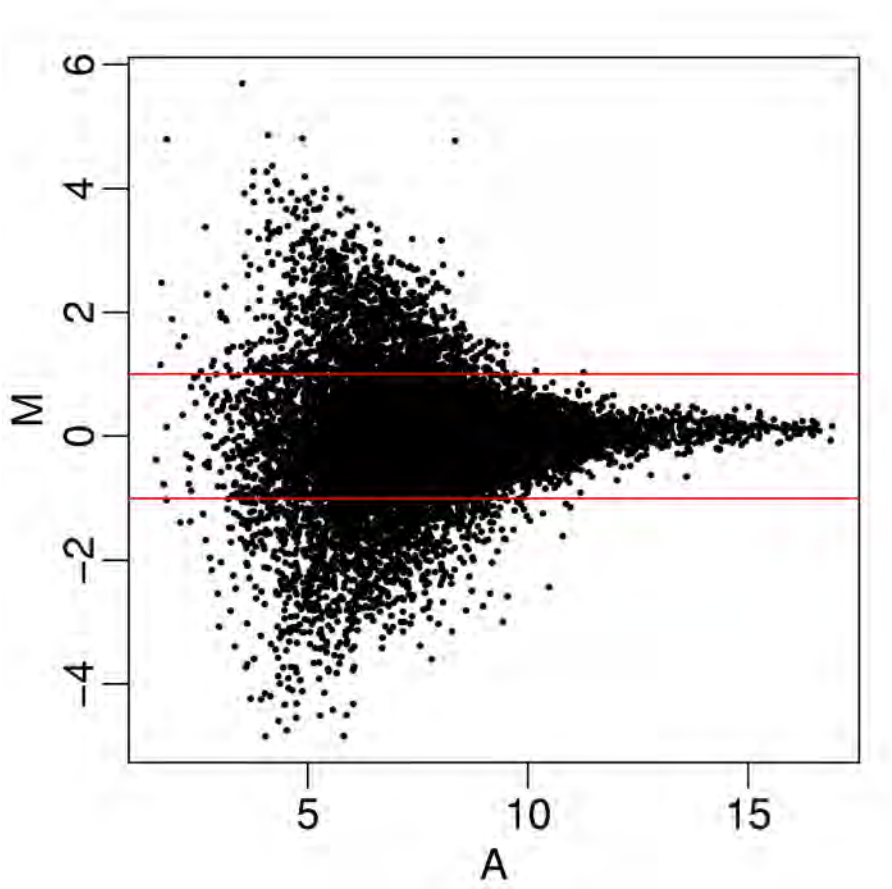
Note: different genes, different variance

Some data and useful plots

Scatter Plot



A 45° rotation highlights a problem

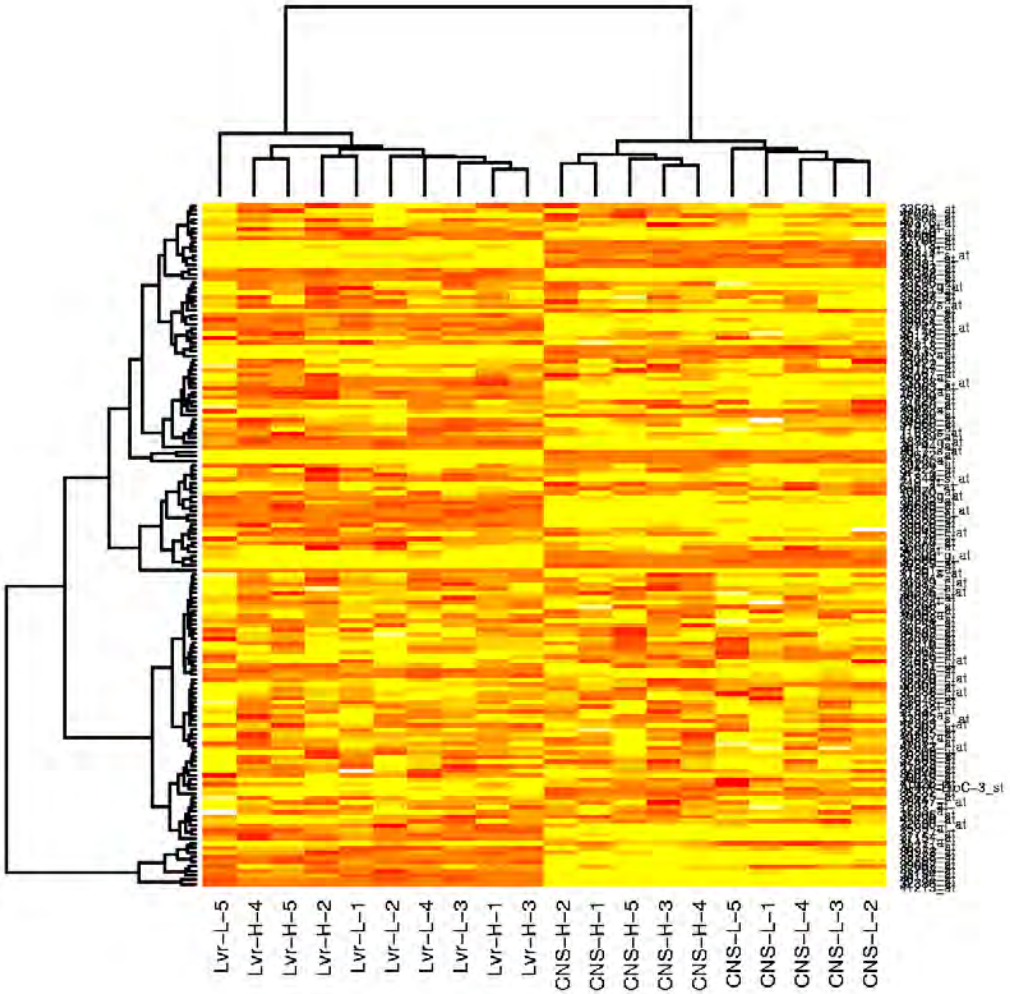


This is referred to as MAplot

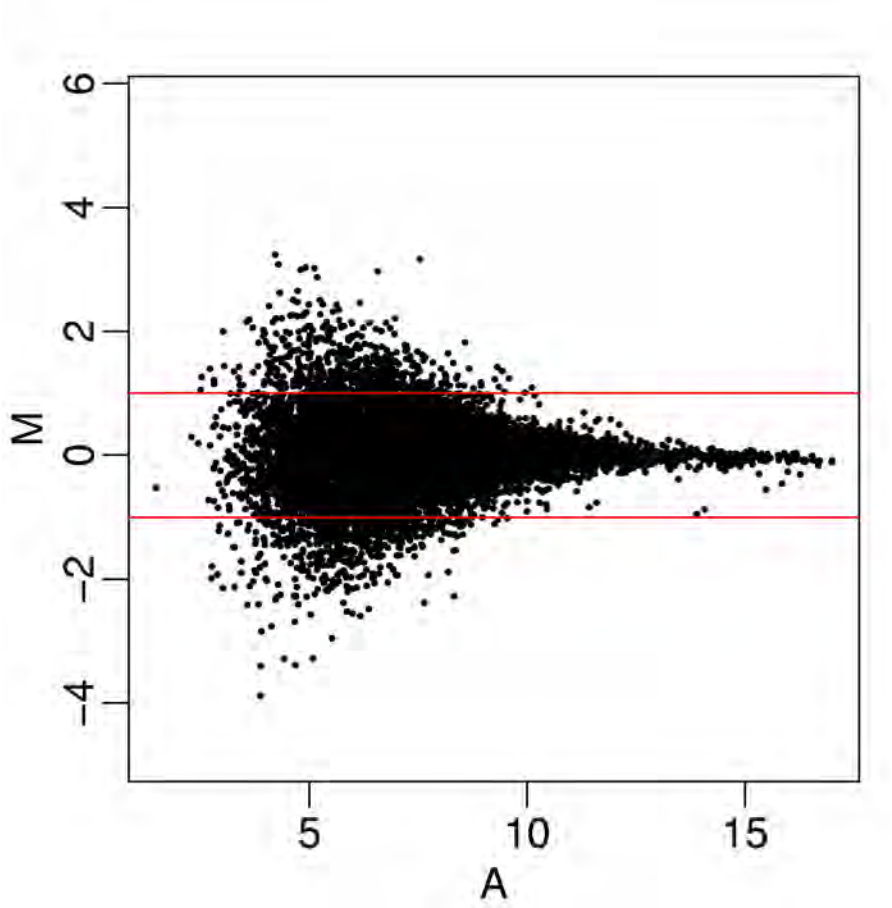
Experiments with replicates

- If we are interested in genes with over-all large fold changes why not look at average (log) fold changes?
- Experience has shown that one usually wants to stratify by over-all expression
- We can make averaged MA plots:
  - M = difference in average log intensities and
  - A = average of log intensities

Heatmap is common

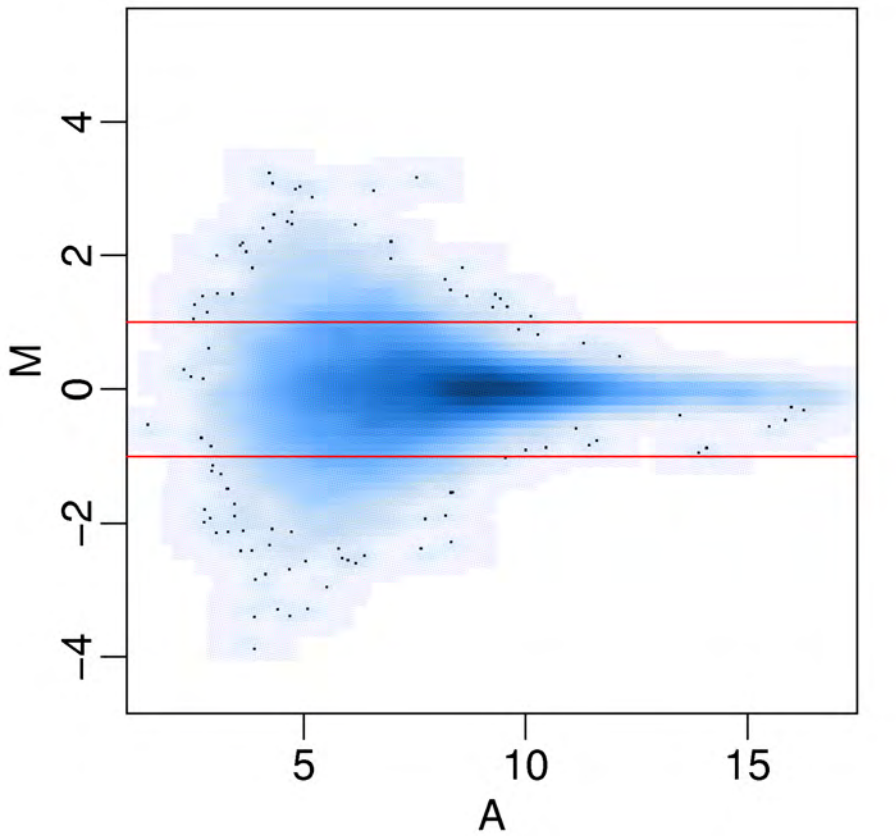


MA plot of average log ratios

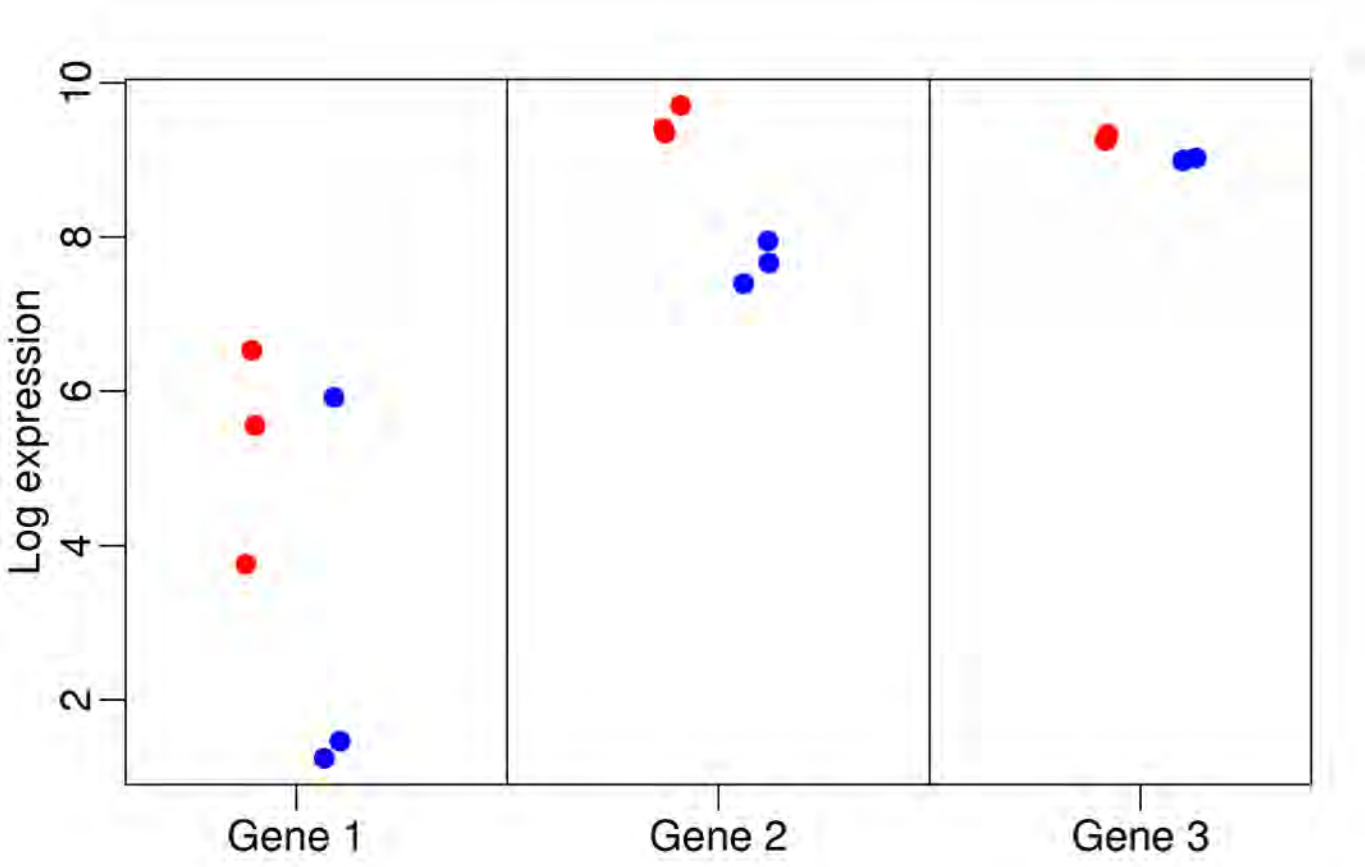


Note: variance reduced… law of large numbers

Scatter Smooth



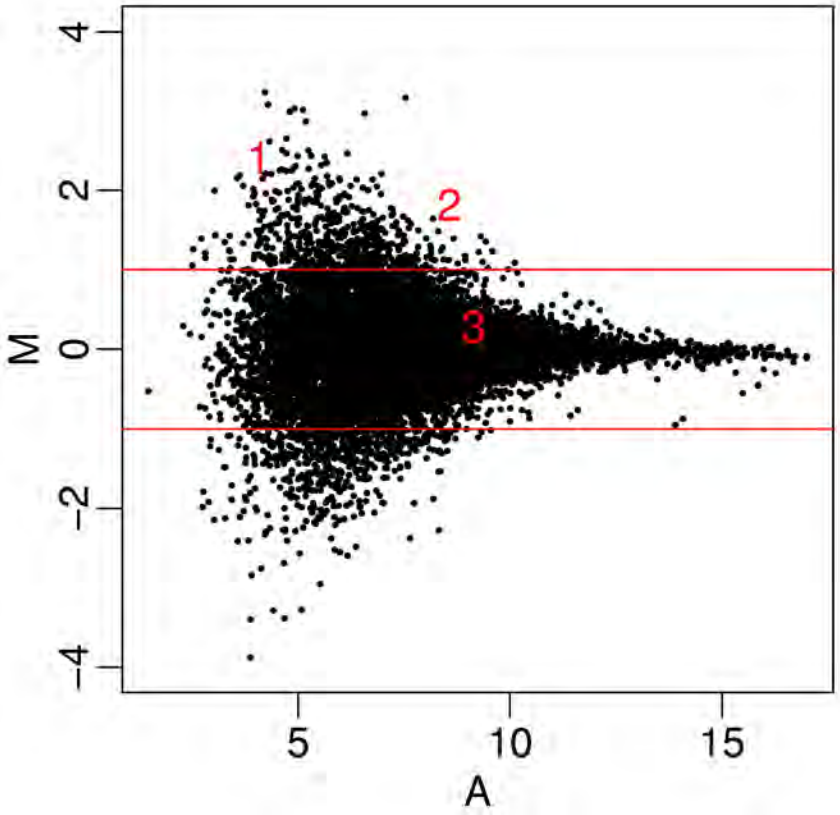
Should we consider gene-specific variance?

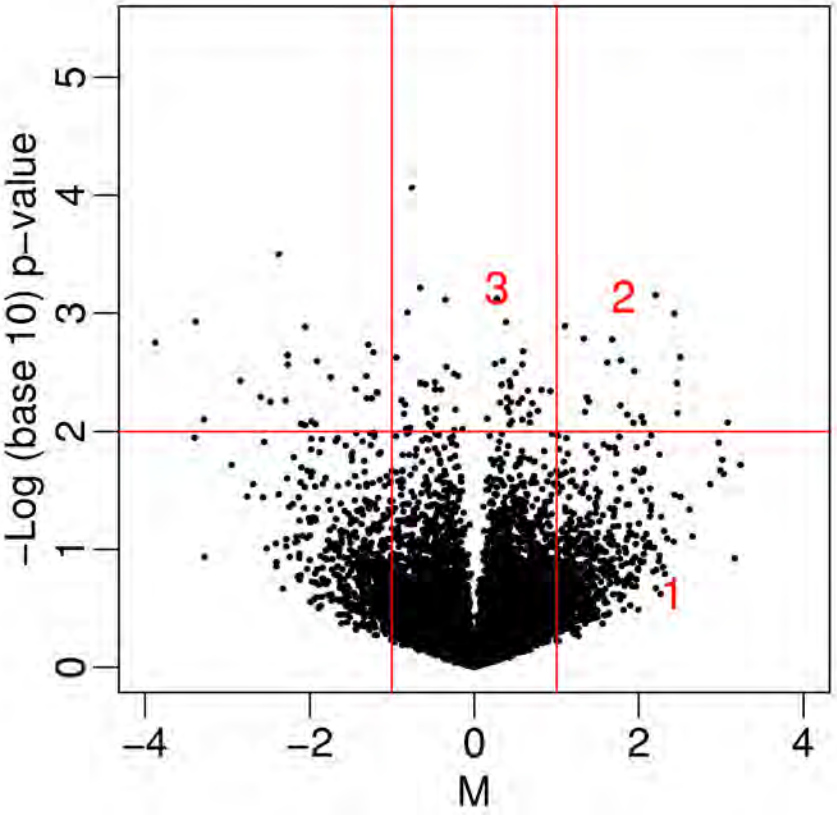


Another useful plot

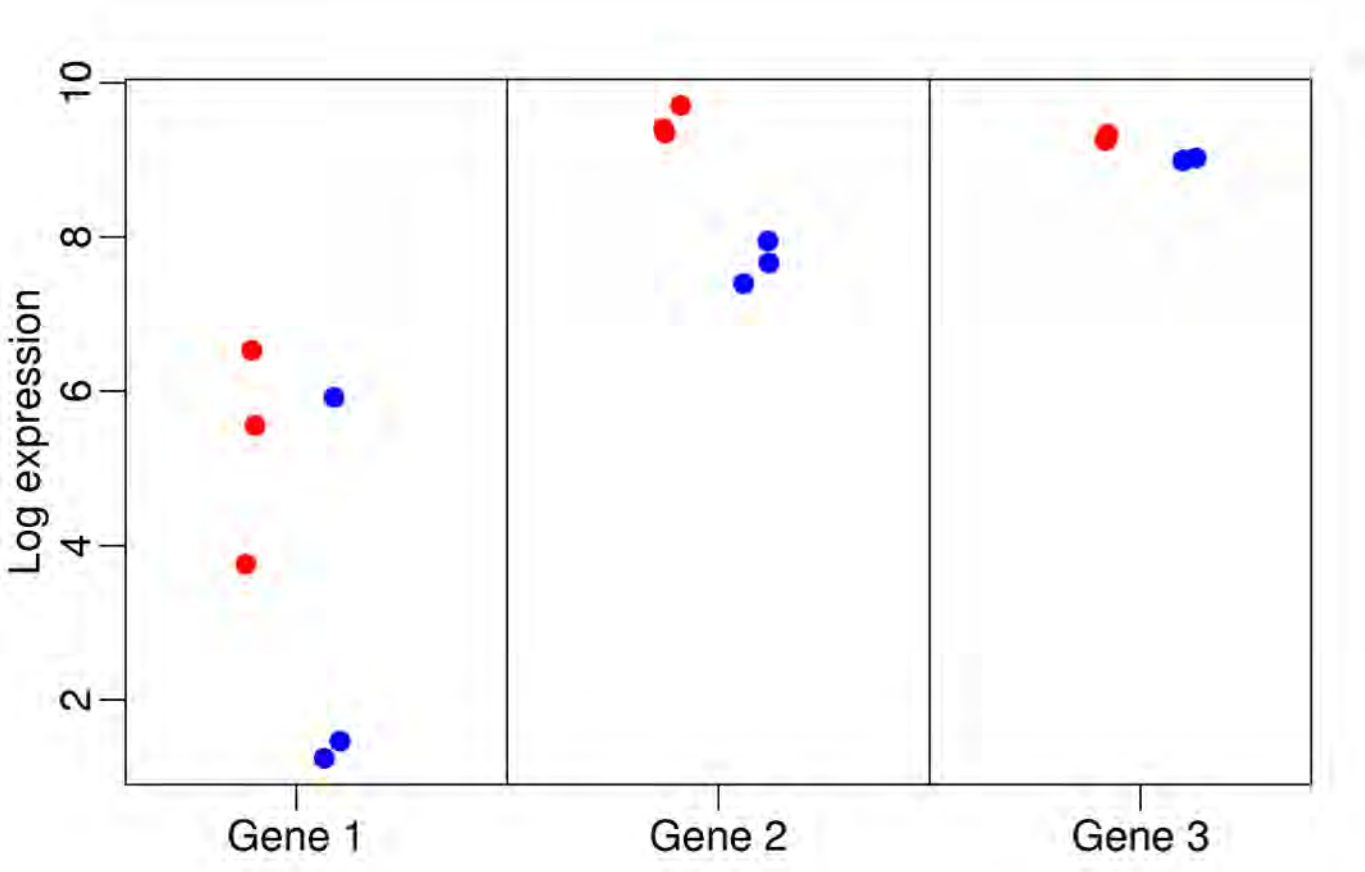
- The volcano plot shows, for a particular test, negative log p-value against the effect size (M)

MA and volcano





Remember these?



Borrowing strength

- Biological intuition tells us variance can't be to small or too big
- We estimate variance for thousands of genes, why not use this information.

- Ad-hoc (SAM), empirical Bayes (limma), and Stein estimators (Churchill's group) are examples of techniques that can help

Introduction to Empirical Bayes

BASIC TWO-STAGE SAMPLING

$$\theta \sim G$$
$$Y \mid \theta \sim f(y \mid \theta)$$

- G is the prior
- f is the sampling distribution
- Use the "rules of probability" to get the:

Posterior Distribution

$$g(\theta \mid Y) = \frac{f(y|\theta)g(\theta)}{f_G(Y)}$$

Marginal Distribution

$$f_G(Y) = \int f(y \mid u)g(u)du$$

THE BASIC GAUSSIAN/GAUSSIAN MODEL

$$\begin{aligned} \text{Prior:} \quad \mathsf{G} &= N(\mu,\, \tau^2) \\ \text{Sampling distn.:} \quad \mathsf{f} &= N(\theta,\, \sigma^2) \\ \text{Marginal distn.:} \quad \mathsf{f}_G &= N(\mu,\, \sigma^2 + \tau^2) \end{aligned}$$

Overdispersion

- If (μ, τ^2, σ^2) are known, the posterior is Gaussian:

$$\begin{aligned} E(\theta|Y) &= B\mu + (1-B)Y \\ &= \mu + (1-B)(Y-\mu) \\ V(\theta|Y) &= (1-B)\sigma^2 \\ B &= \frac{\sigma^2}{\sigma^2+\tau^2} \end{aligned}$$

- The Gaussian prior is conjugate
- Shrinkage and variance reduction
- Increasing σ^2 or decreasing τ^2 produces greater shrinkage

Modeling Relative Expression

Courtesy of Gordon Smyth

Hierarchical Model

Normal Model

$$\hat{\beta}_{gj} \sim N(\beta_{gj}, c_{gj}\sigma_g^2)$$

$$s_g^2 \sim \sigma_g^2 \chi_{d_g}^2$$

Prior

$$P(\beta_{gj} \neq 0) = p$$

$$\beta_{gj} \mid \beta_{gj} \neq 0 \sim N(0, c_{0j}\sigma_g^2)$$

$$\sigma_g^2 \sim s_0^2 \left(\chi_{d_0}^2 / d_0\right)^{-1}$$

Reparametrization of Lönnstedt and Speed 2002

Normality, independence assumptions are wrong but convenient, resulting methods are useful

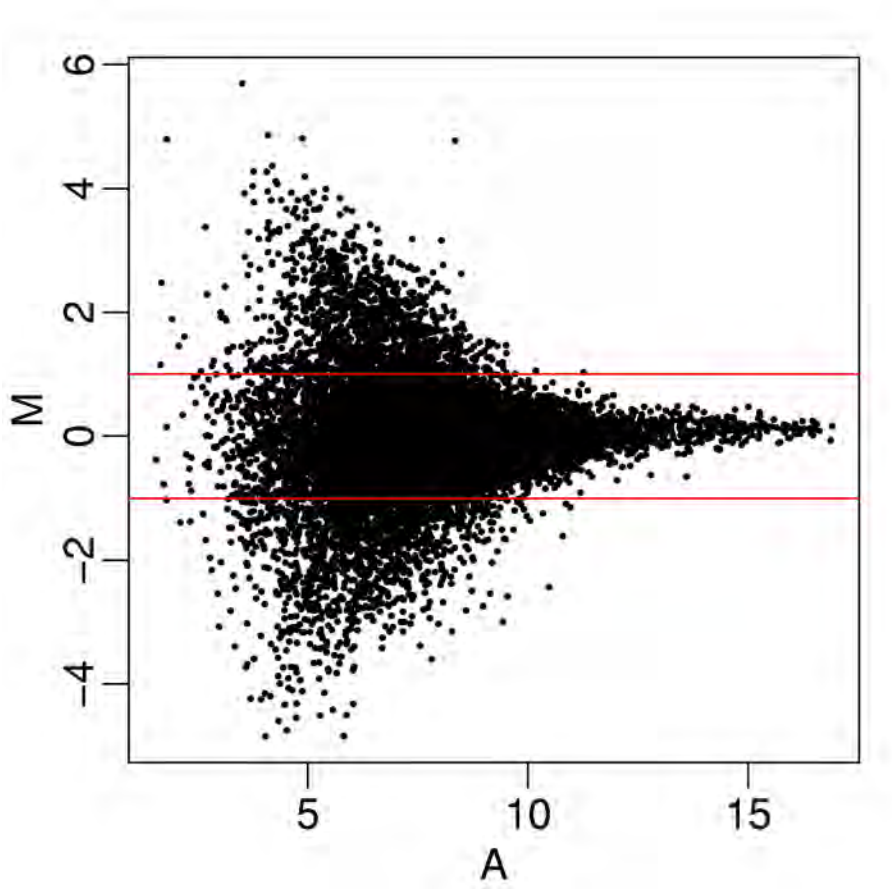
One final problem

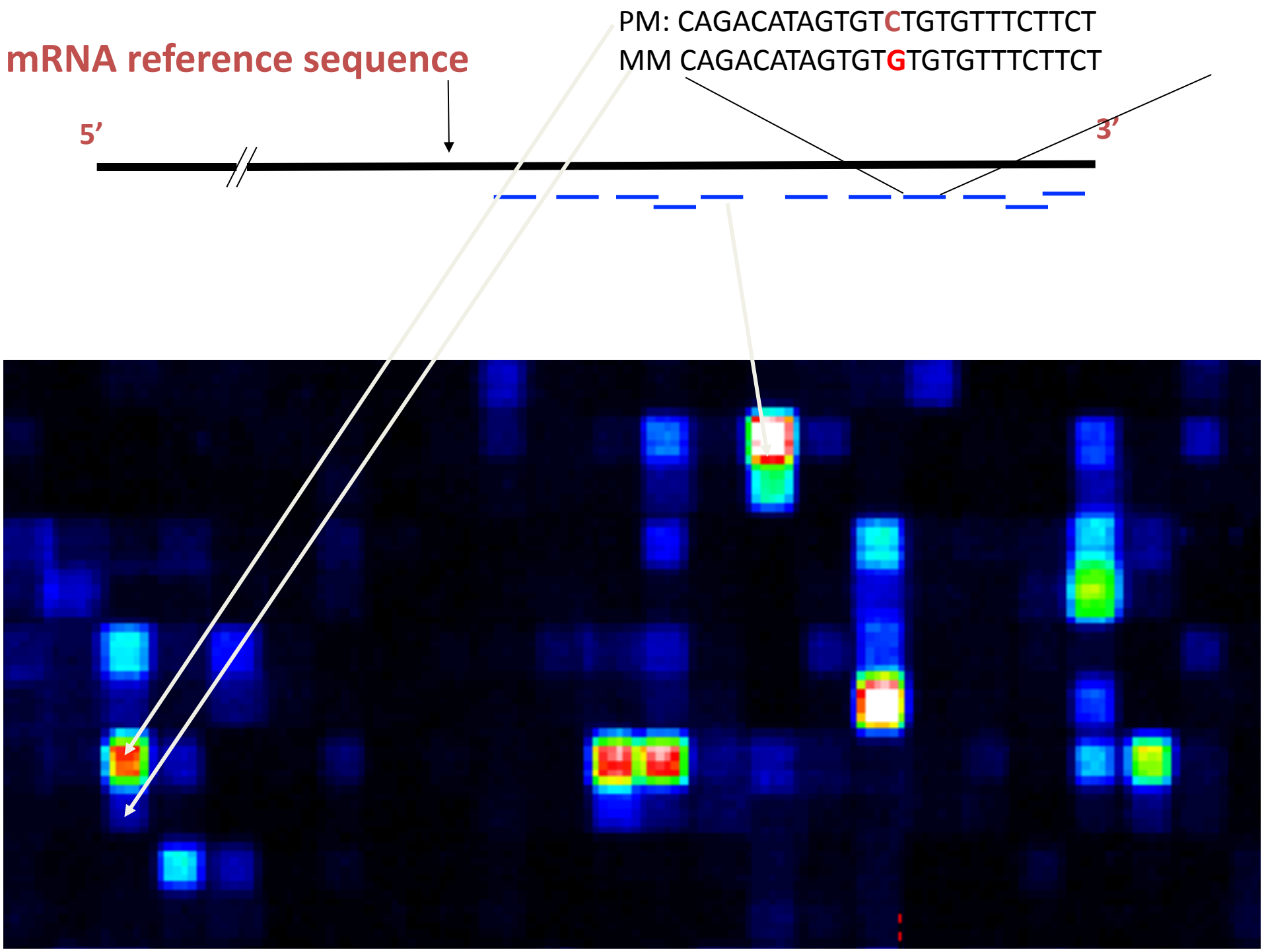
- If we are independently testing 20,000 genes…
- If no gene is differentially expressed, how many will attain p-values smaller than 0.01?
- 200 genes! p-value no longer have same interpretation
- FDR and methods inspired by it, such as q-values, more useful
- See papers by 1) Benjamini 2) Storey and 3) Dudoit

Pre-processing

Background correction and normalization

Why so much noise?




mRNA reference sequence
PM: CAGACATAGTGTCTGTGTTTCTTCT
MM CAGACATAGTGTGTGTGTTTCTTCT
5’
3’

Affymetrix Spike In Experiment

Spike-in Experiment

- Throughout we will be using Data from Affymetrix's spike-in experiment
- Replicate RNA was hybridized to various arrays
- Some probesets were spiked in at different concentrations across the different arrays
- This gives us a way to assess precision and accuracy
- Done for HGU95 and HGU133 chips
- Available from Bioconductor experimental data package: *SpikeIn*

Spikein Experiment (HG-U95)

Probeset

Array	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	0	0.25	0.5	1	2	4	8	16	32	64	128	0	512	1024	256	32
B	0.25	0.5	1	2	4	8	16	32	64	128	256	0.25	1024	0	512	64
C	0.5	1	2	4	8	16	32	64	128	256	512	0.5	0	0.25	1024	128
D	1	2	4	8	16	32	64	128	256	512	1024	1	0.25	0.5	0	256
E	2	4	8	16	32	64	128	256	512	1024	0	2	0.5	1	0.25	512
F	4	8	16	32	64	128	256	512	1024	0	0.25	4	1	2	0.5	1024
G	8	16	32	64	128	256	512	1024	0	0.25	0.5	8	2	4	1	0
H	16	32	64	128	256	512	1024	0	0.25	0.5	1	16	4	8	2	0.25
I	32	64	128	256	512	1024	0	0.25	0.5	1	2	32	8	16	4	0.5
J	64	128	256	512	1024	0	0.25	0.5	1	2	4	64	16	32	8	1
K	128	256	512	1024	0	0.25	0.5	1	2	4	8	128	32	64	16	2
L	256	512	1024	0	0.25	0.5	1	2	4	8	16	256	64	128	32	4
M	512	1024	0	0.25	0.5	1	2	4	8	16	32	512	128	256	64	8
N	512	1024	0	0.25	0.5	1	2	4	8	16	32	512	128	256	64	8
O	512	1024	0	0.25	0.5	1	2	4	8	16	32	512	128	256	64	8
P	512	1024	0	0.25	0.5	1	2	4	8	16	32	512	128	256	64	8
Q	1024	0	0.25	0.5	1	2	4	8	16	32	64	1024	256	512	128	16
R	1024	0	0.25	0.5	1	2	4	8	16	32	64	1024	256	512	128	16
S	1024	0	0.25	0.5	1	2	4	8	16	32	64	1024	256	512	128	16
T	1024	0	0.25	0.5	1	2	4	8	16	32	64	1024	256	512	128	16

Spikein Experiment (HG-U133)

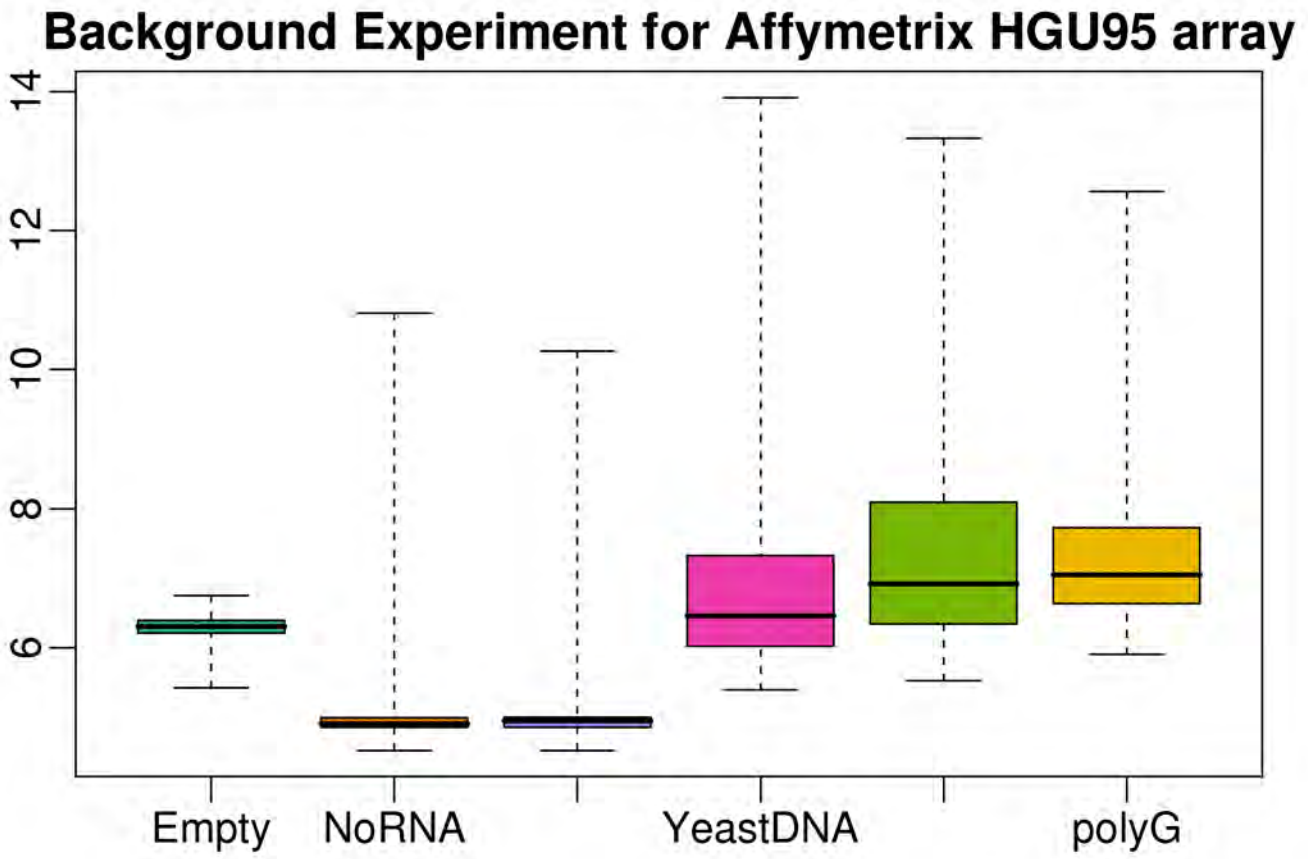
- A similar experiment was repeated for a newer chip
- The 1024 picoMolar concentration was not used. 1/8 was used instead.
- No groups of 12
- Note: More spike-ins to come!

Background Effects Experiments

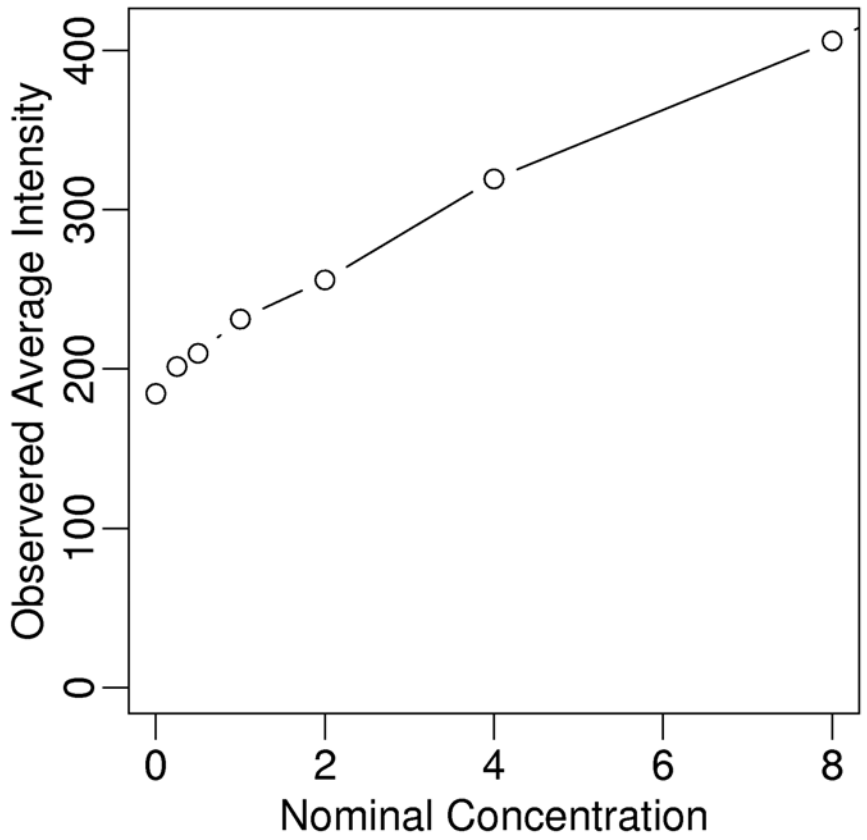
Learn about optical effect and NSB

	label	sample type
Empty	0	empty
NoRNA	1	no RNA
NoLabel	0	human
YeastDNA	1	yeast genomic DNA
polyC	1	poly C
polyG	1	poly G

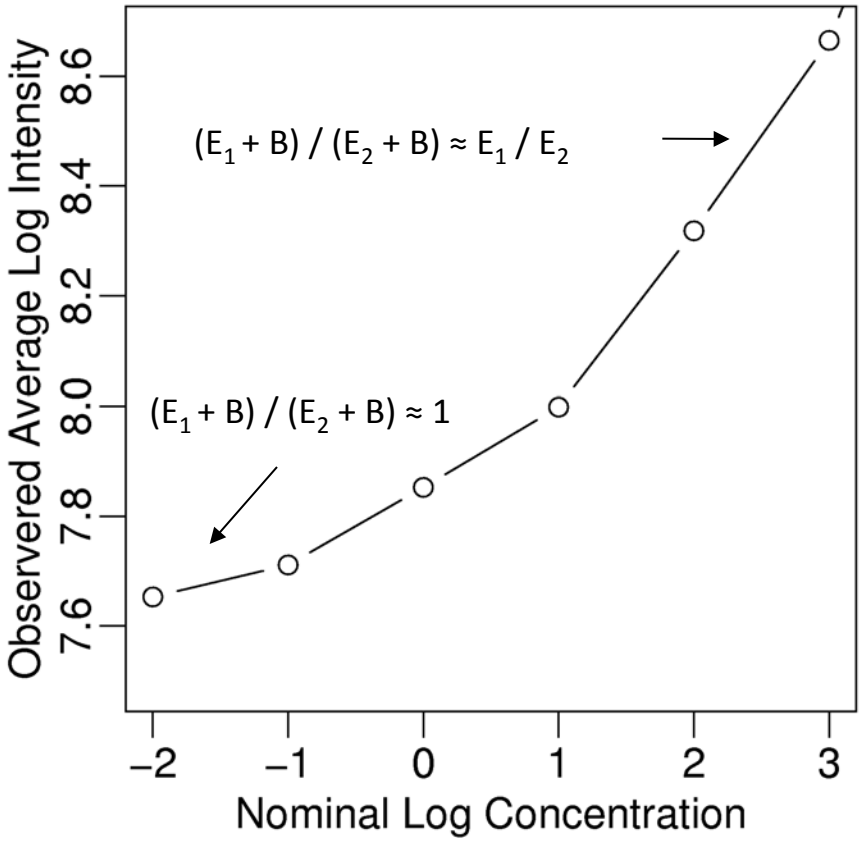
Background Effect



Why Adjust for Background?

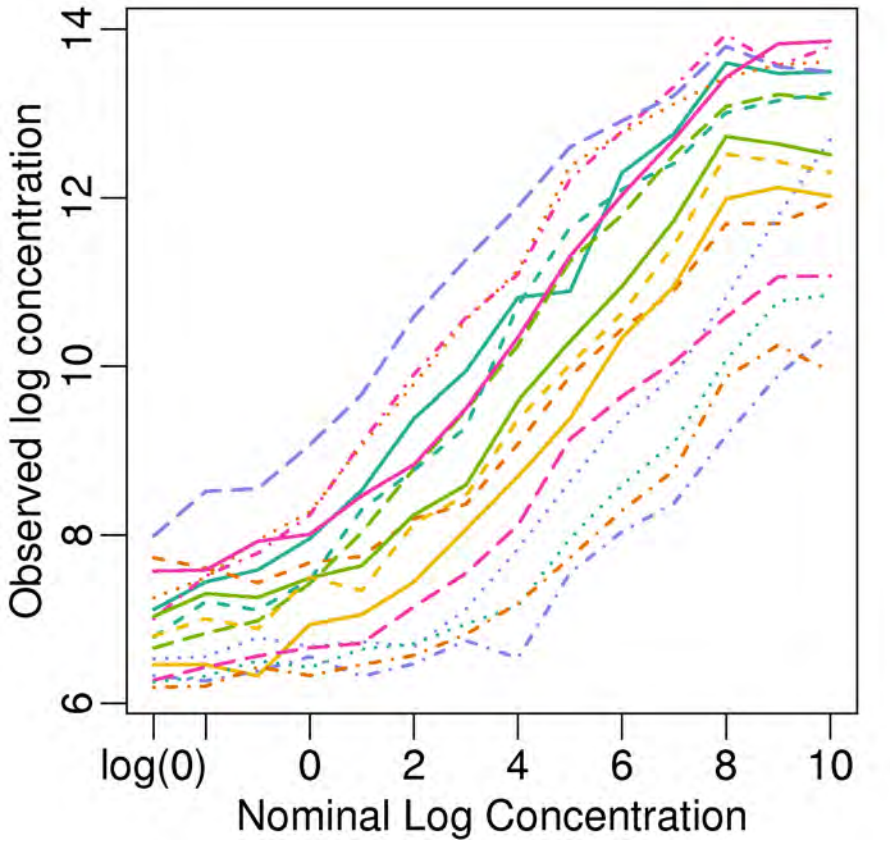


Why Adjust for Background?



Notice local slope decrease as the nominal concentration becomes small

Probe-specific NSB



Solutions

The fundamental problem

We want: $\frac{e_1}{e_2}$ But see: $\frac{b_1 + e_1}{b_2 + e_2}$

The company used $\frac{b_1 + e_1 - B_1}{b_2 + e_2 - B_2}$

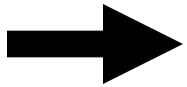
Which we can show has large variance when *e is* small

Direct Measurement Strategy

The hope is that:

$PM = B + S$

$MM = B$

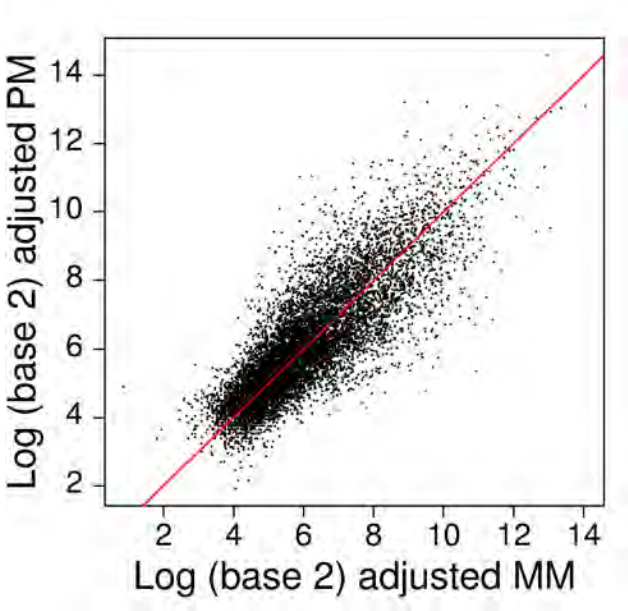


$PM - MM = S$

But this is not correct!

Notice

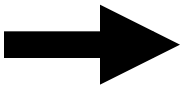
- We care about ratios
- We usually take log of S



Stochastic Model

Better to assume:

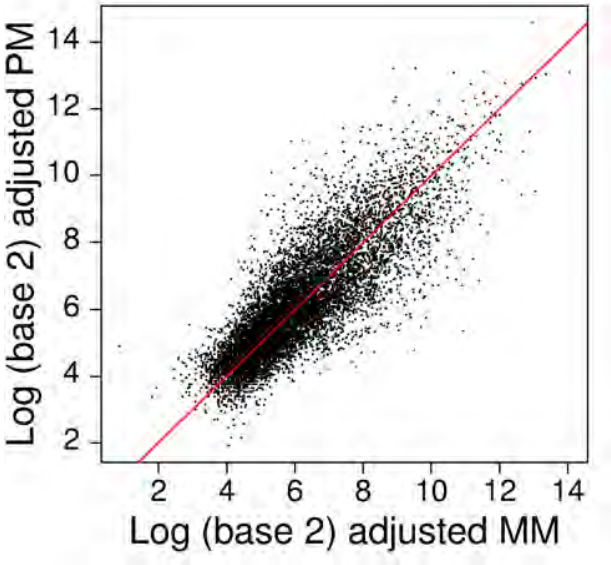
$PM = B_{PM} + S$
$MM = B_{MM}$
$Cor[\log(B_{PM}), \log(B_{MM})] = 0.7$



$Var[\log(PM - MM^{*})] \sim 1/S^{2}$

Alternative solution:

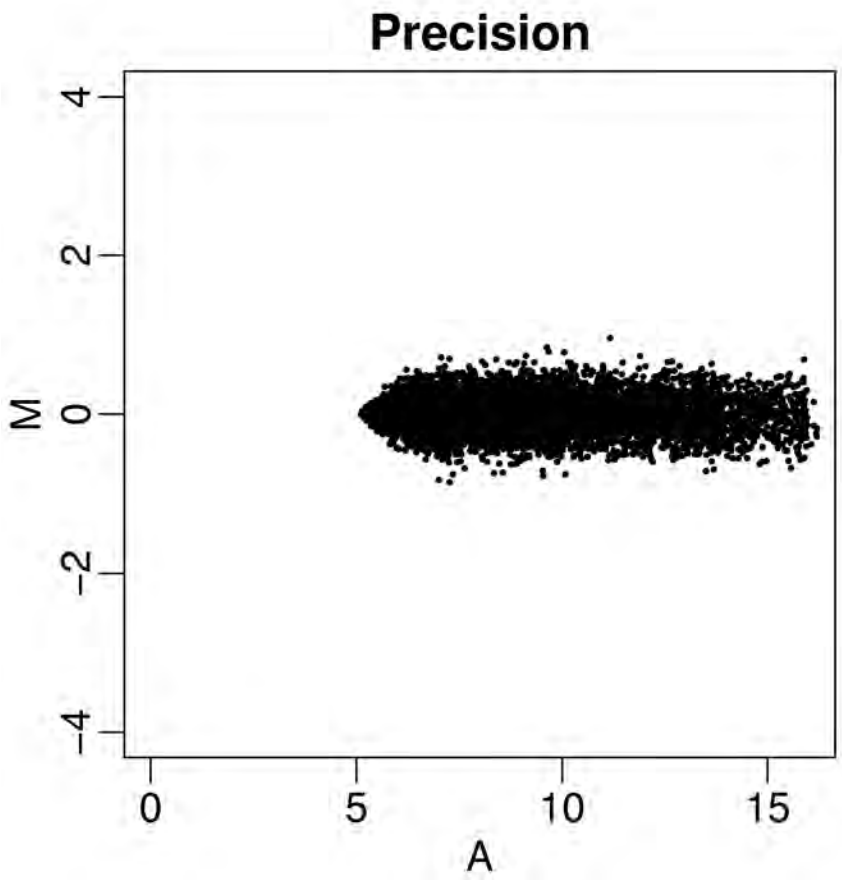
$E[\, S \mid PM \,]$

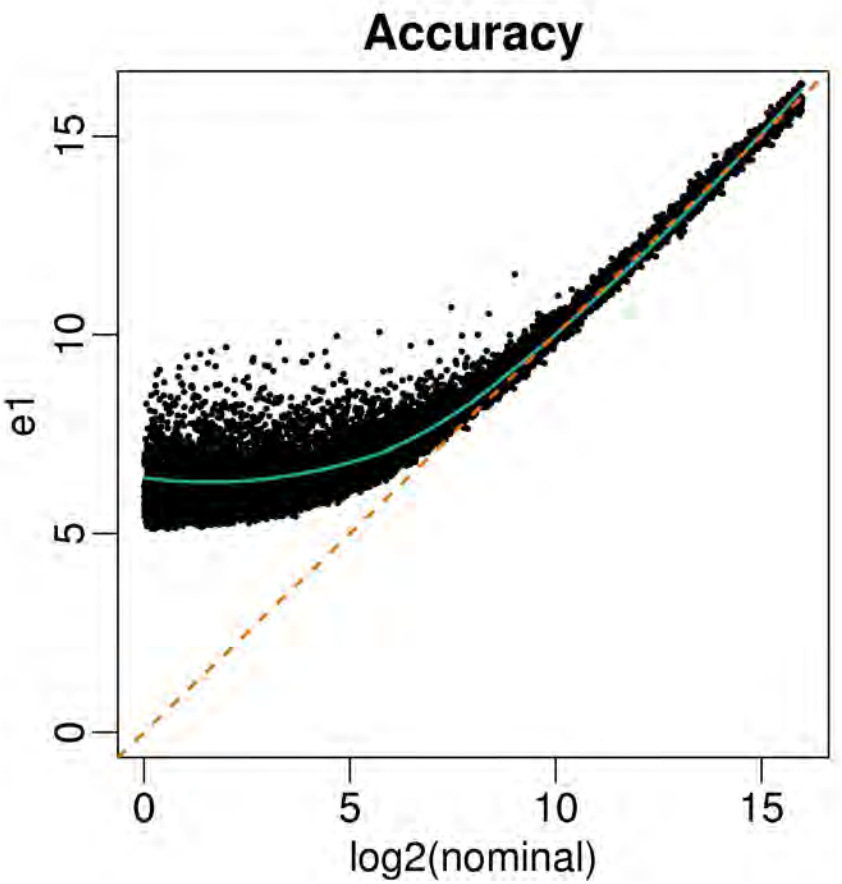


Simulation

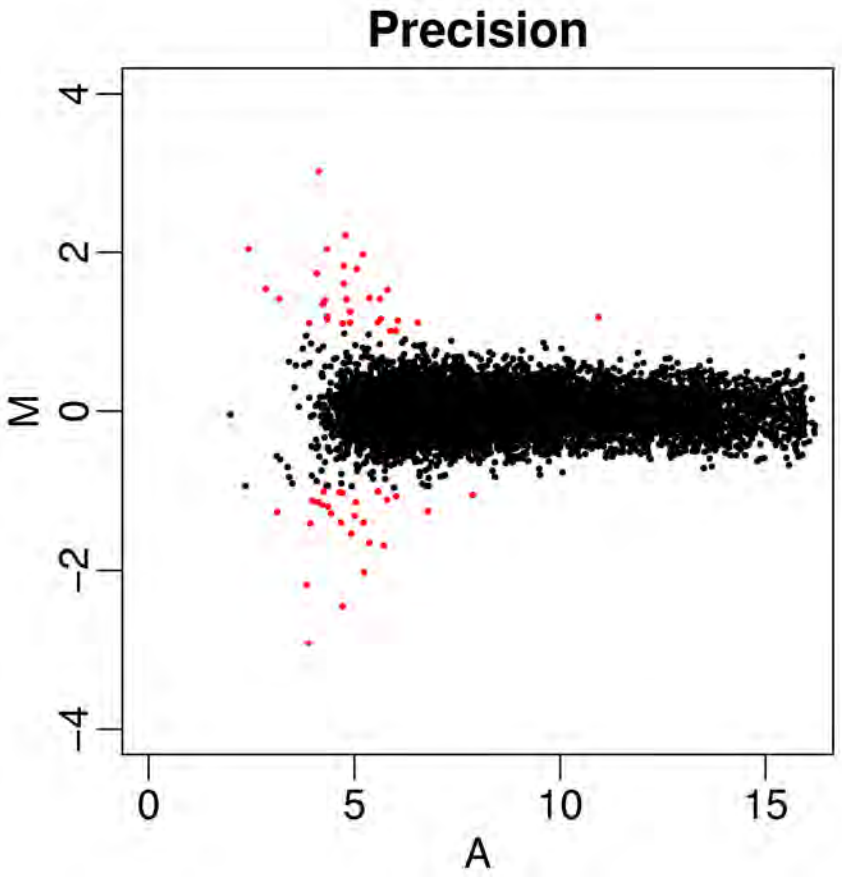
- We create some feature level data for two replicate arrays
- Then compute *Y=log(PM-kMM)* for each array
- We make an MA using the *Y*s for each array
- We make a observed concentration versue known concentration plot
- We do this for various values of k. The following “movie” shows *k* moving from 0 to 1.

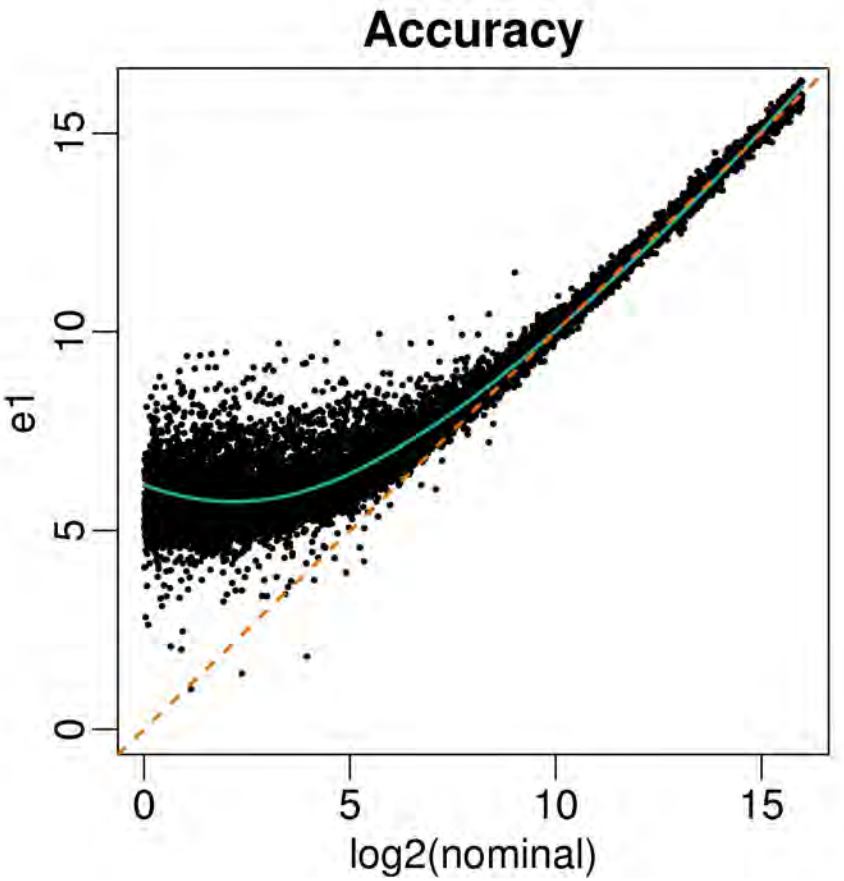
k=0


Precision
M
A
4
2
0
−2
−4
0
5
10
15

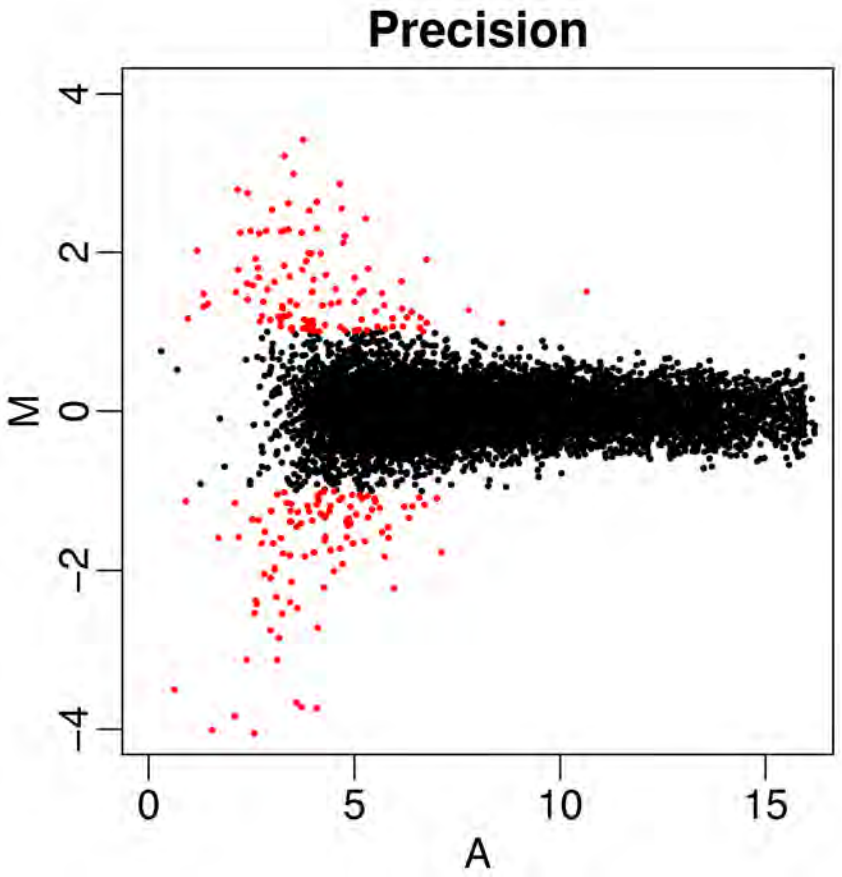

Accuracy
e1
log2(nominal)
15
10
5
0
0
5
10
15

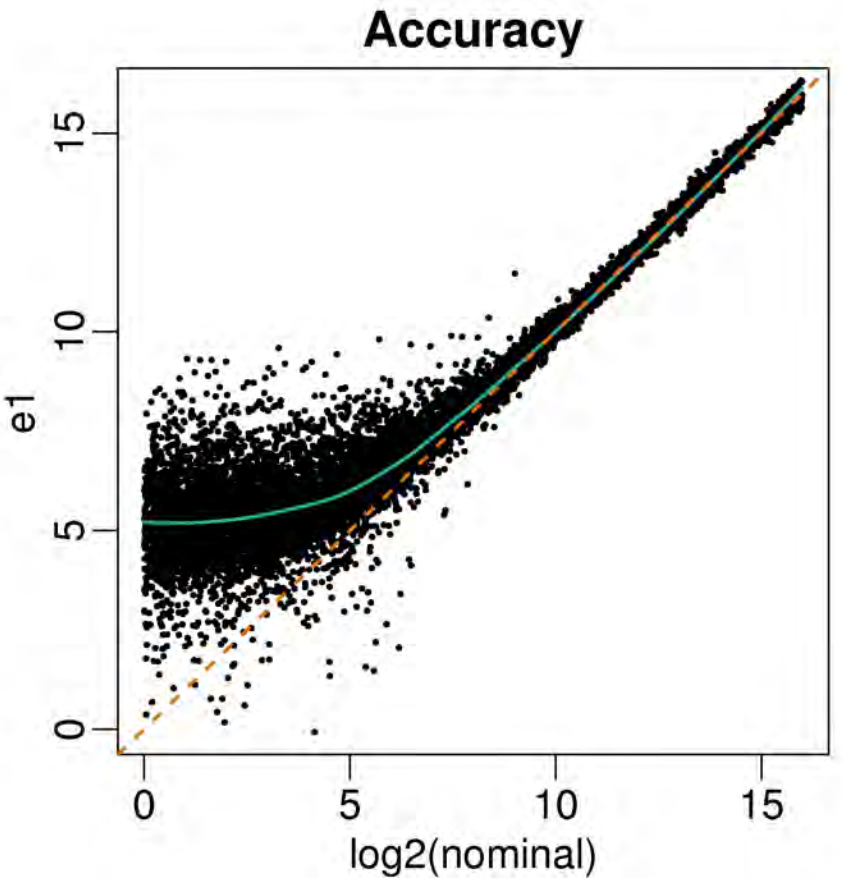
k=1/4



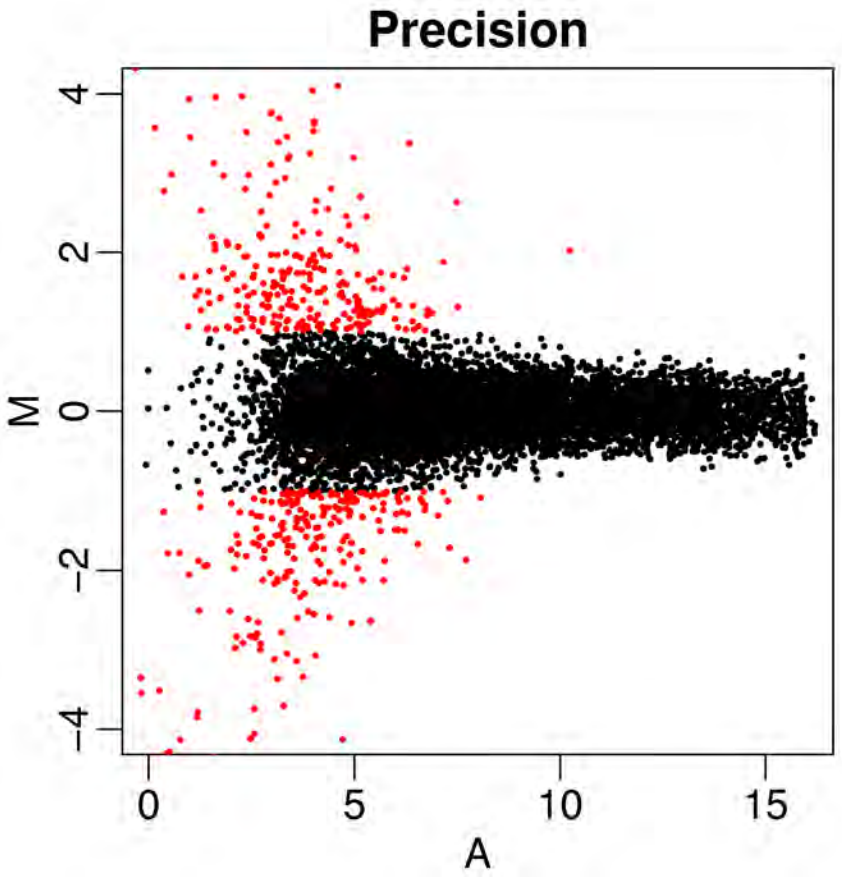


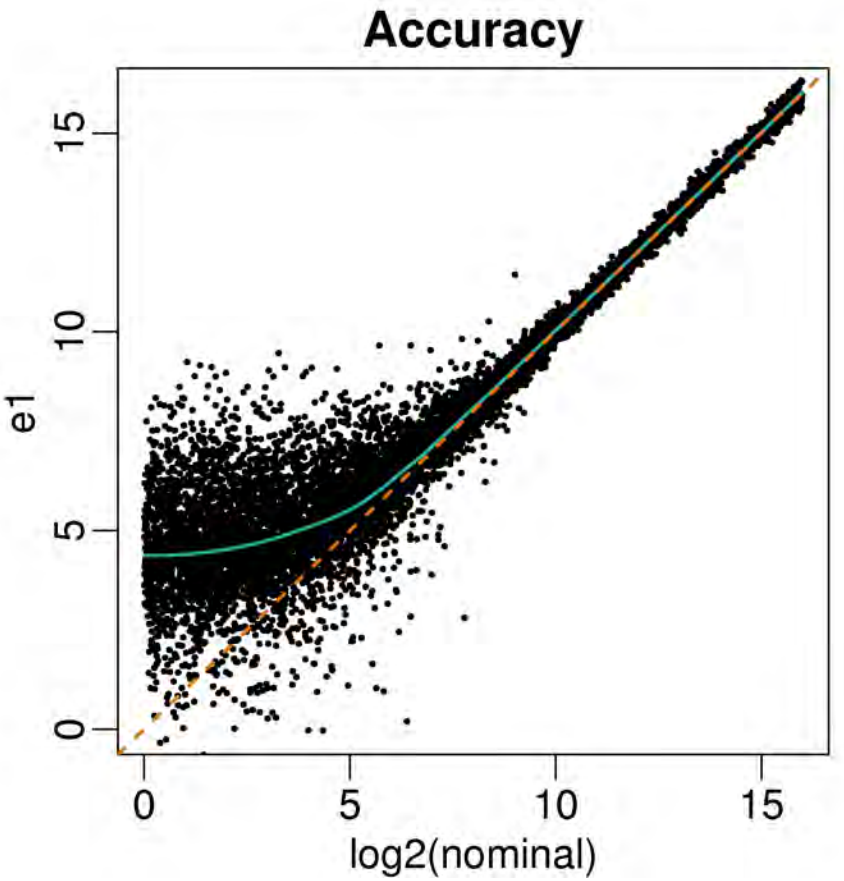
k=1/2


Precision
M
A
4
2
0
−2
−4
0
5
10
15

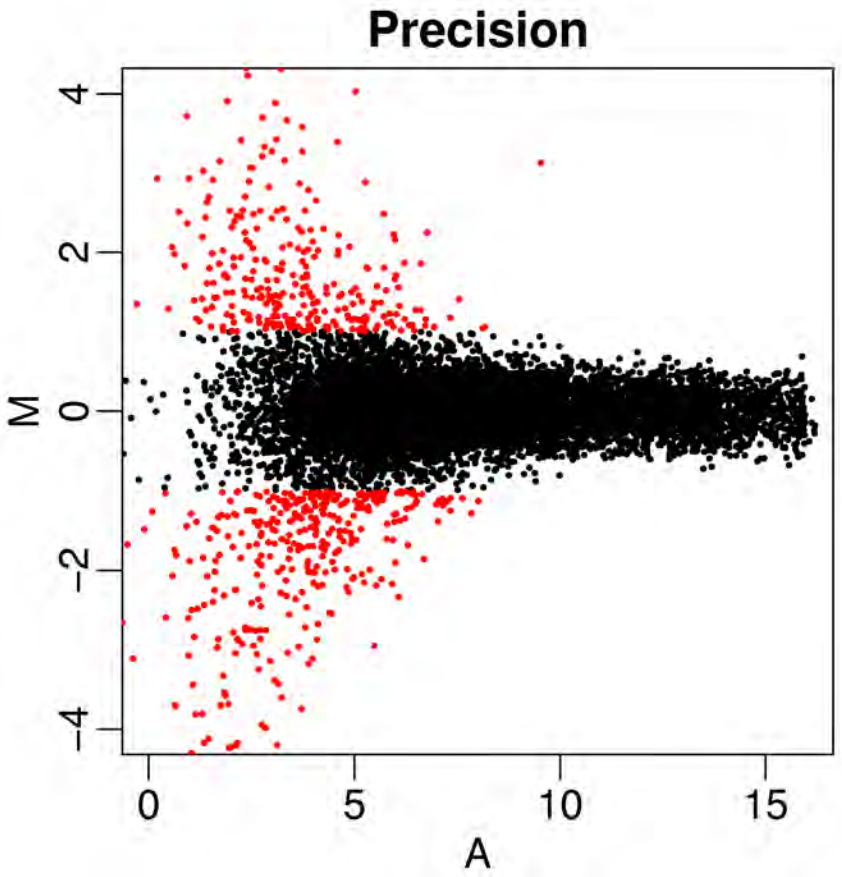

Accuracy
e1
log2(nominal)
0
5
10
15
0
5
10
15

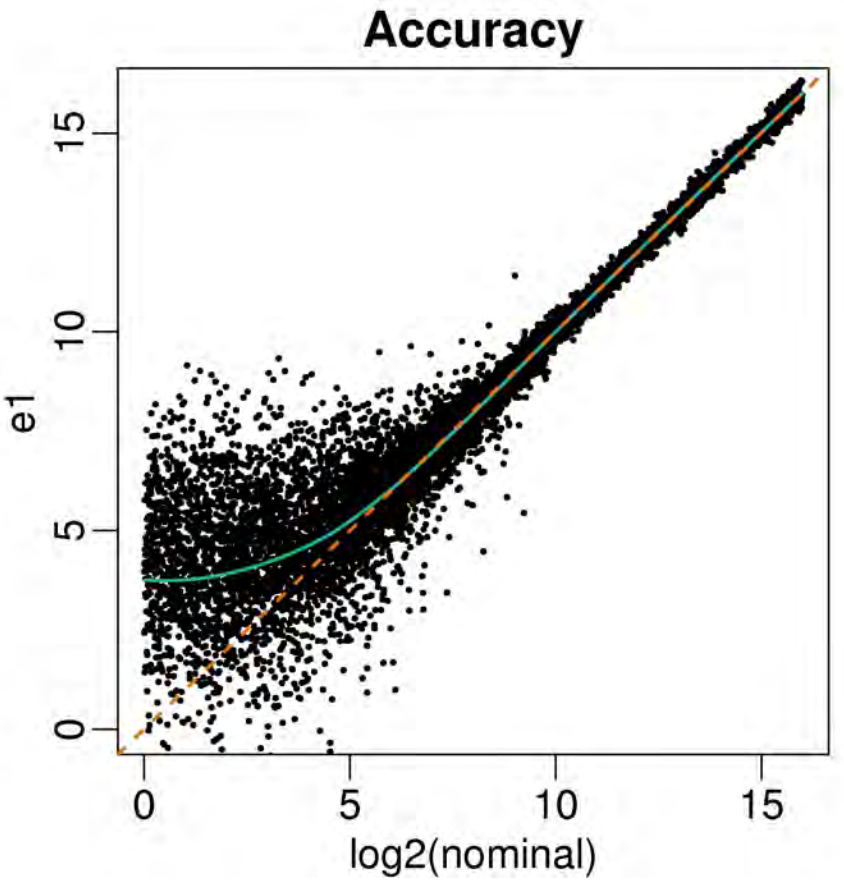
k=3/4


Precision
M
A
4
2
0
−2
−4
0
5
10
15

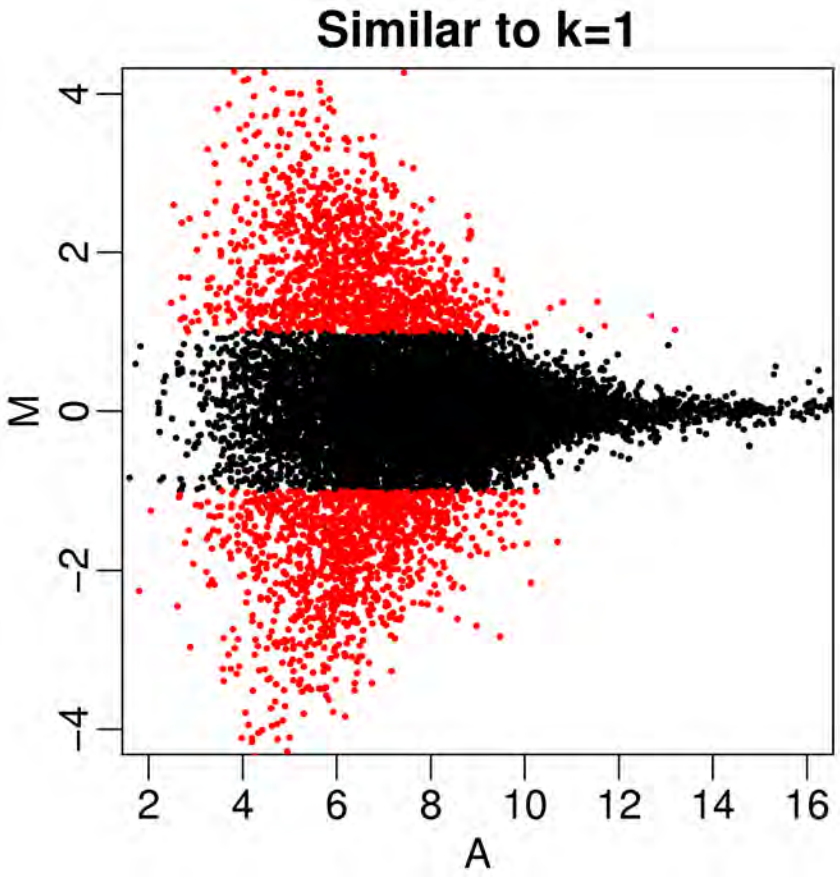

Accuracy
e1
log2(nominal)
0
5
10
15
0
5
10
15

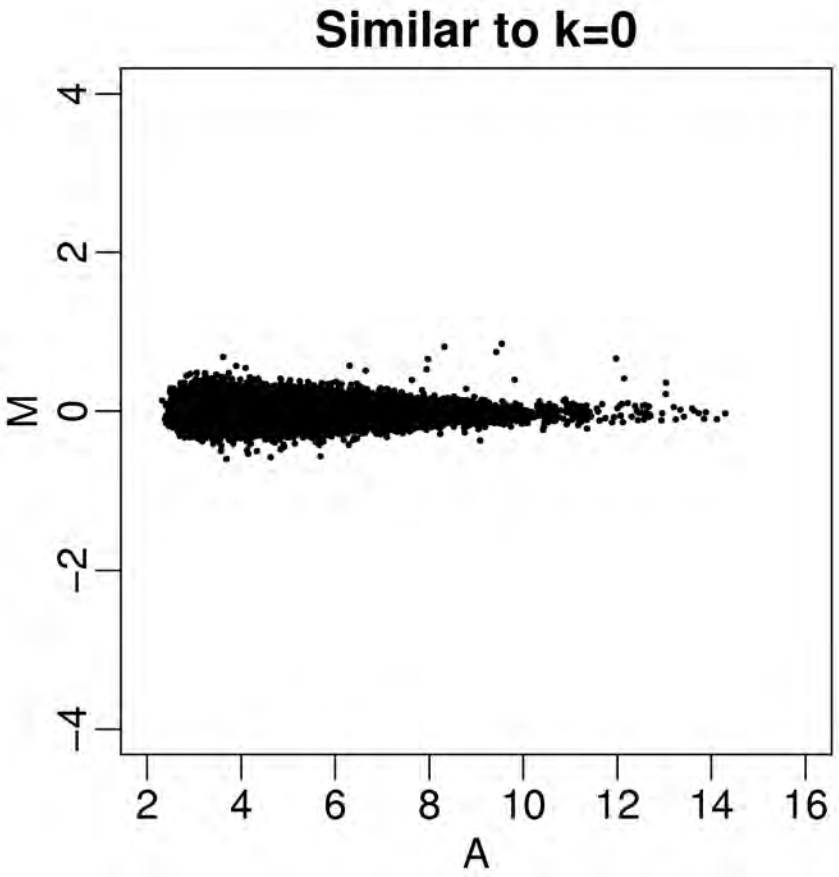
k=1





Real Data


Similar to k=1
M
A
4
2
0
−2
−4
2
4
6
8
10
12
14
16


Similar to k=0
M
A
4
2
0
−2
−4
2
4
6
8
10
12
14
16

RMA Background Adjustment

The Basic Idea:

PM=B+S

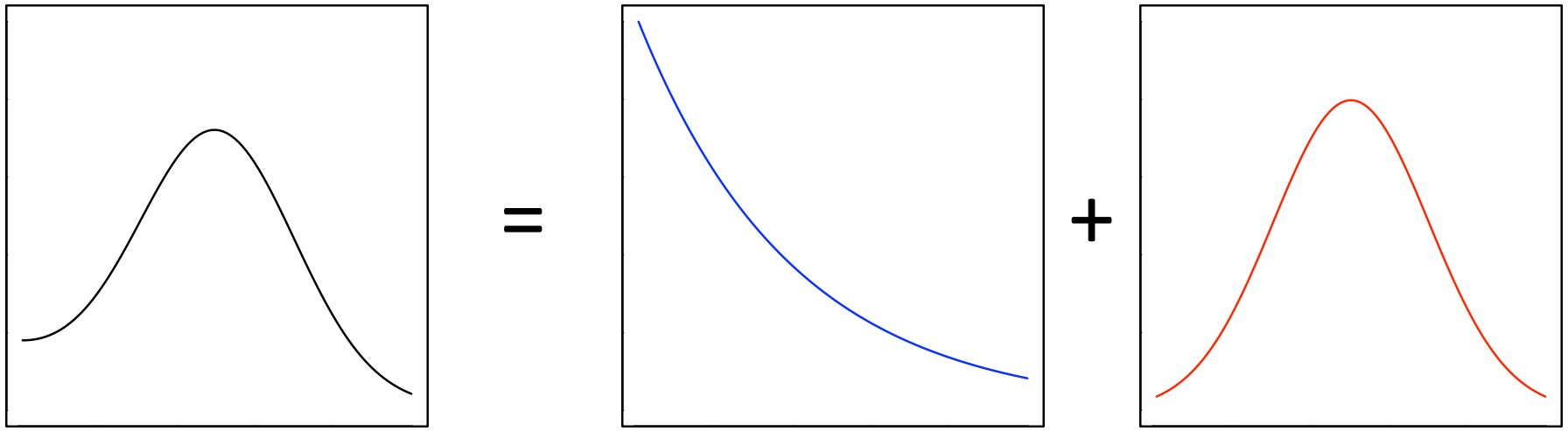
Observed: PM

Of interest: S

Pose a statistical model and use it to predict S from the observed PM

Background Correction

Observed Intensities = Signal + Background Noise



Use the data from all probes to estimate signal/noise distributions

The Basic Idea

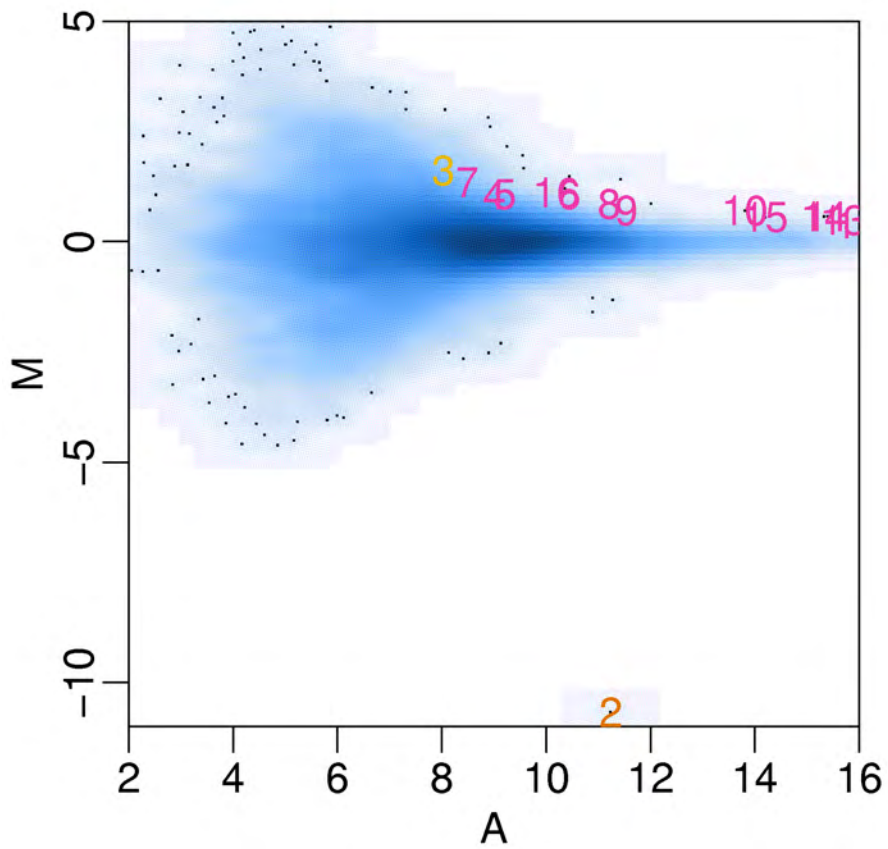
PM=B+S

- A mathematically convenient, useful model
  - *B ~ Normal (μ, σ)*
    S ~ Exponential (λ)

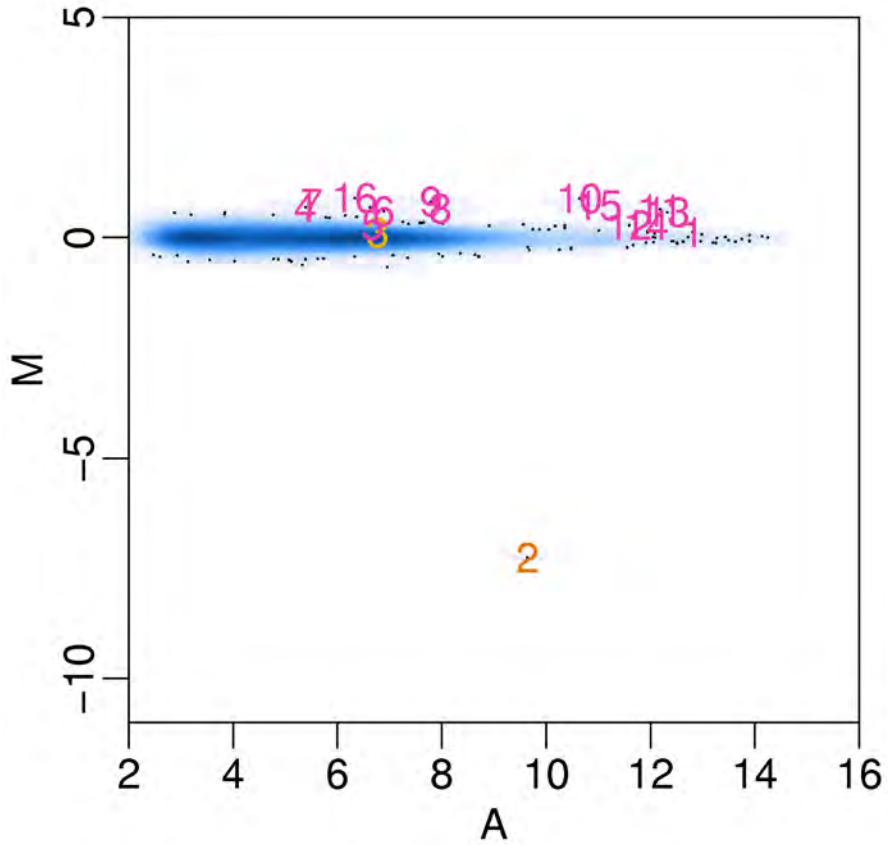
$$\hat{S} = \mathrm{E}[S \mid PM]$$

  - Borrowing strength across probes

MAS 5.0



RMA



Notice improved precision but worst accuracy

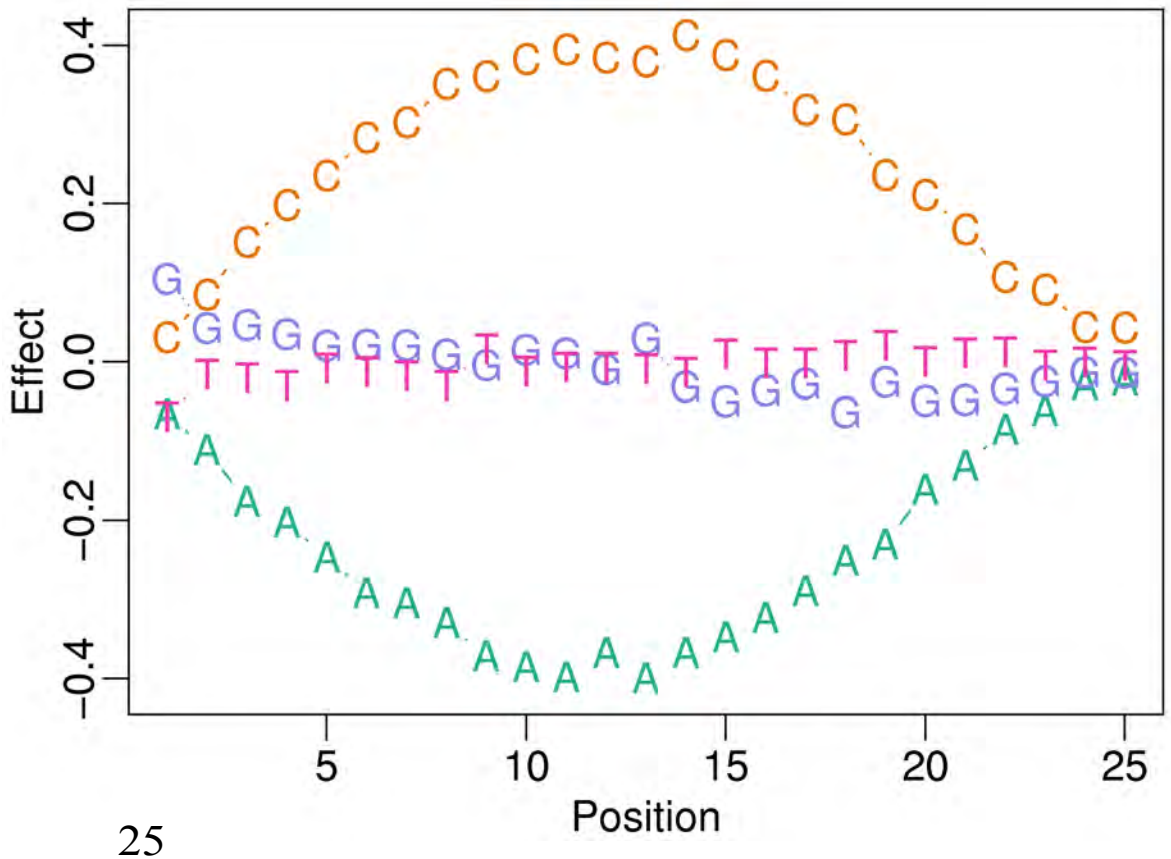
Problem

- Global background correction ignores probe-specific NSB
- MM have problems
- Another possibility: Use probe sequence

Sequence effect

Naef & Magnasco (2003) Nucleic. Acids Res. 31 7

Wu et al. (2004) JASA

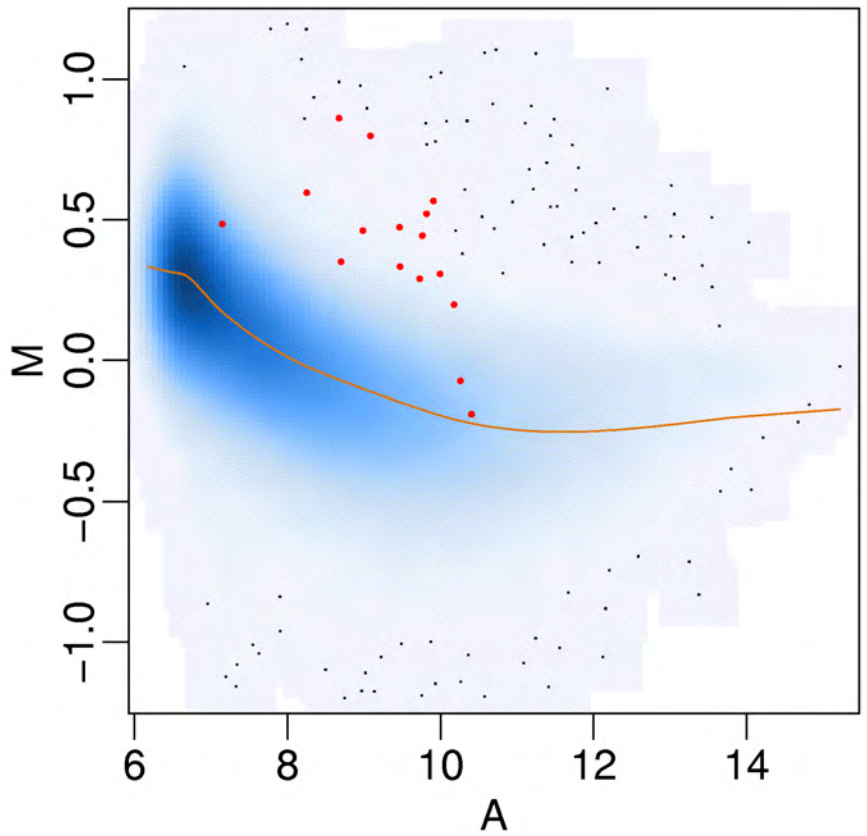


$$Affinity = \sum_{k=1}^{25} \sum_{j \in \{A,T,G,C\}} \mu_{j,k} 1_{b_k = j}$$

$\mu_{j,k}$ ~ *smooth function of k*

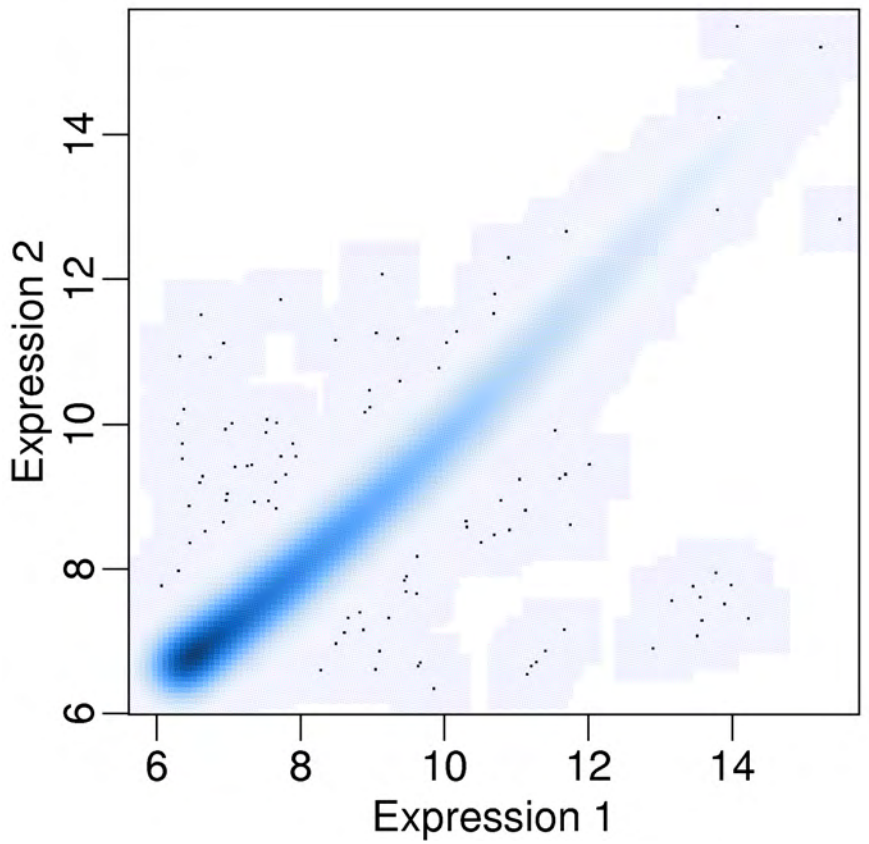
Normalization

Most Common Problem



Intensity dependent effect: Different background level most likely culprit

Scatter Plot



Demonstrates importance of MA plot

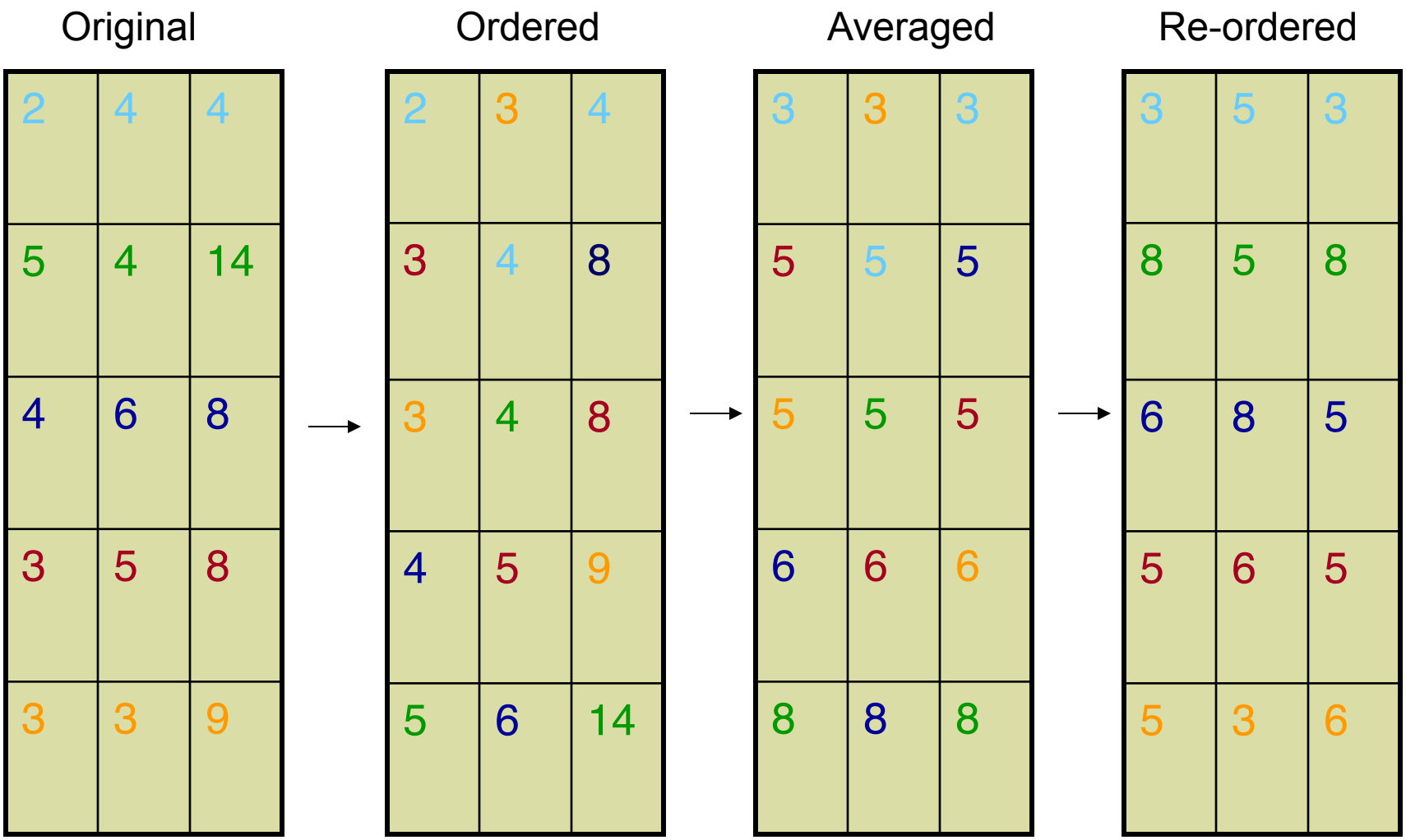
Some Solutions

- Proposed solutions
  - Force distributions (not just medians) to be the same:
    - Amaratunga and Cabrera (2001)
    - Bolstad et al. (2003)
  - Use curve estimators, e.g. loess, to adjust for the effect:
    - Li and Wong (2001) Note: they also use a rank invariant set
    - Colantuoni et al (2002)
    - Dudoit et al (2002)
  - Use adjustments based on additive/multiplicative model:
    - Rocke and Durbin (2003)
    - Huber et al (2002)
    - Cui et al (2003)

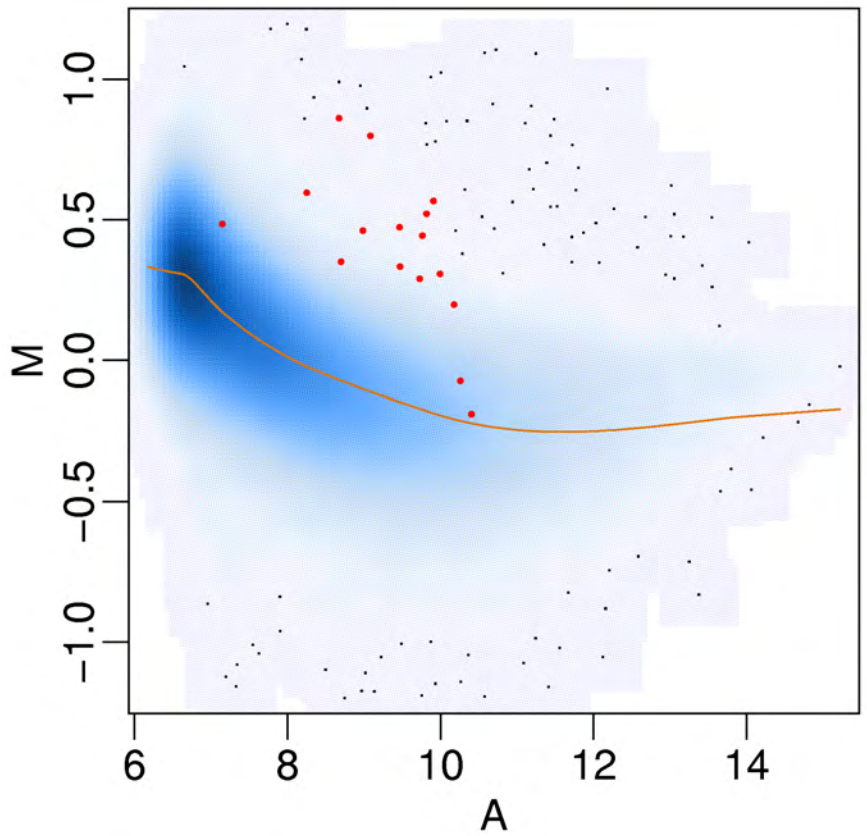
Quantile normalization

- All these non-linear methods perform similarly
- Quantiles is my favorite because its fast and conceptually simple
- Basic idea:
  - order value in each array
  - take average across probes
  - Substitute probe intensity with average
  - Put in original order

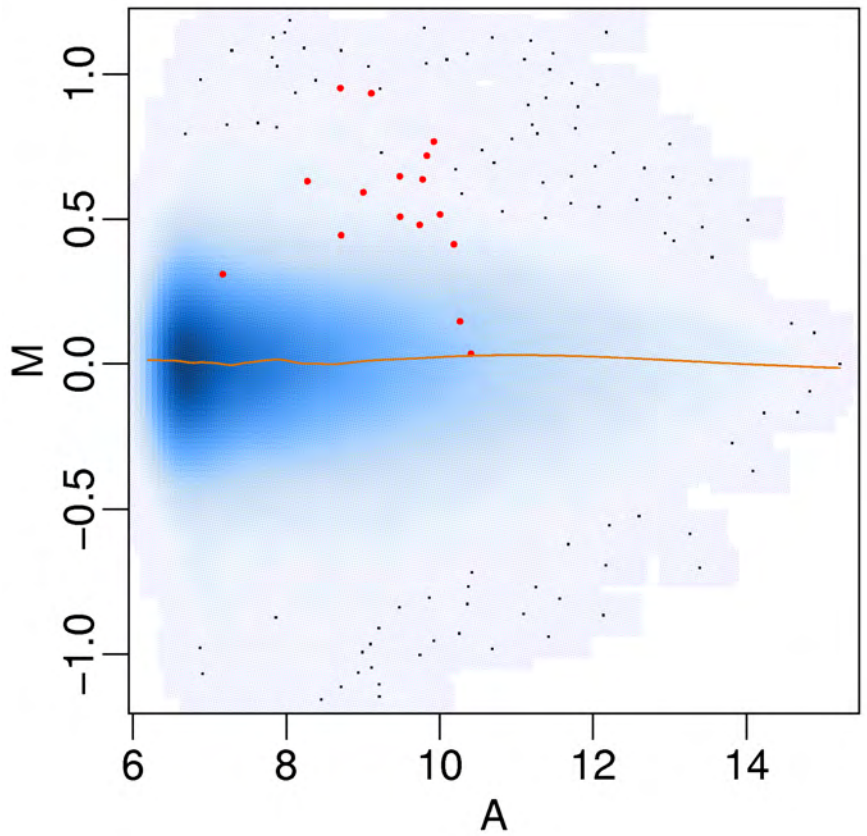
Example of quantile normalization



Before Quantile Normalization



After Quantile Normalization

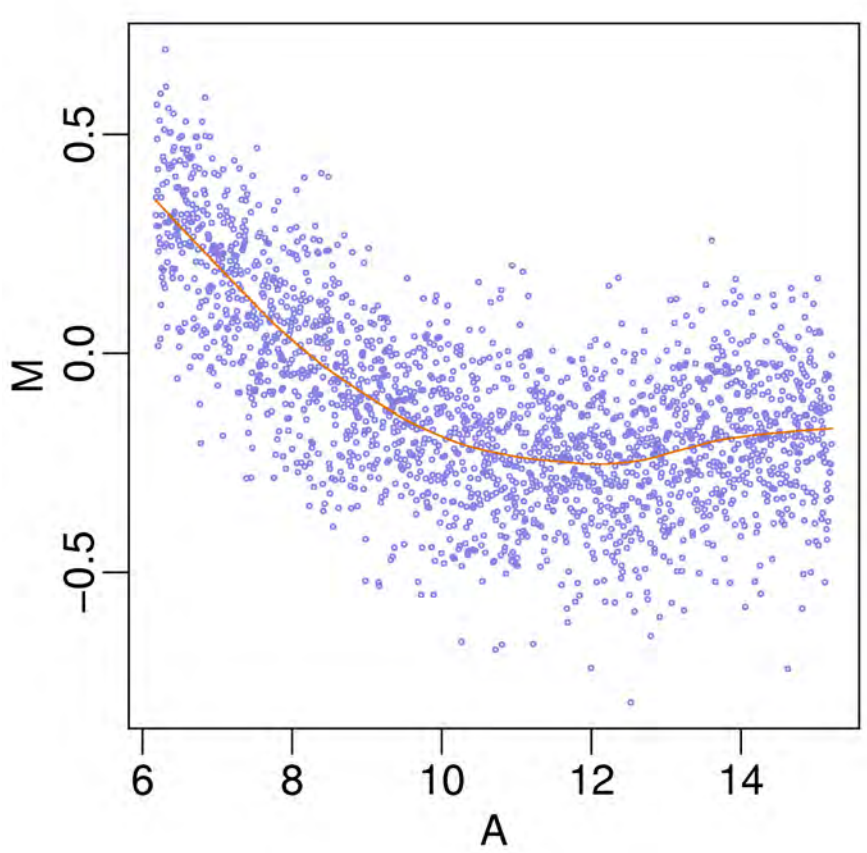


A worry is that it over corrects

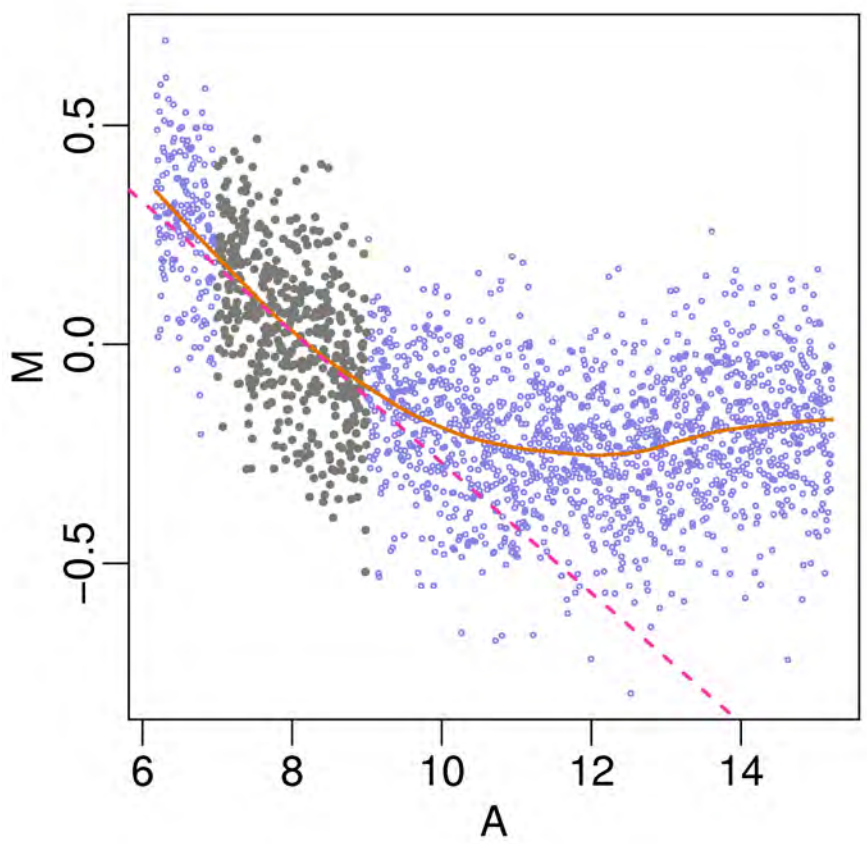
For two color arrays we want to take advantage of pairing

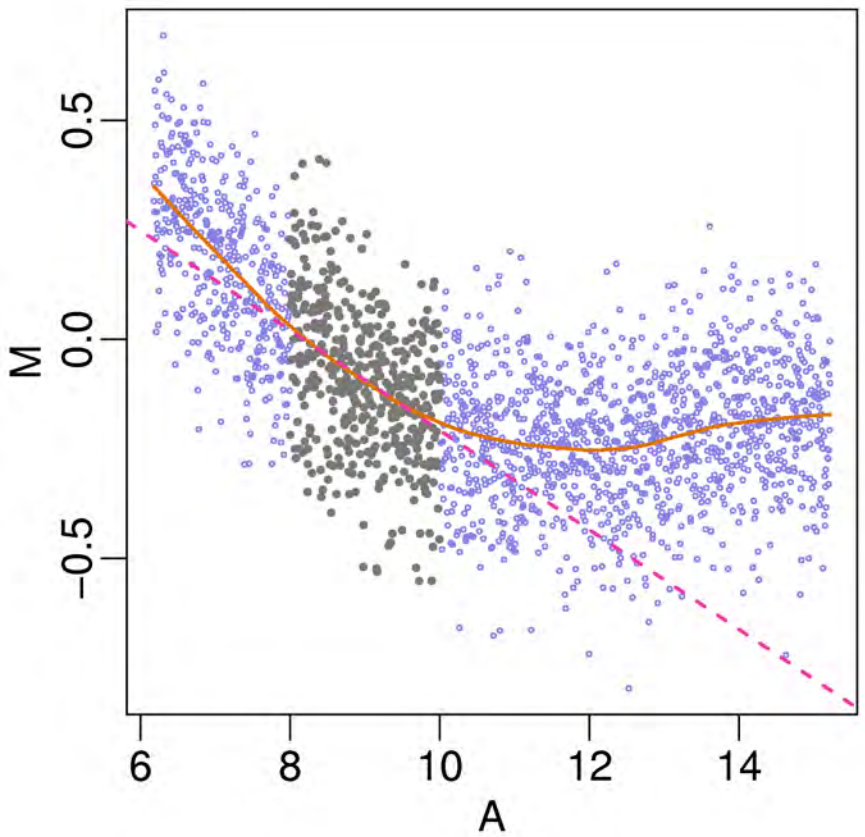
Loess is used to take curves ou

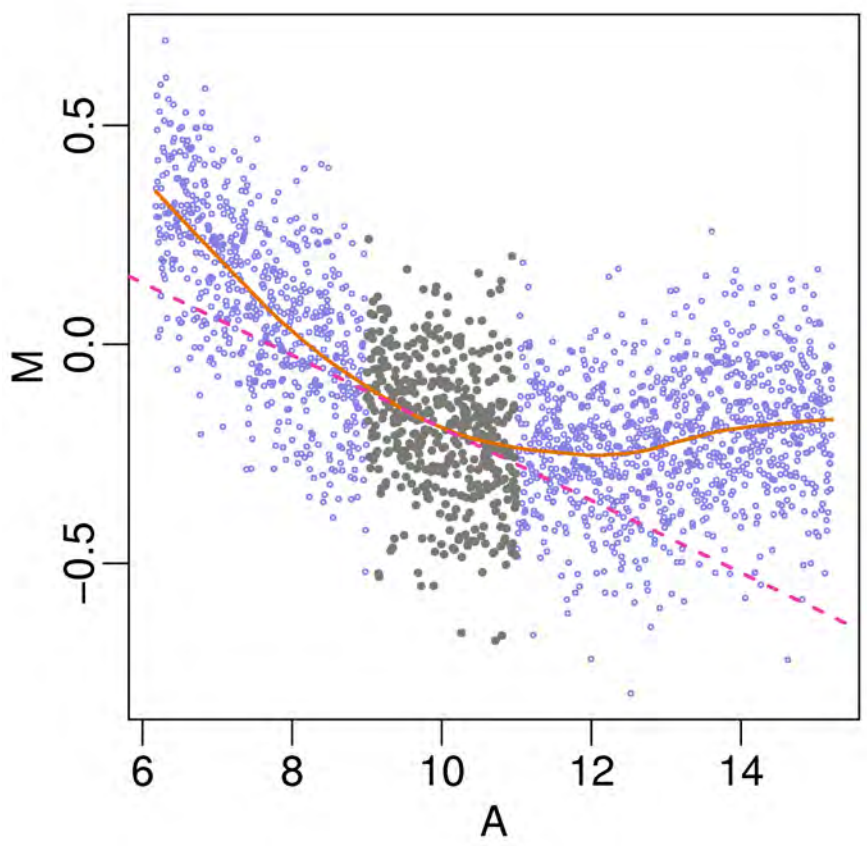
Loess

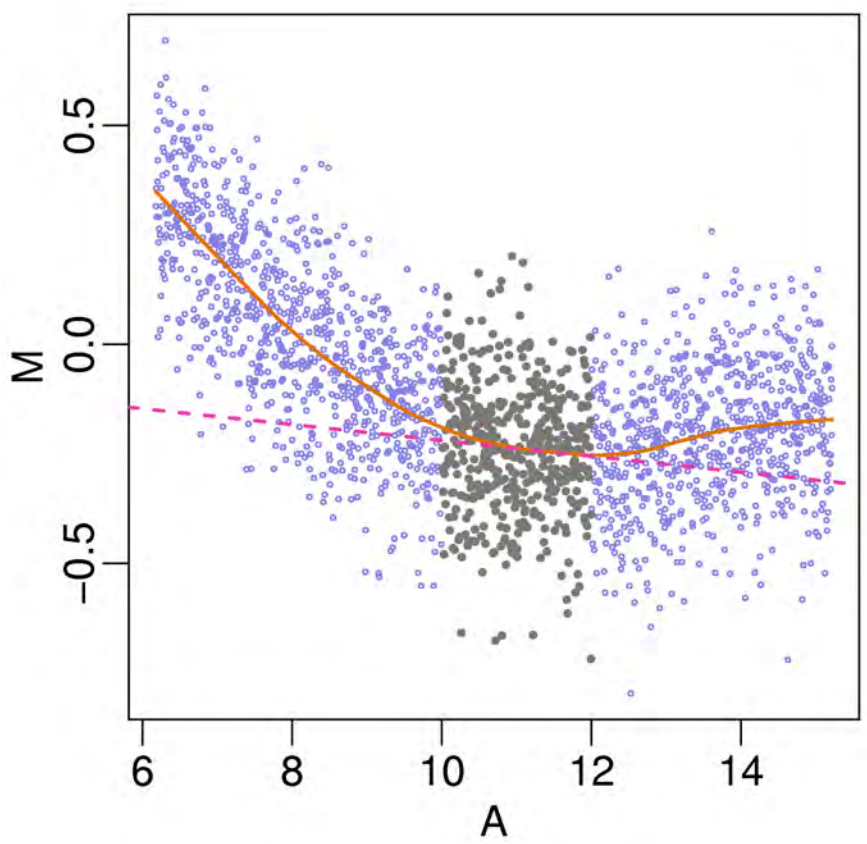


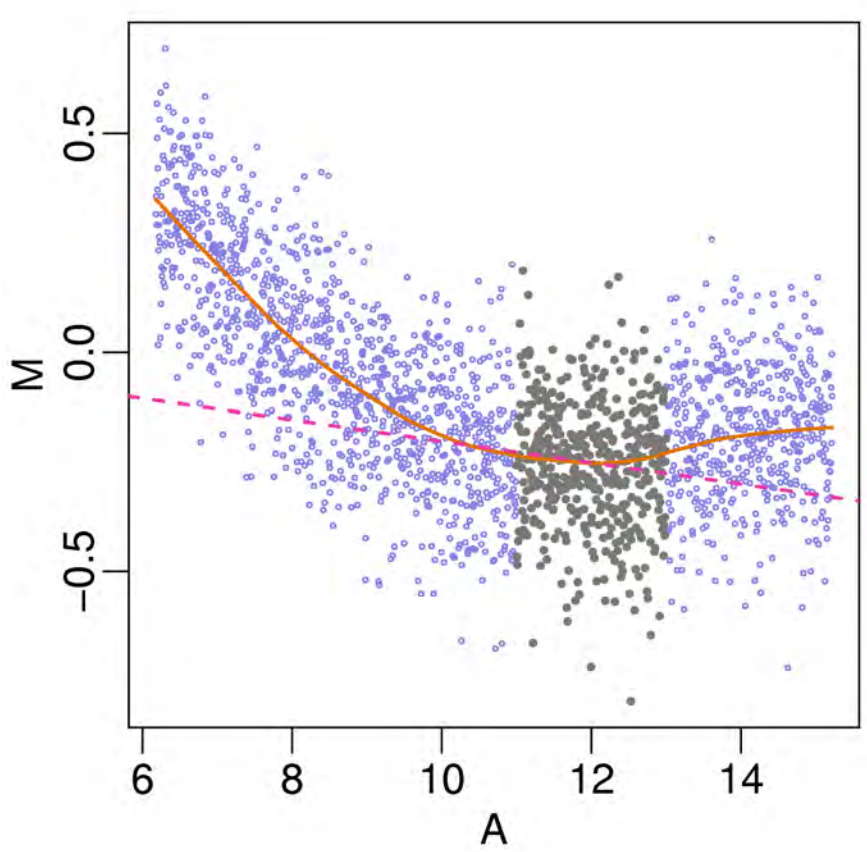
For details look for papers by Terry Speed and Gordon Smyth


0.5
0.0
−0.5
M
6
8
10
12
14
A


0.5
0.0
−0.5
M
6
8
10
12
14
A


0.5
0.0
−0.5
M
6
8
10
12
14
A


0.5
0.0
−0.5
M
6
8
10
12
14
A


0.5
0.0
−0.5
M
6
8
10
12
14
A

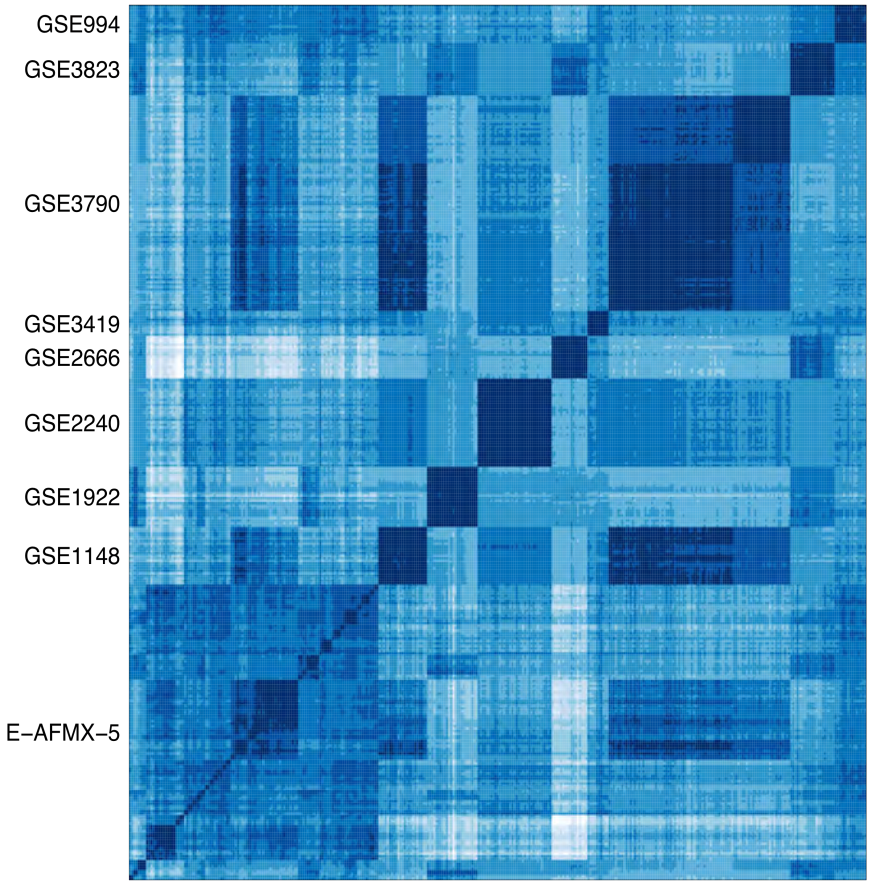
Probe and Batch effects

For details and more

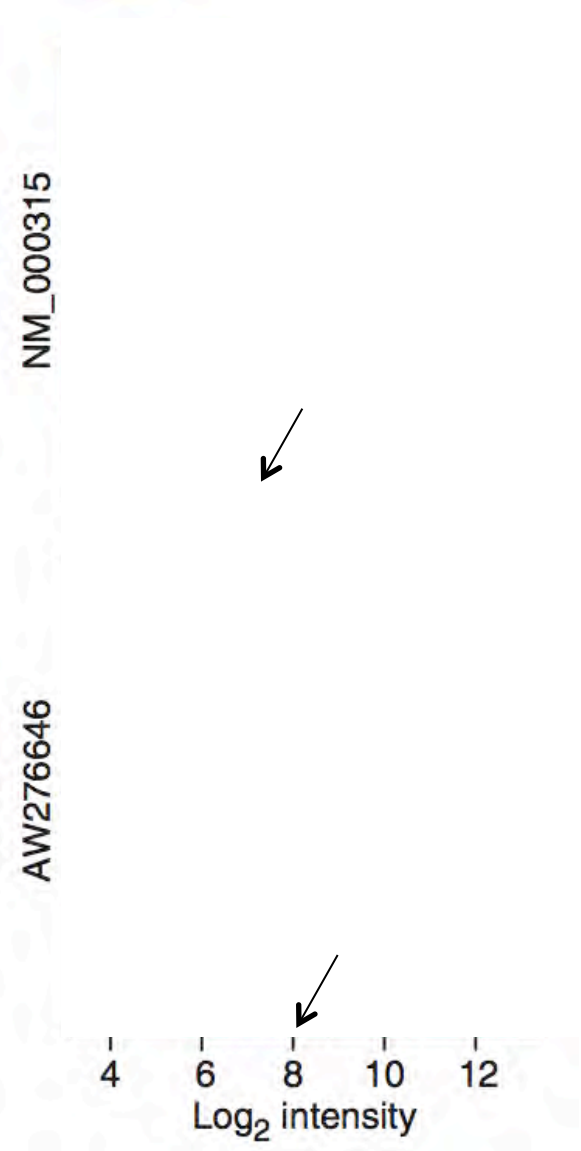
- Zilliox MJ and Irizarry RA (2007) *Nature Methods* 4(11):911-913

- http://rafalab.org

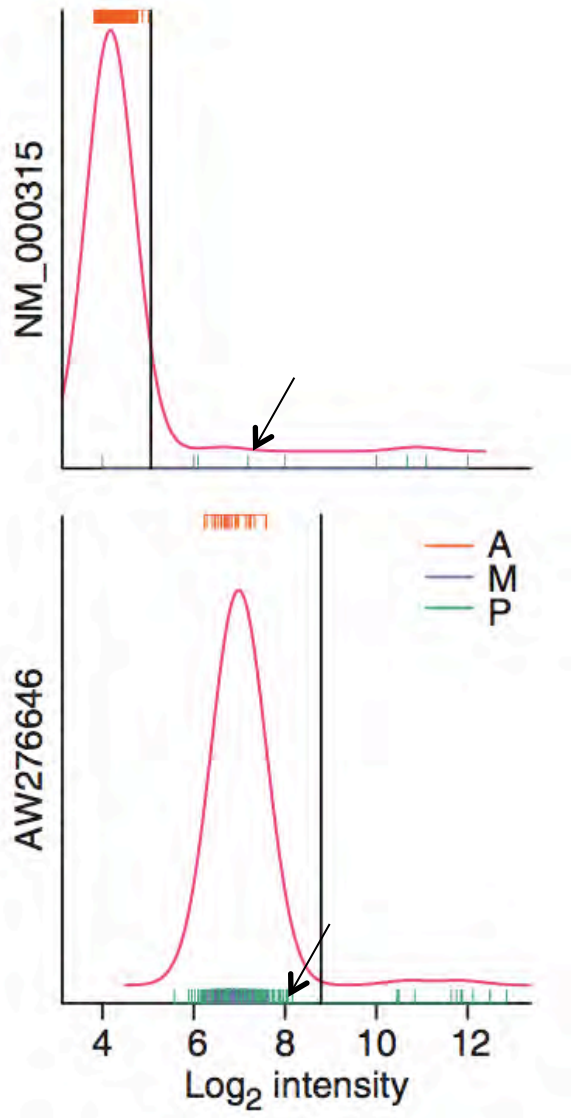
Between Array Correlation



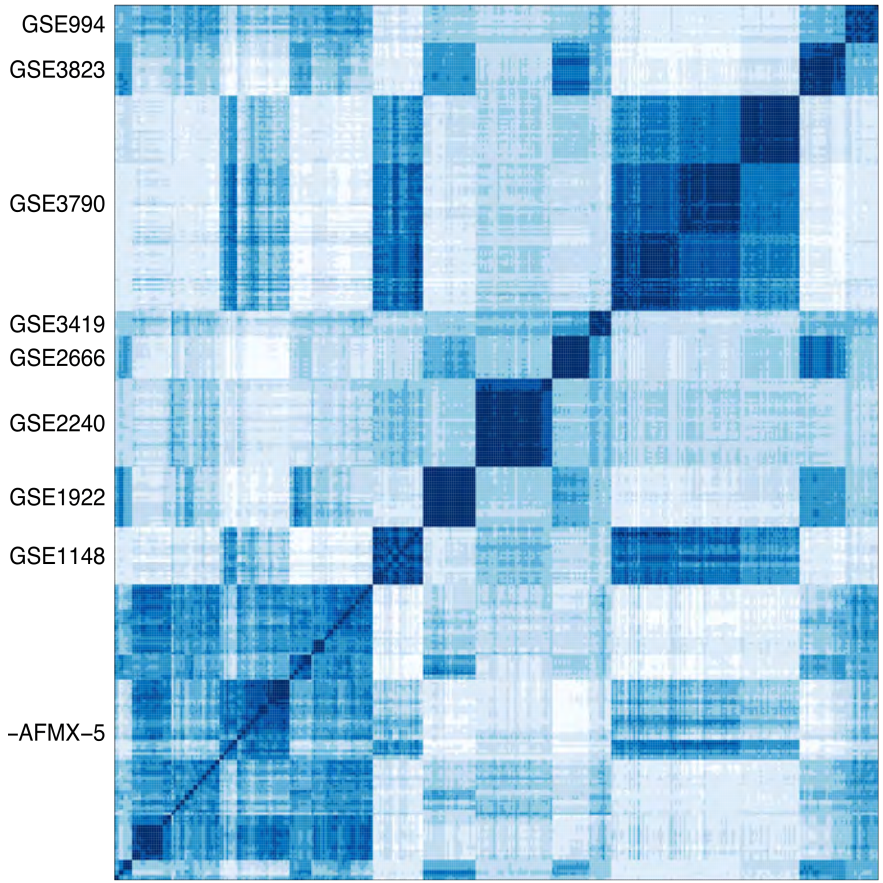
Probe effect



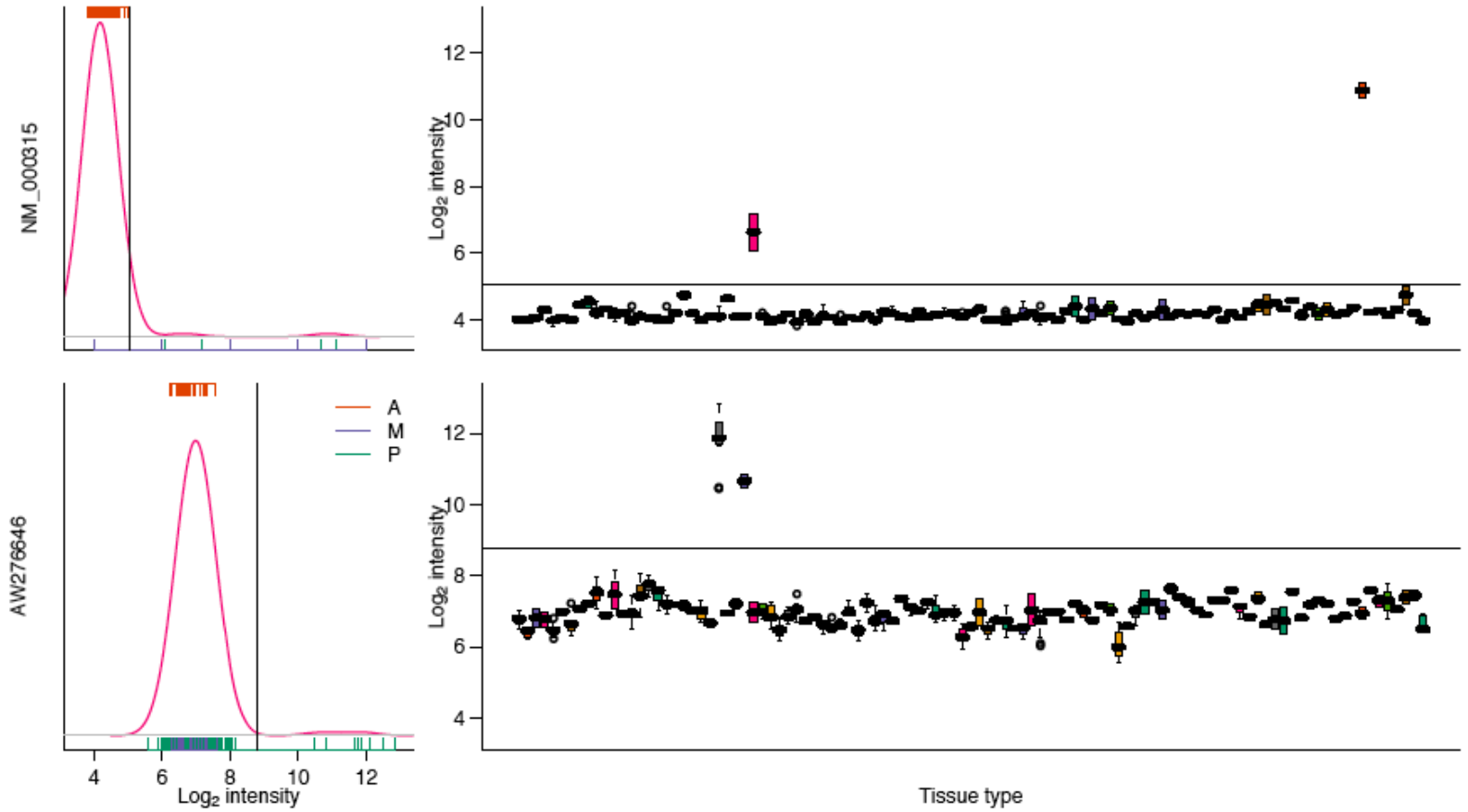
Probe effect



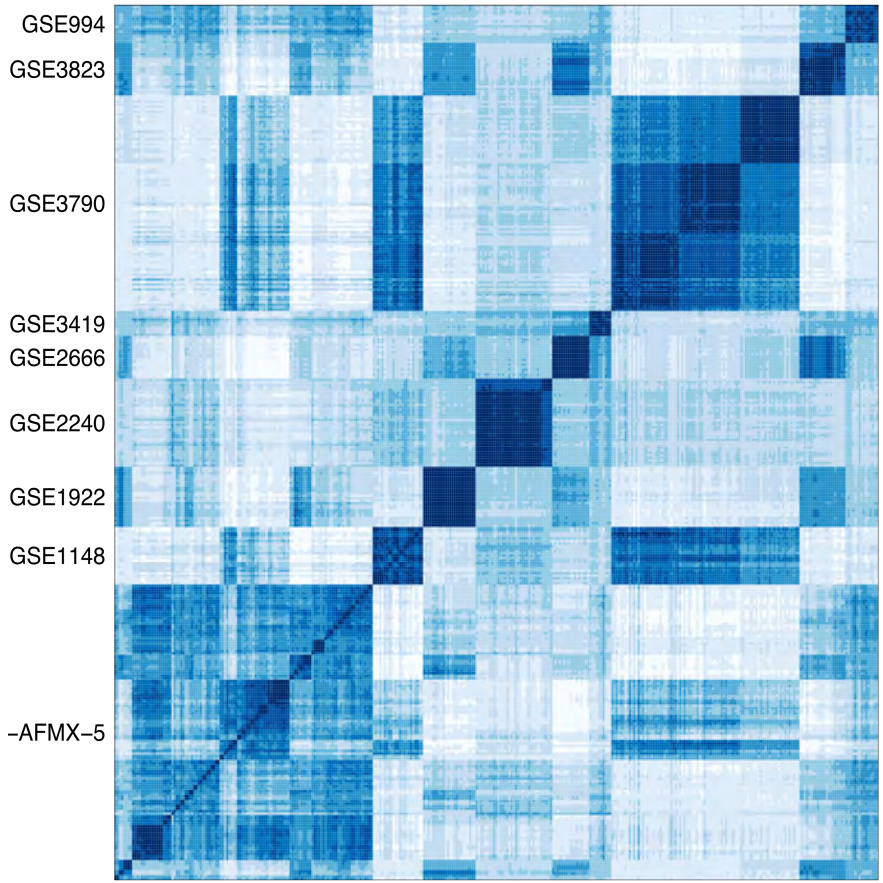
After centering



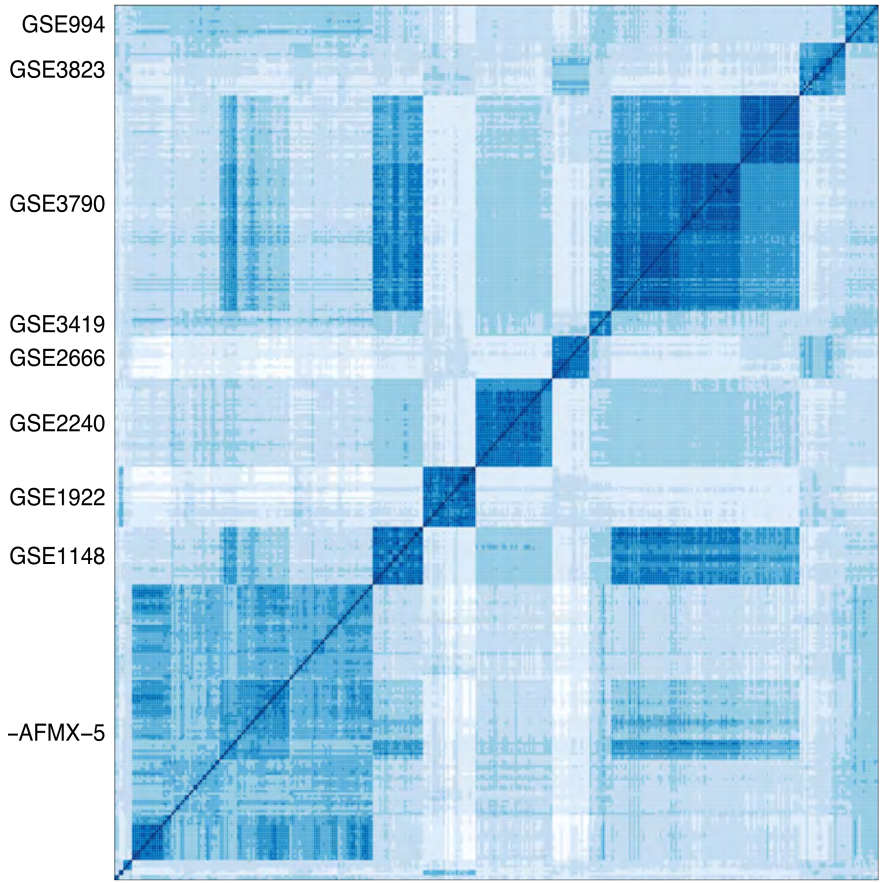
Barcode


NM_000315
AW276646
A
M
P
Log2 intensity
4
6
8
10
12
Log2 intensity
Tissue type

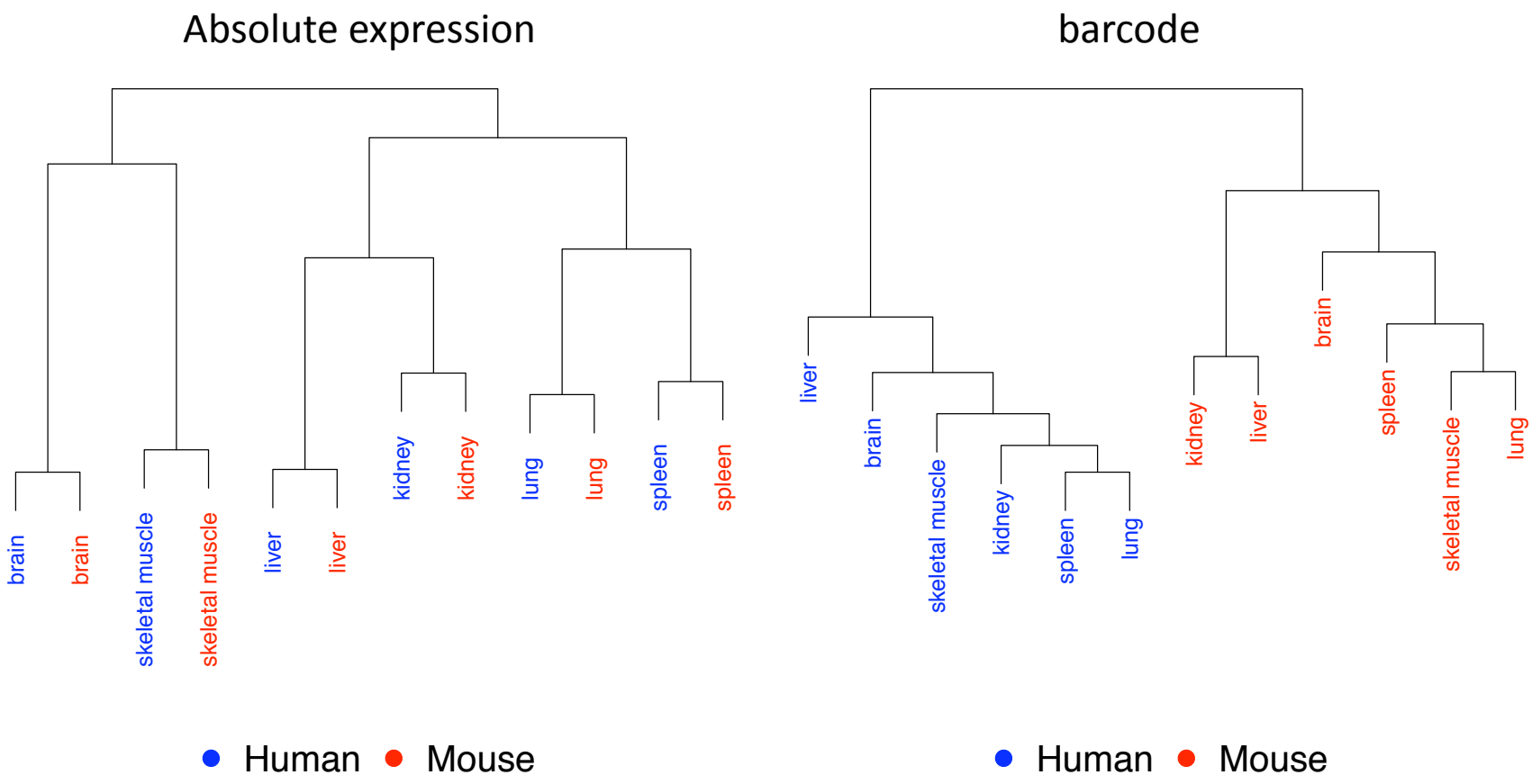
After centering



Barcode

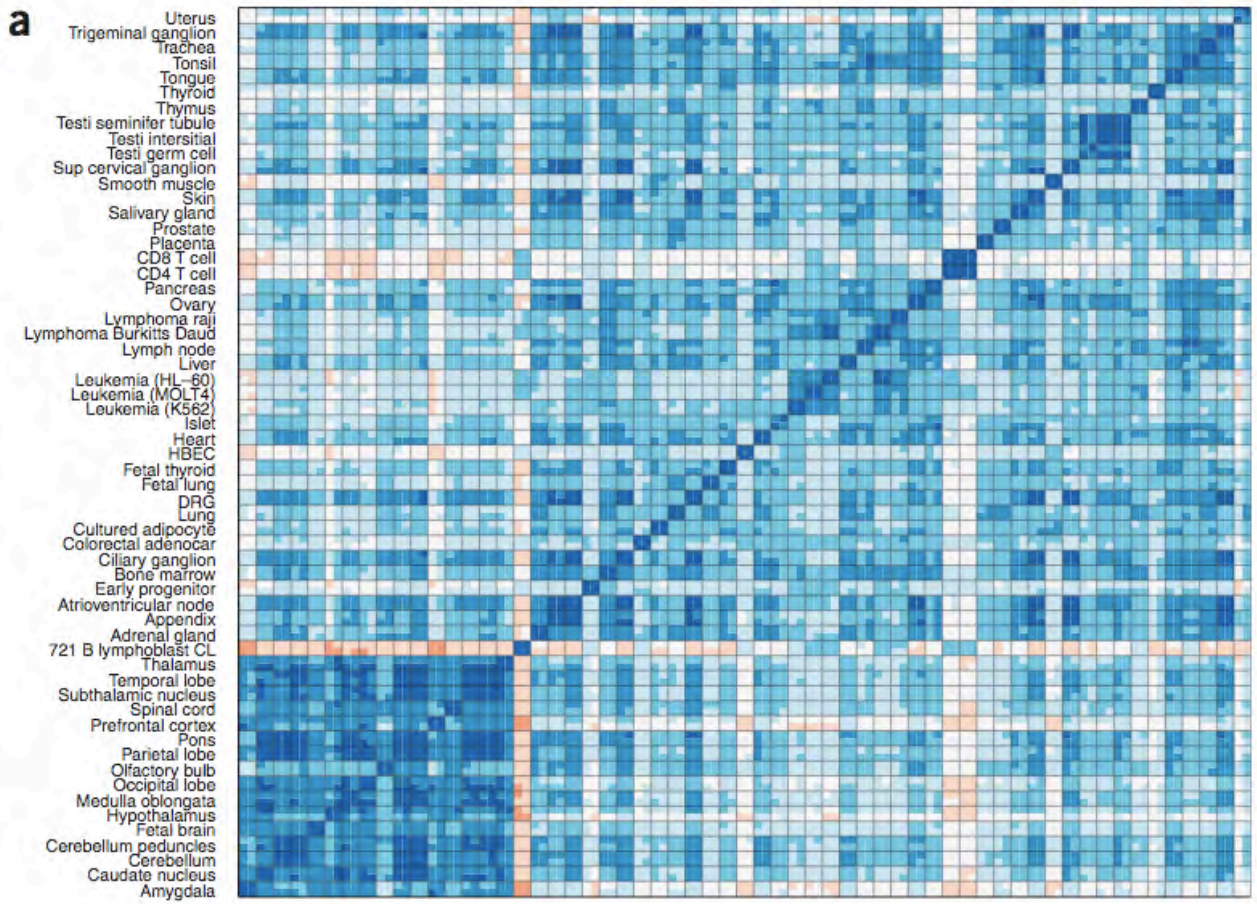

GSE994
GSE3823
GSE3790
GSE3419
GSE2666
GSE2240
GSE1922
GSE1148
-AFMX-5

Consequences

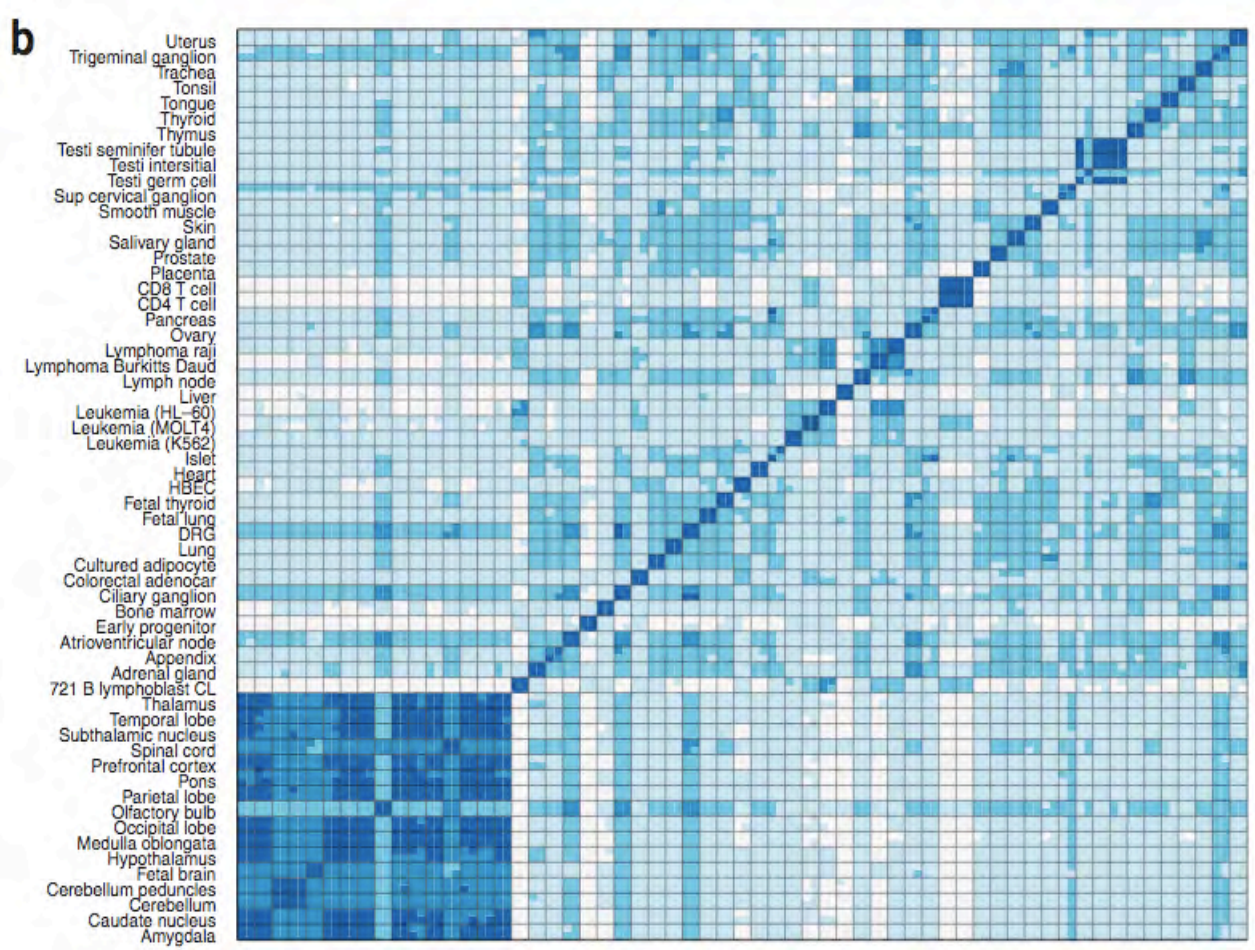


Batch Effect

Close-up (log expression)



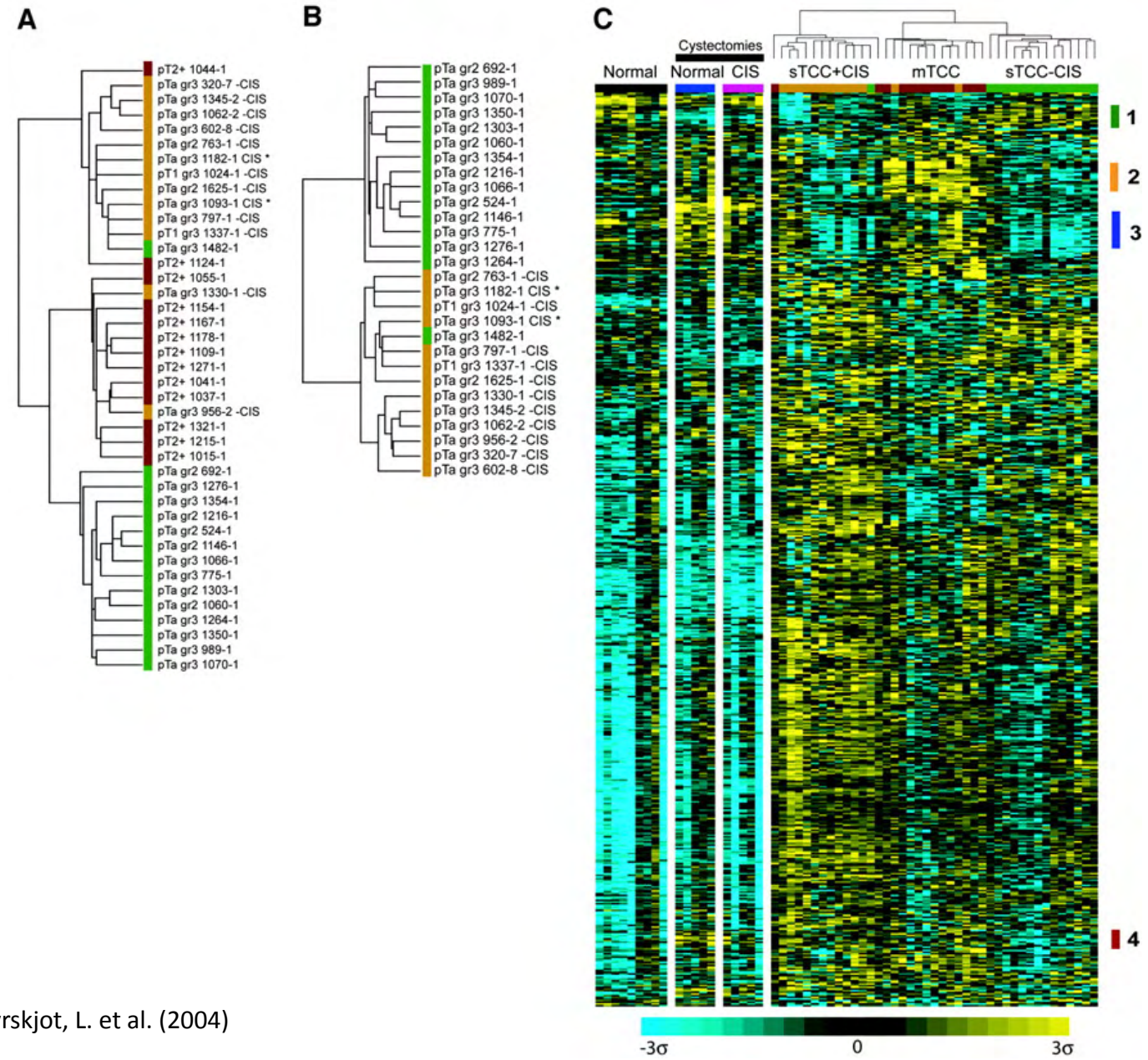
Close-up (barcode)



Stumped by 2 datasets

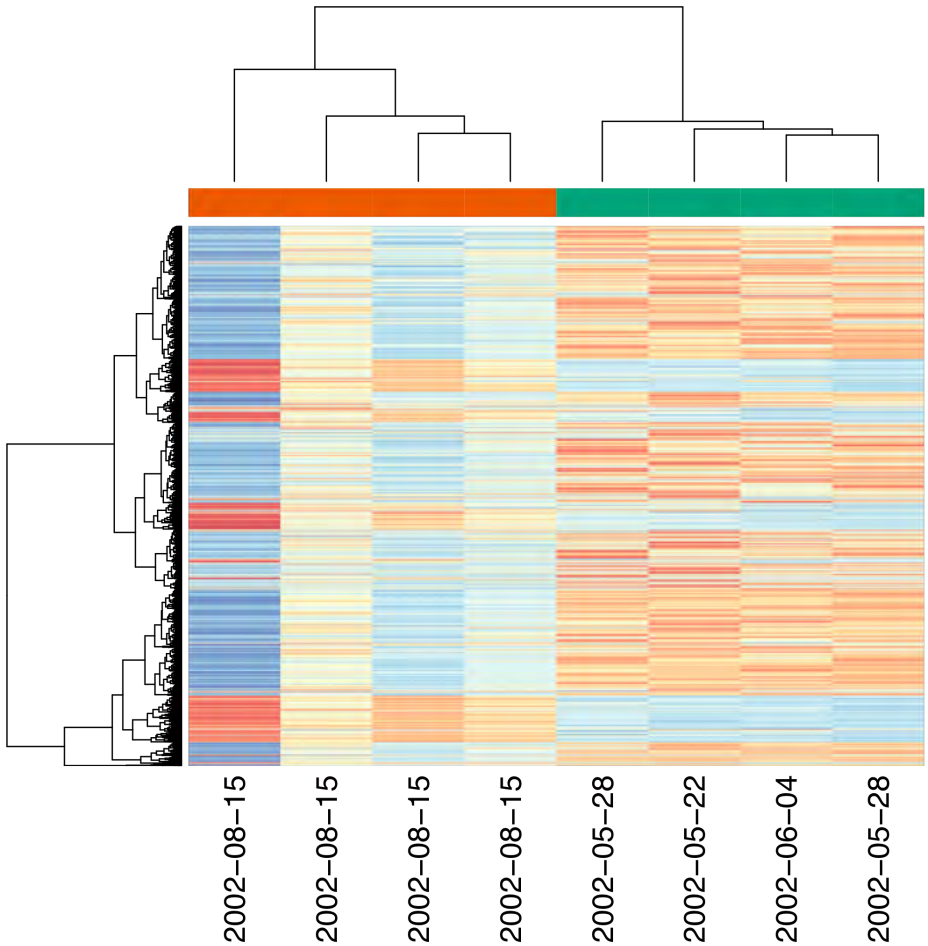
- Potti et al. (2006) Genomic signatures to guide the use of chemotherapeutics. *Nat Med*

- Dyrskjot, L. et al. (2004) Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. *Cancer Res.*

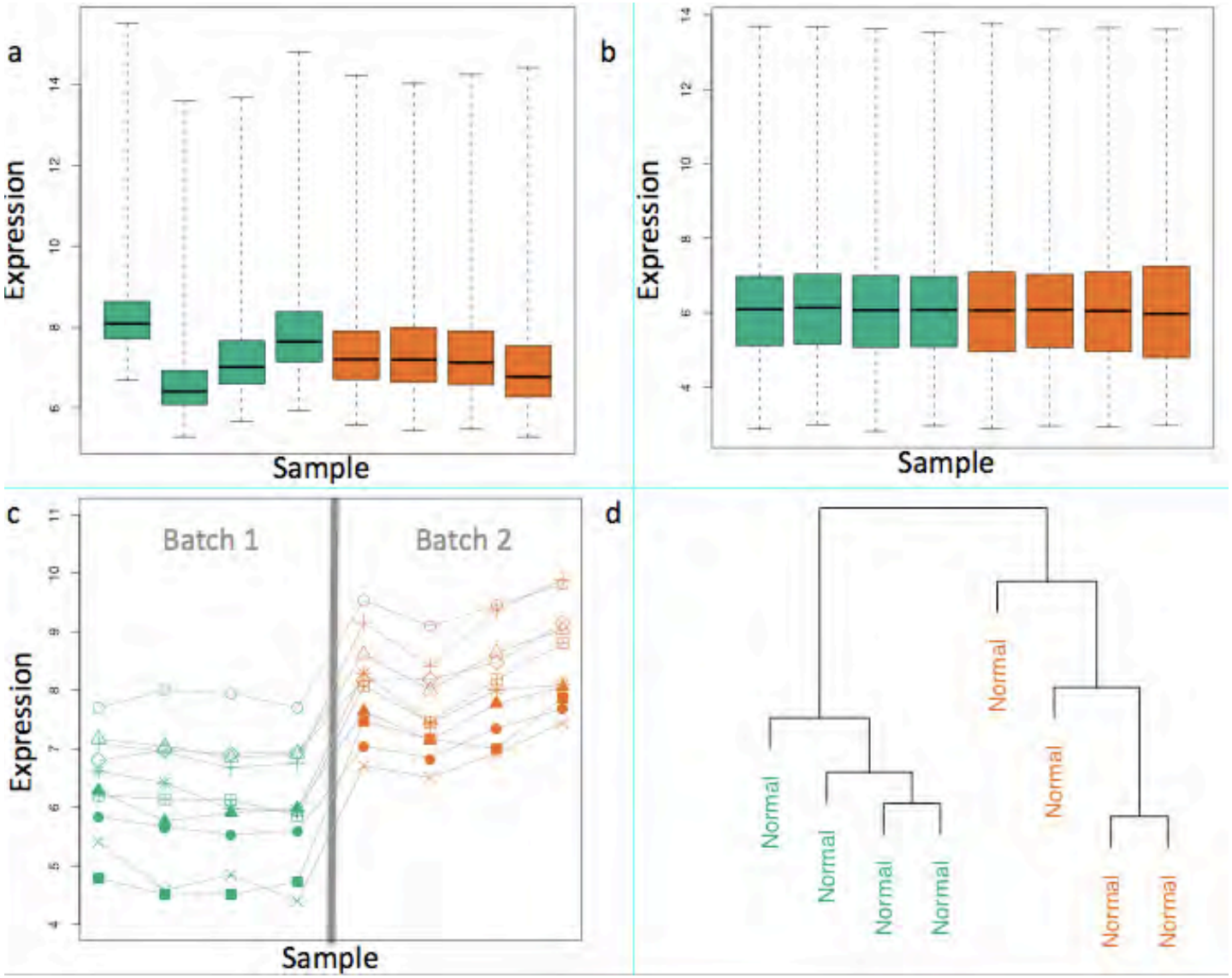


Dyrskjot, L. et al. (2004)

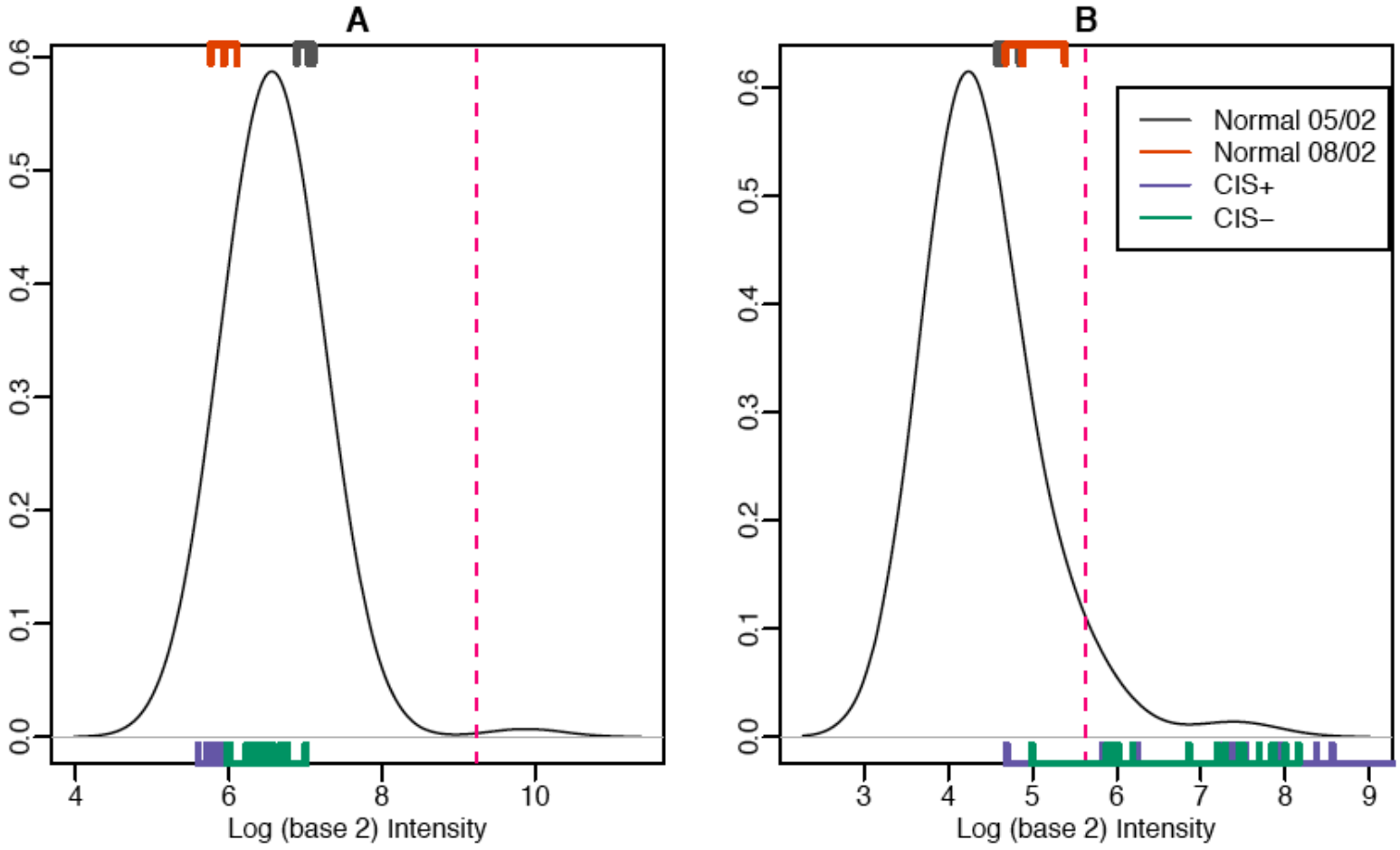
Clustering of normals



Batch Effects survives normalization



Consequence: Artifact versus signal



Probe and batch effect in SNP chips

Affymetrix SNP chip terminology

SNP

Genomic DNA: TACATAGCCATCGGTANGTACTCAATGATGATA

A

G

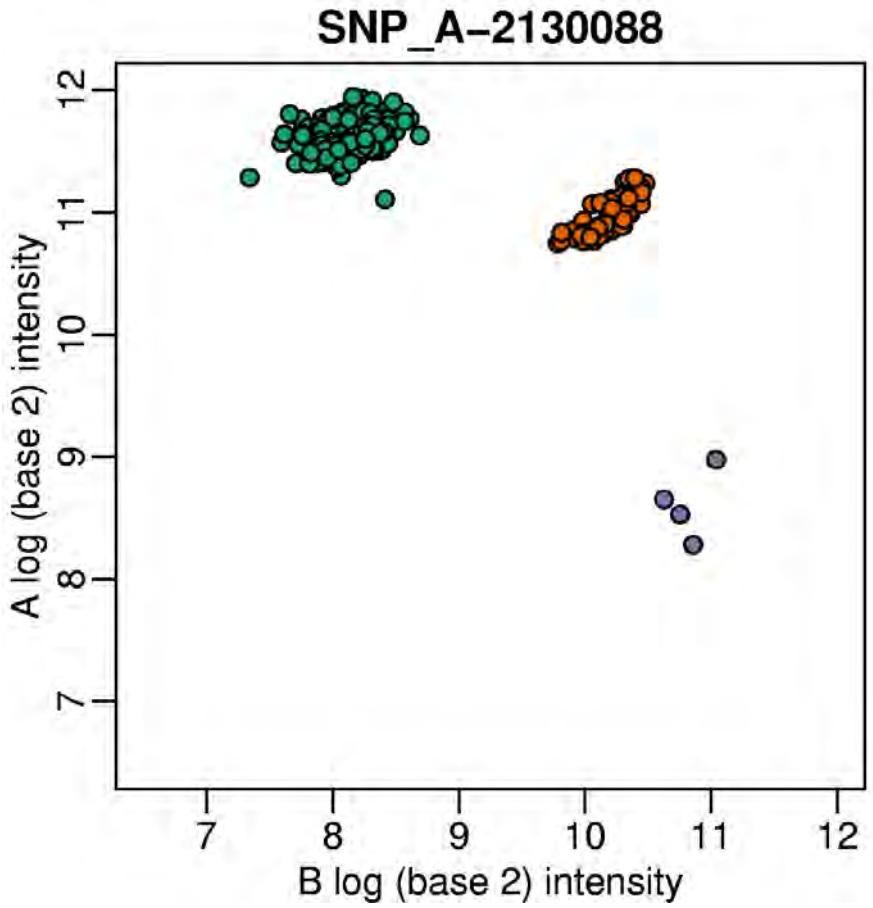
PM probe for Allele A: ATCGGTAGCCATTCATGAGTTACTA

PM probe for Allele B: ATCGGTAGCCATCCATGAGTTACTA

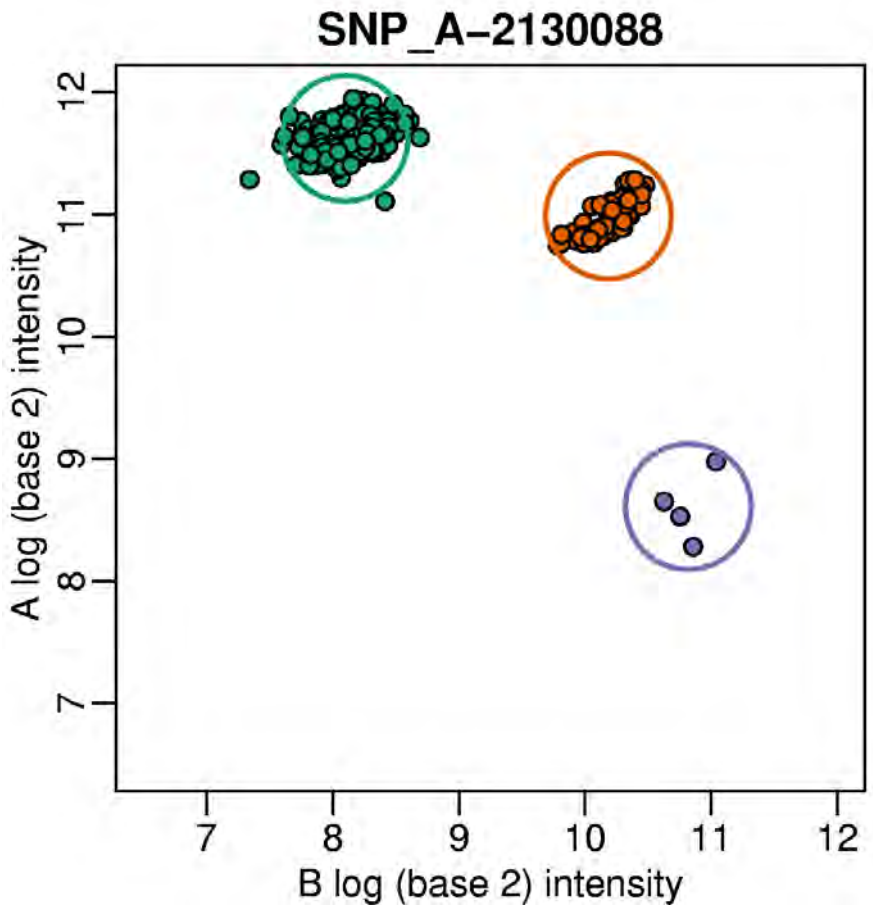
Genotyping: answering the question about the two copies of the chromosome on which the SNP is located:

Is a person AA , AG or GG at this Single Nucleotide Polymorphism?

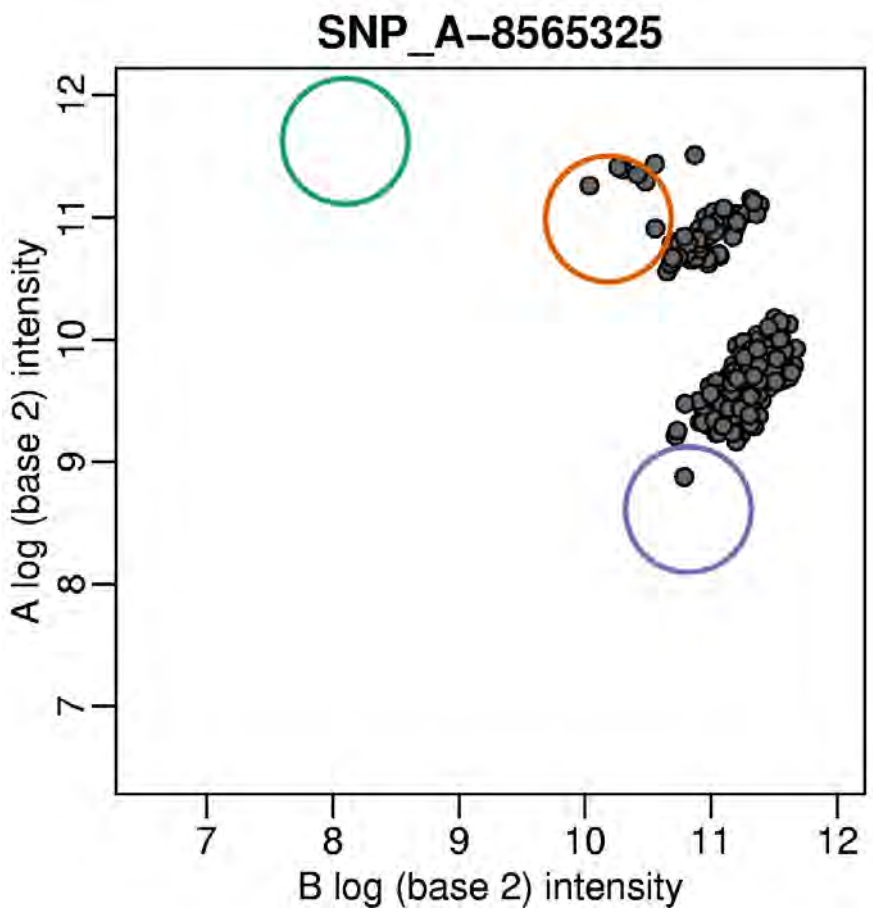
Infer genotype from data



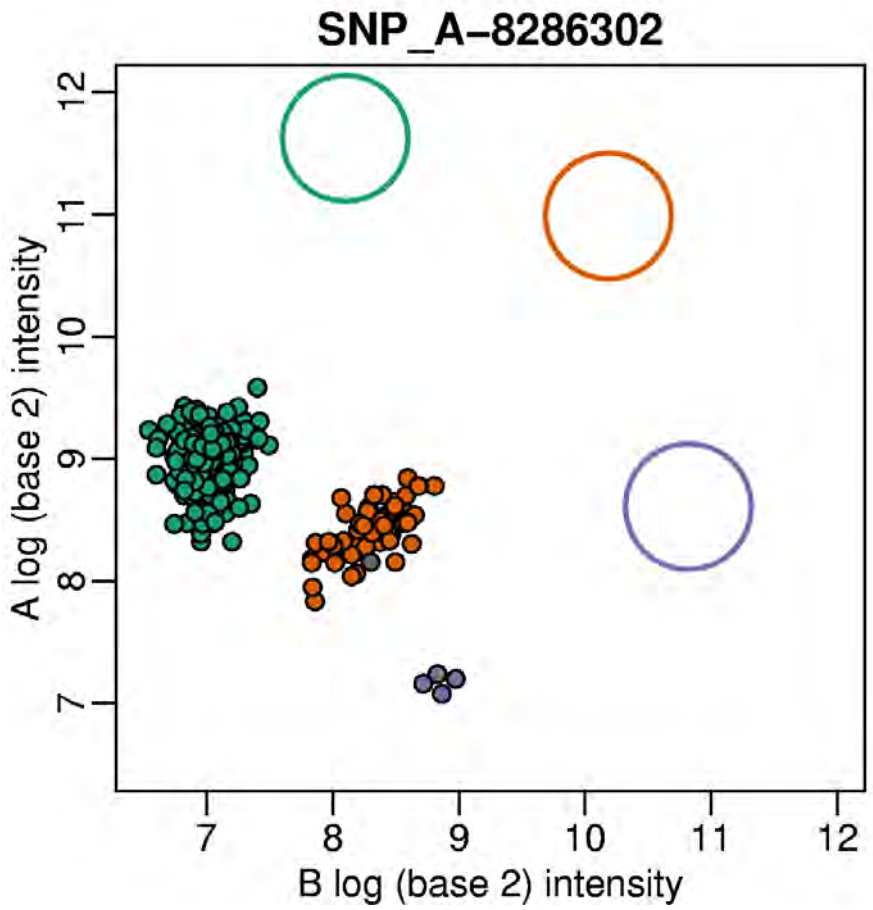
Using regions



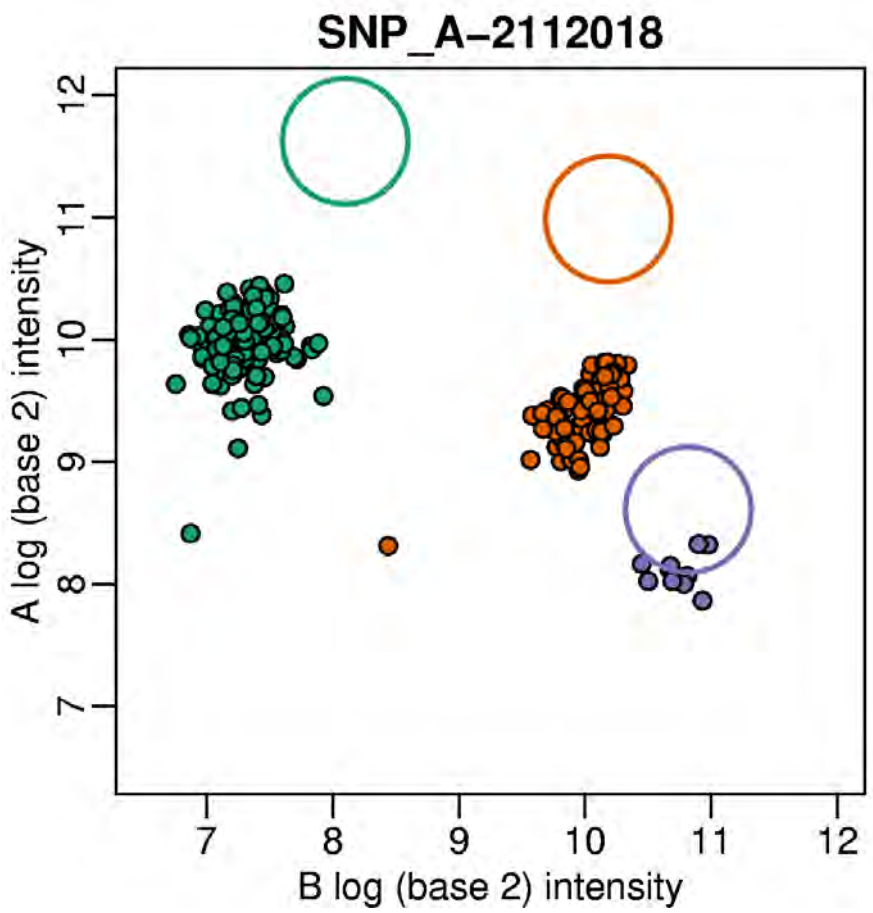
Probe effect



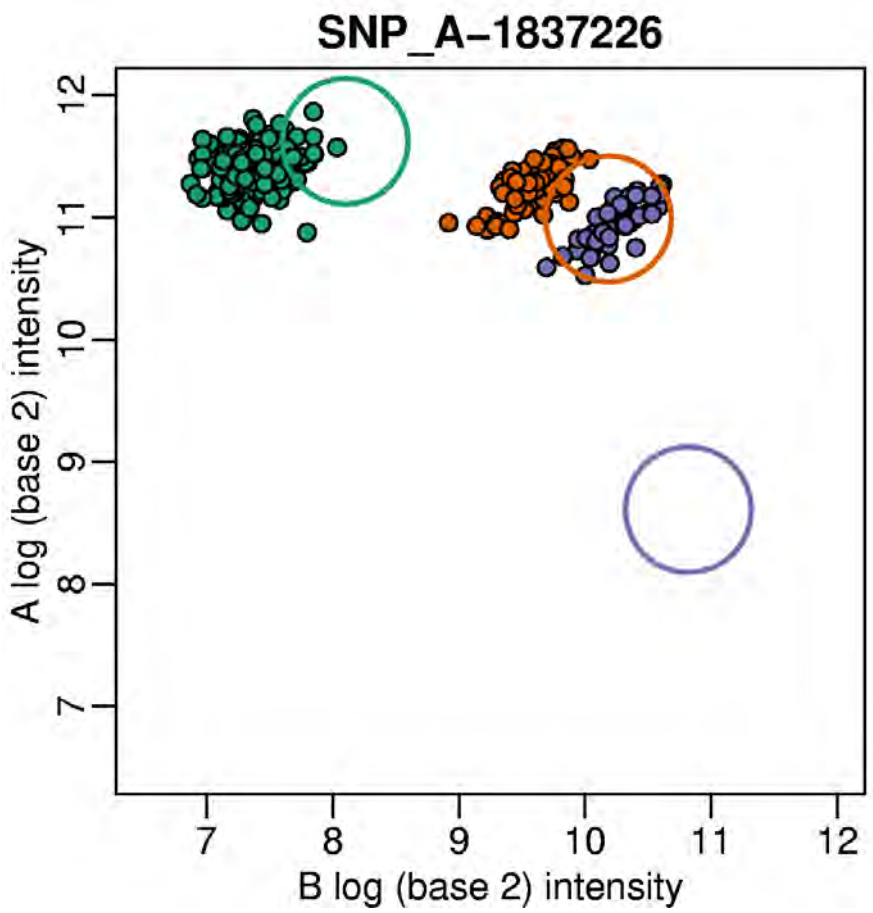
Probe effect



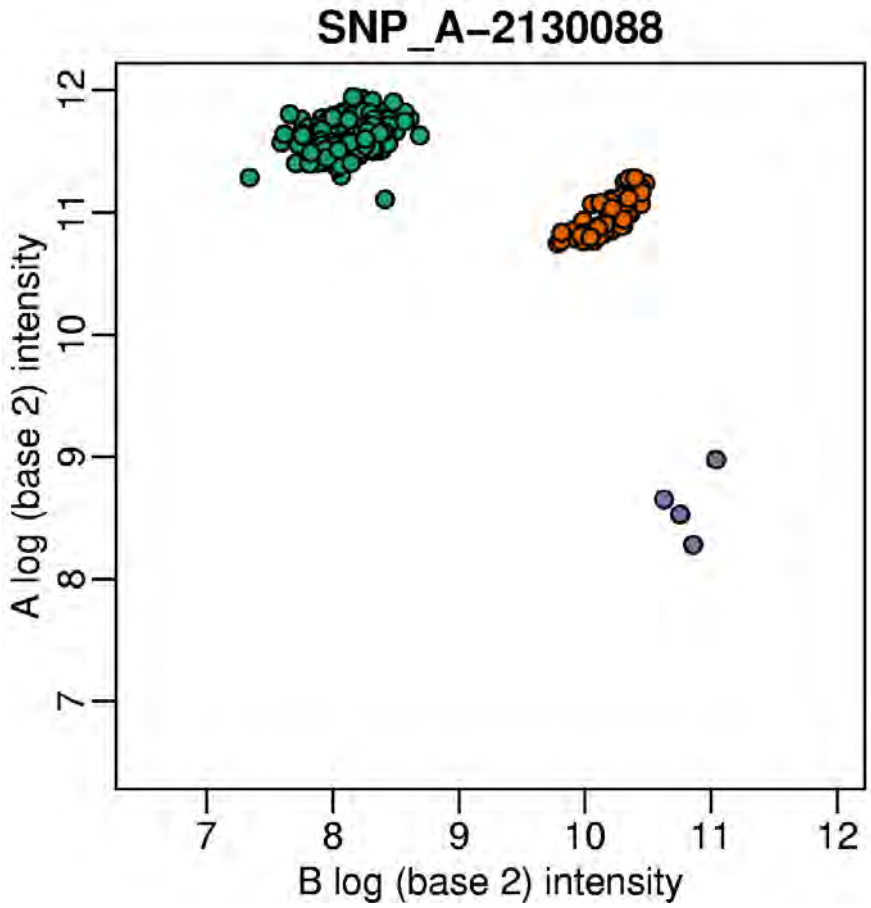
Probe effect



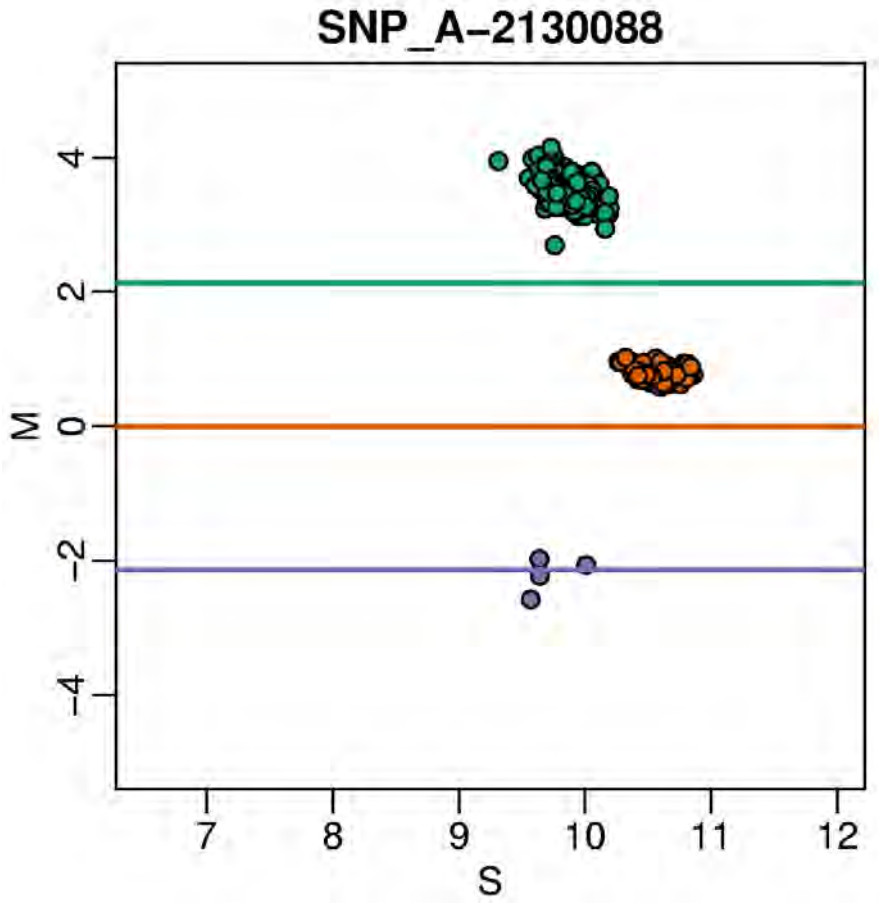
Probe effect



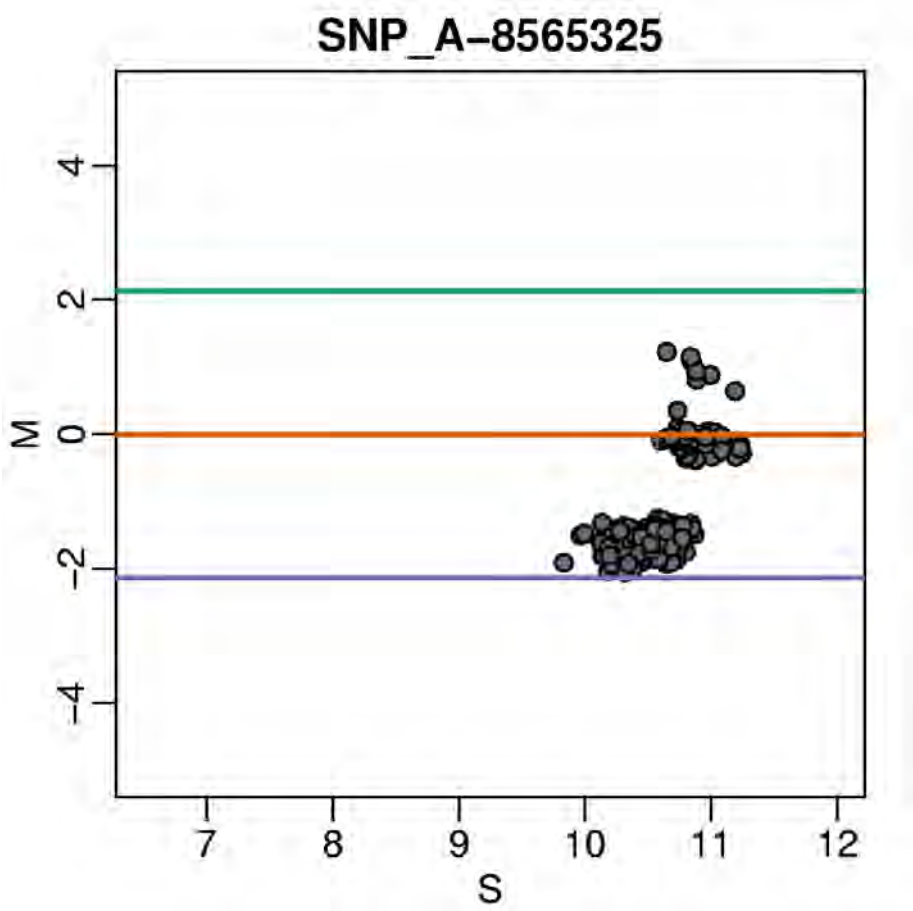
Most information is in *M*



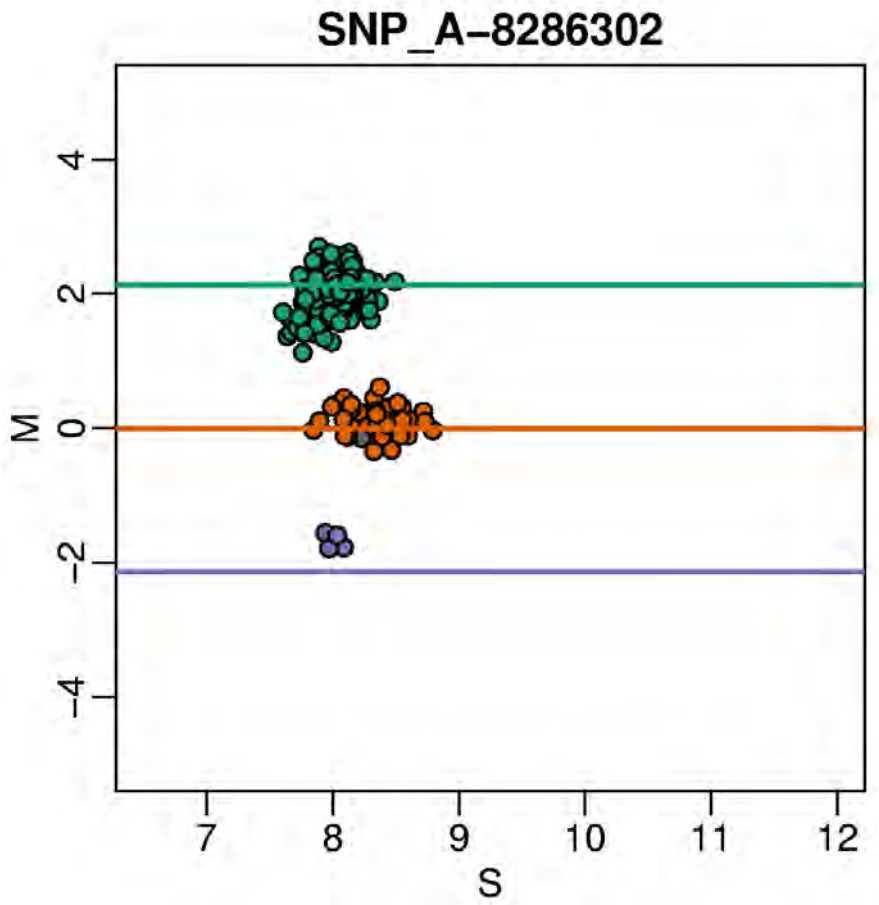
Most information is in *M*



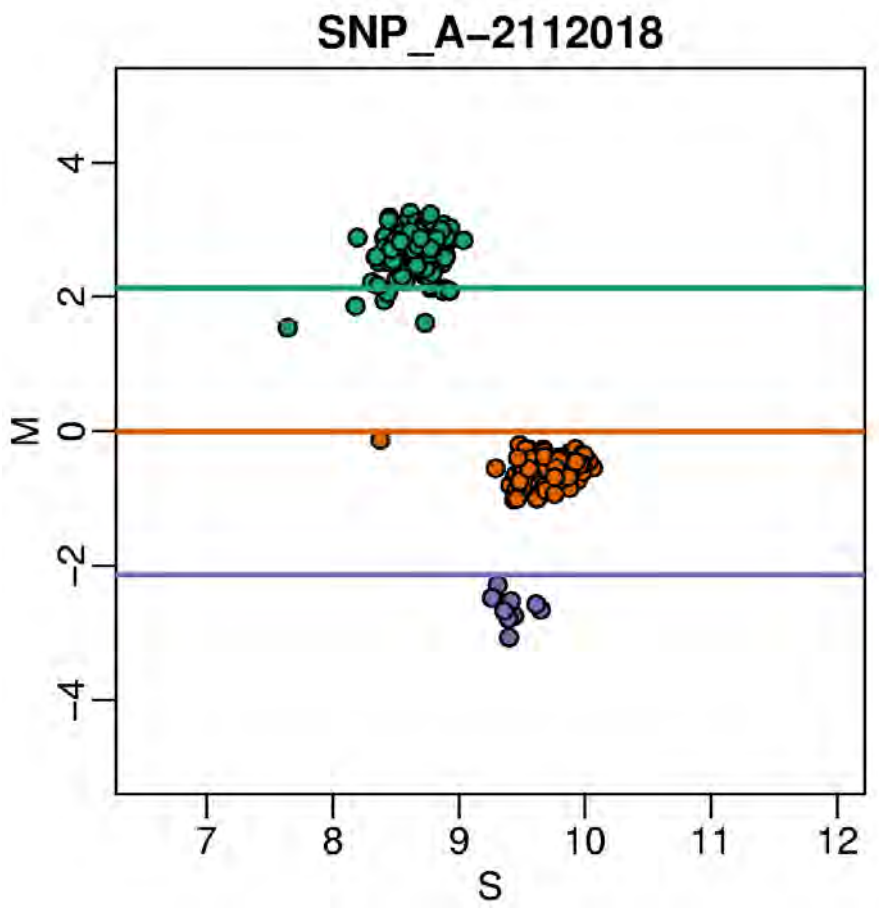
M more stable than *S*



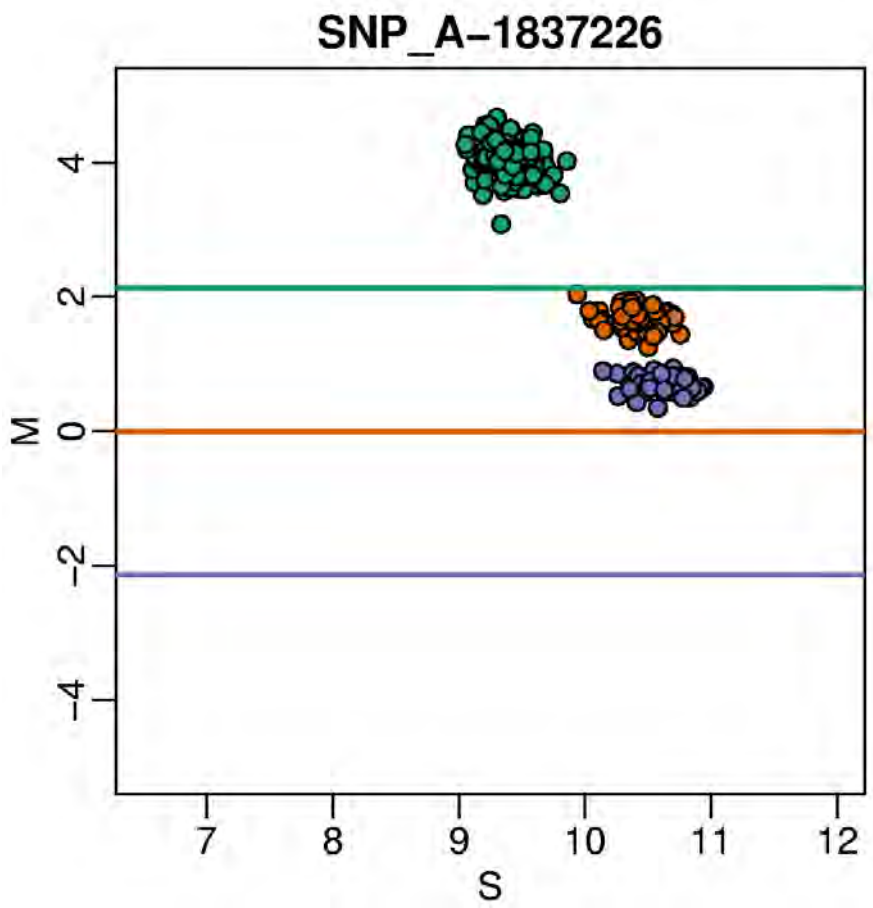
M more stable than *S*



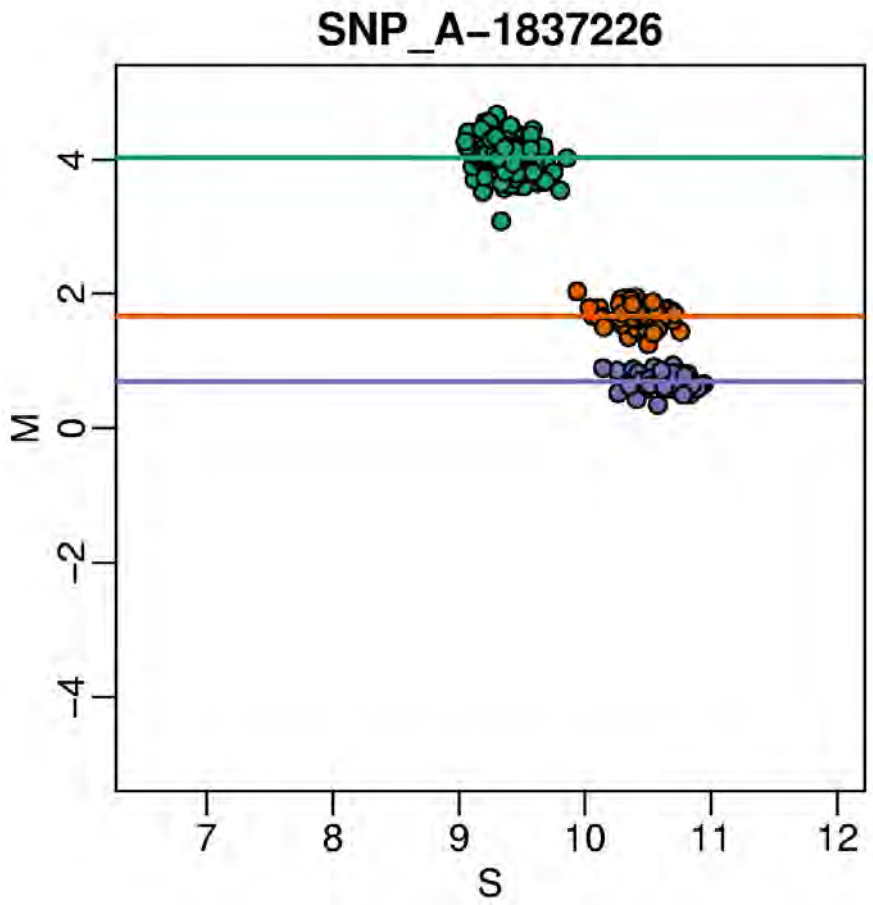
M more stable than *S*



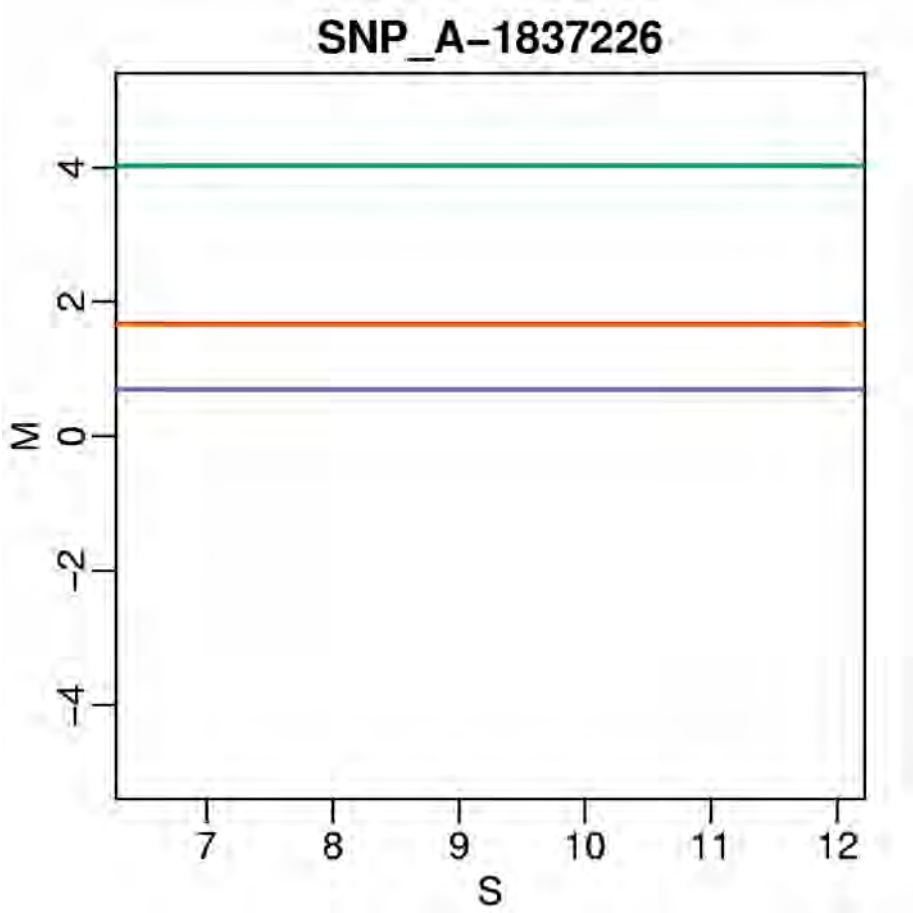
Doh!



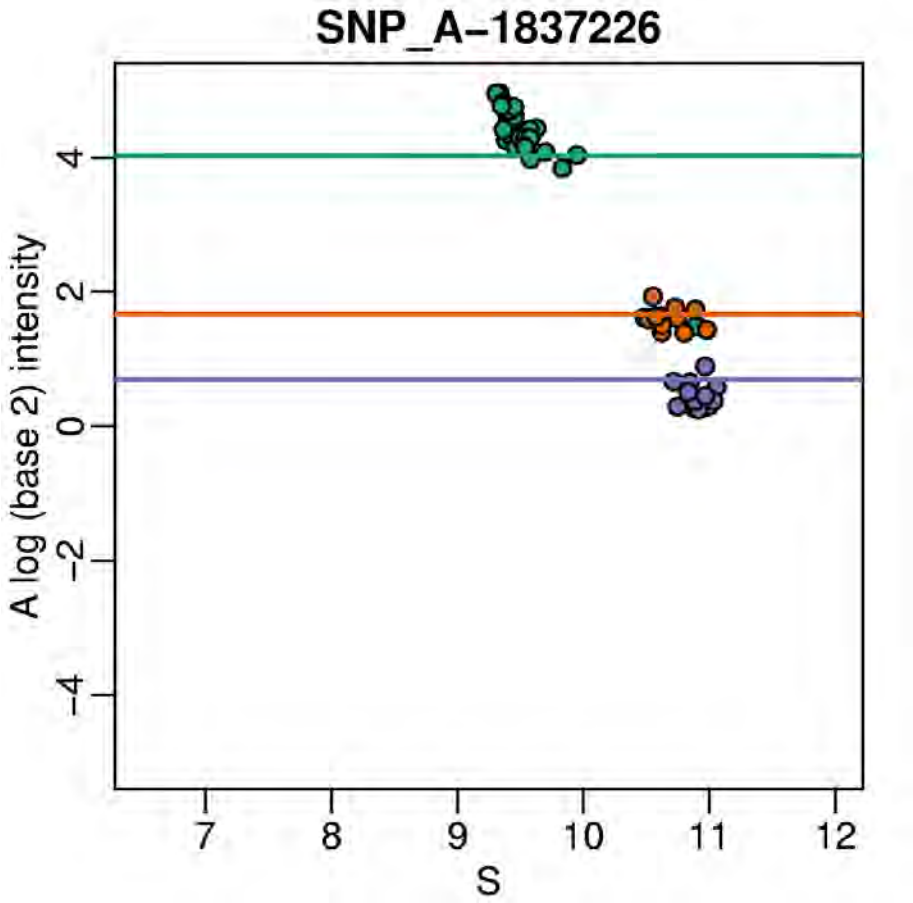
CRLMM uses training data



Does it extrapolate cross study?

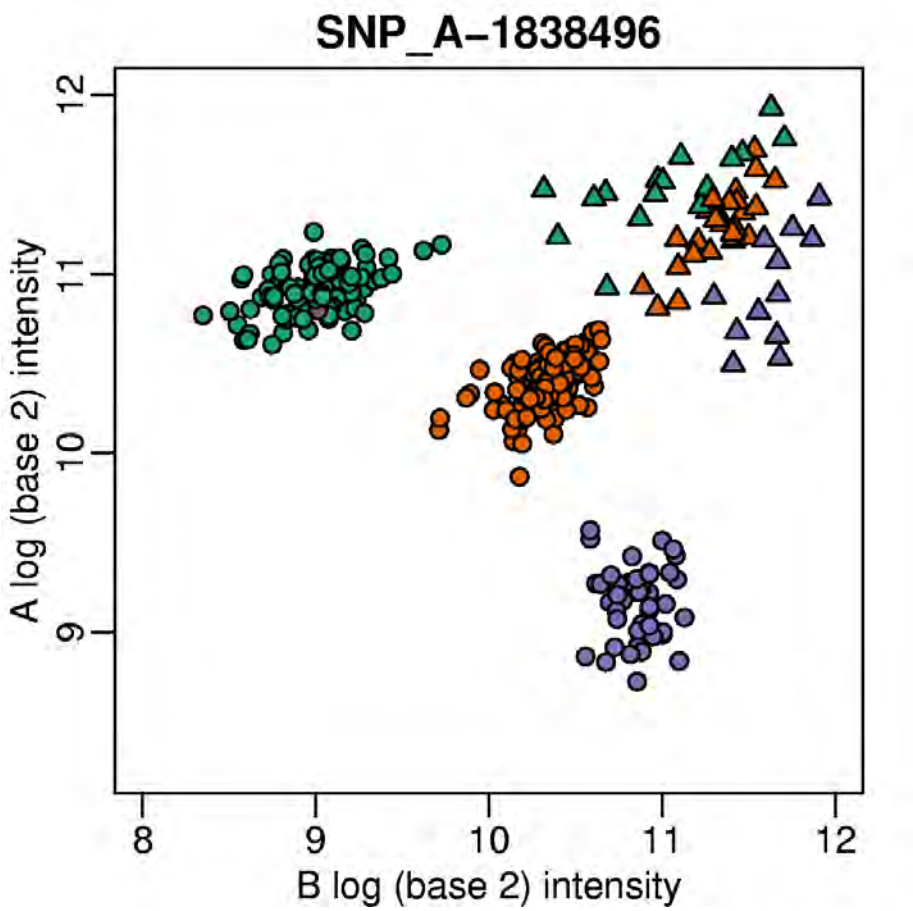


For the most part, yes!

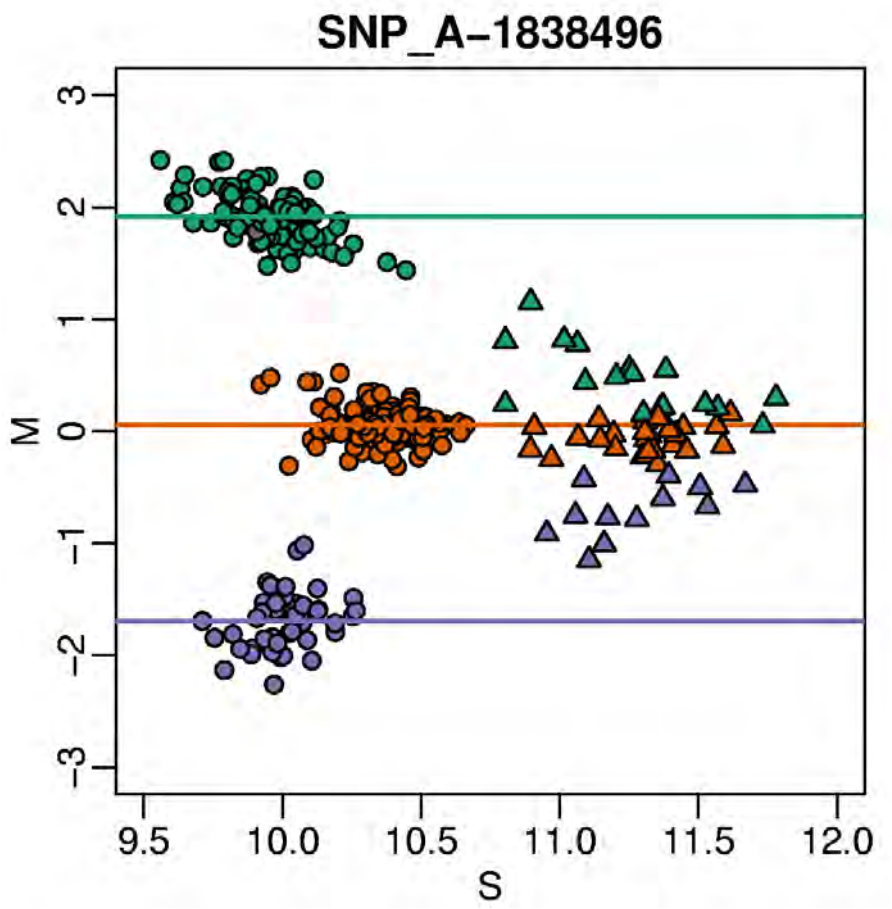


Why is Birdseed having problems across batch?

Batch effec to strong



Again, *M* more stable

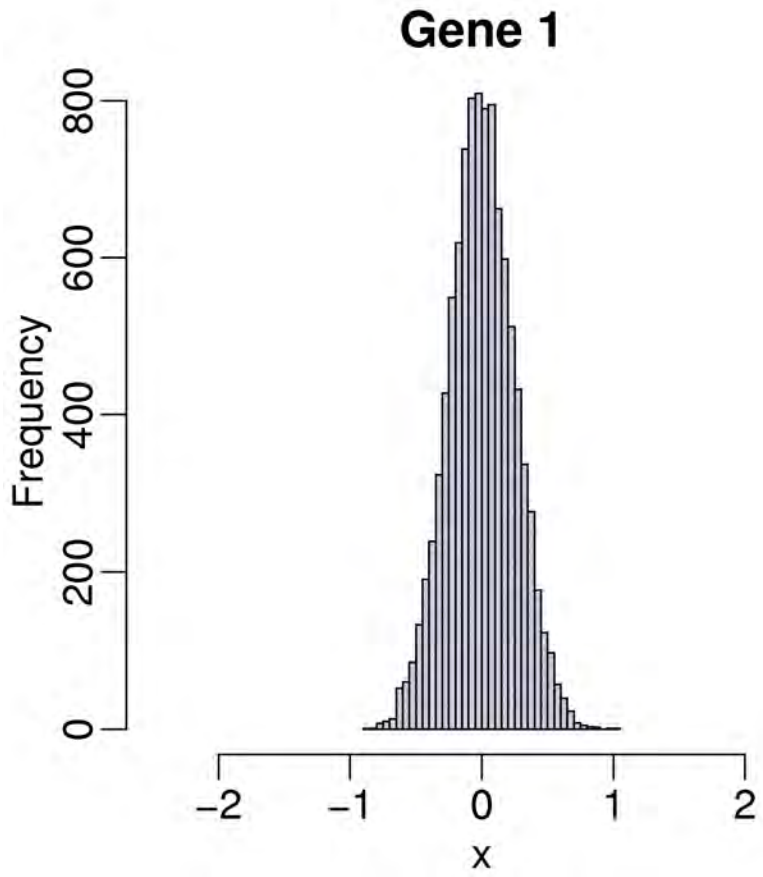


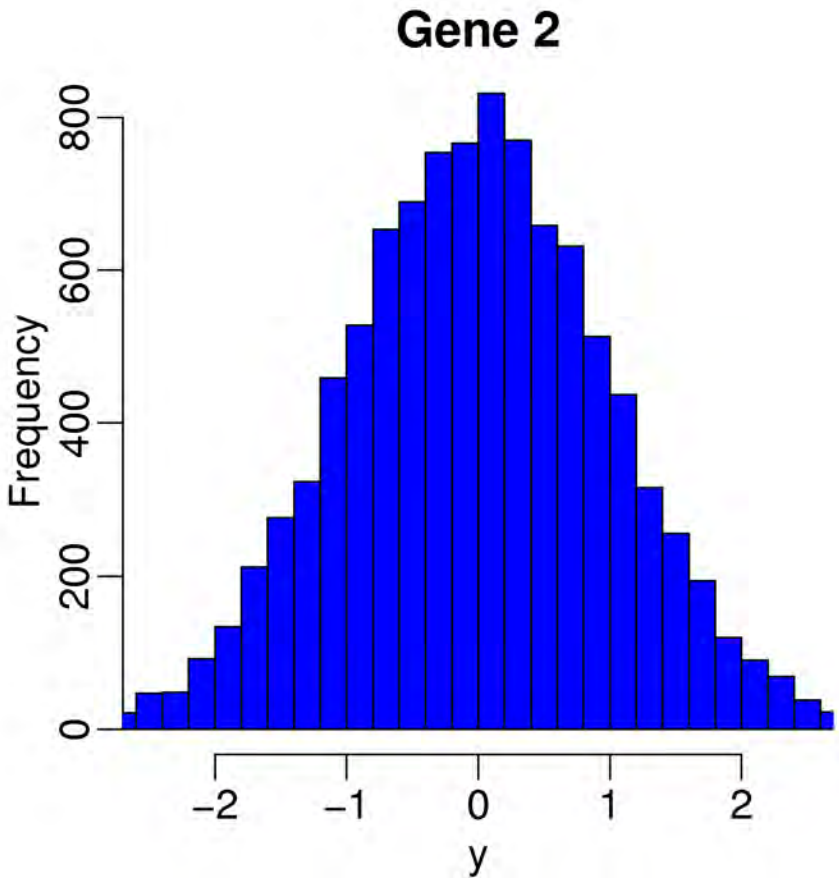
But we need to correct

The End

Supplementary Figures

Repeated Experiment





Why logs

- The intensity distribution has a fat right tail
- Log of ratios are symmetric around 0:
  - Average of *1/10* and *10* is about *5*
  - Average of *log(1/10)* and *10* is *0*
  - *Averaging ratios is almost always a bad idea!*

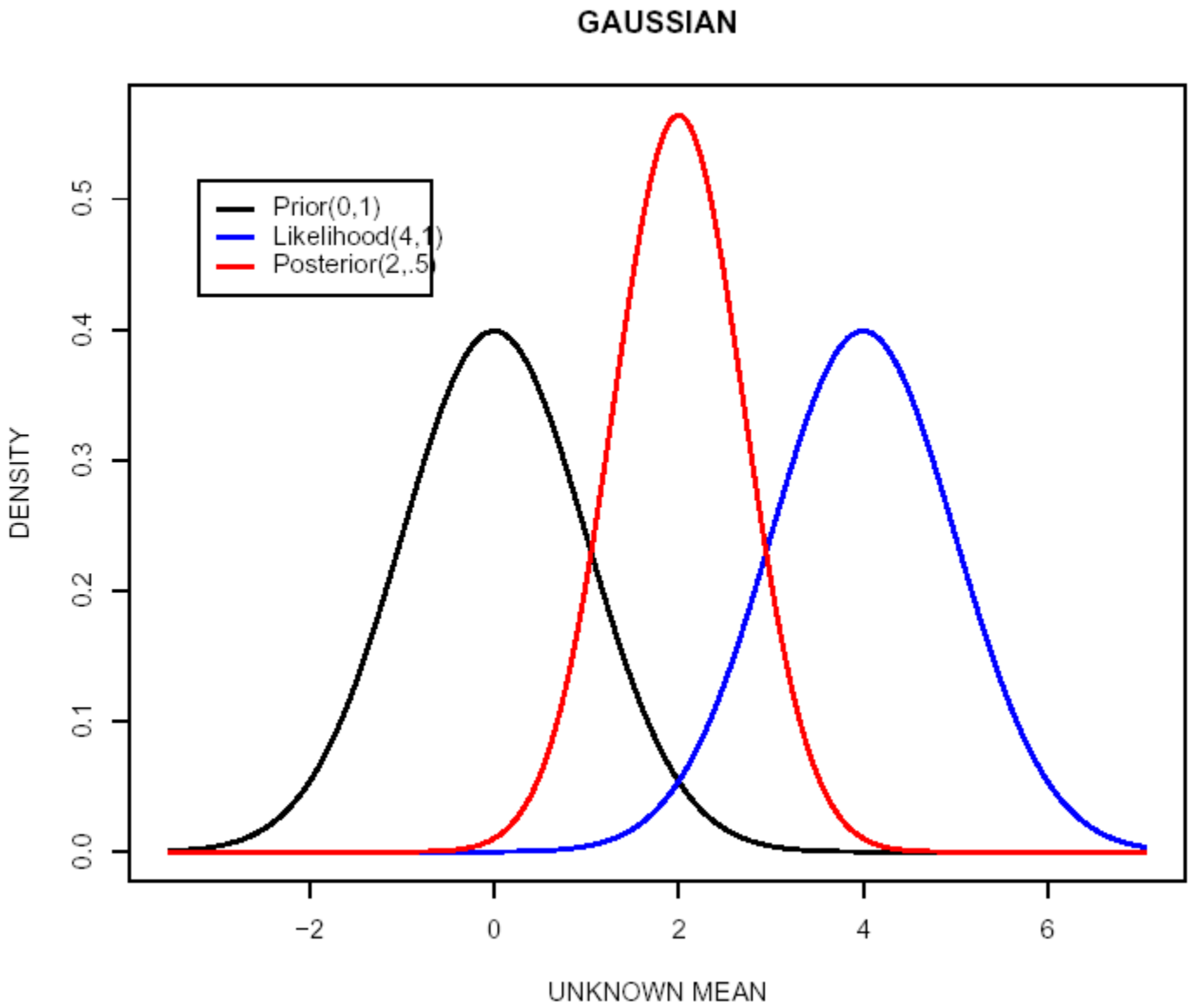
Facts you must remember:

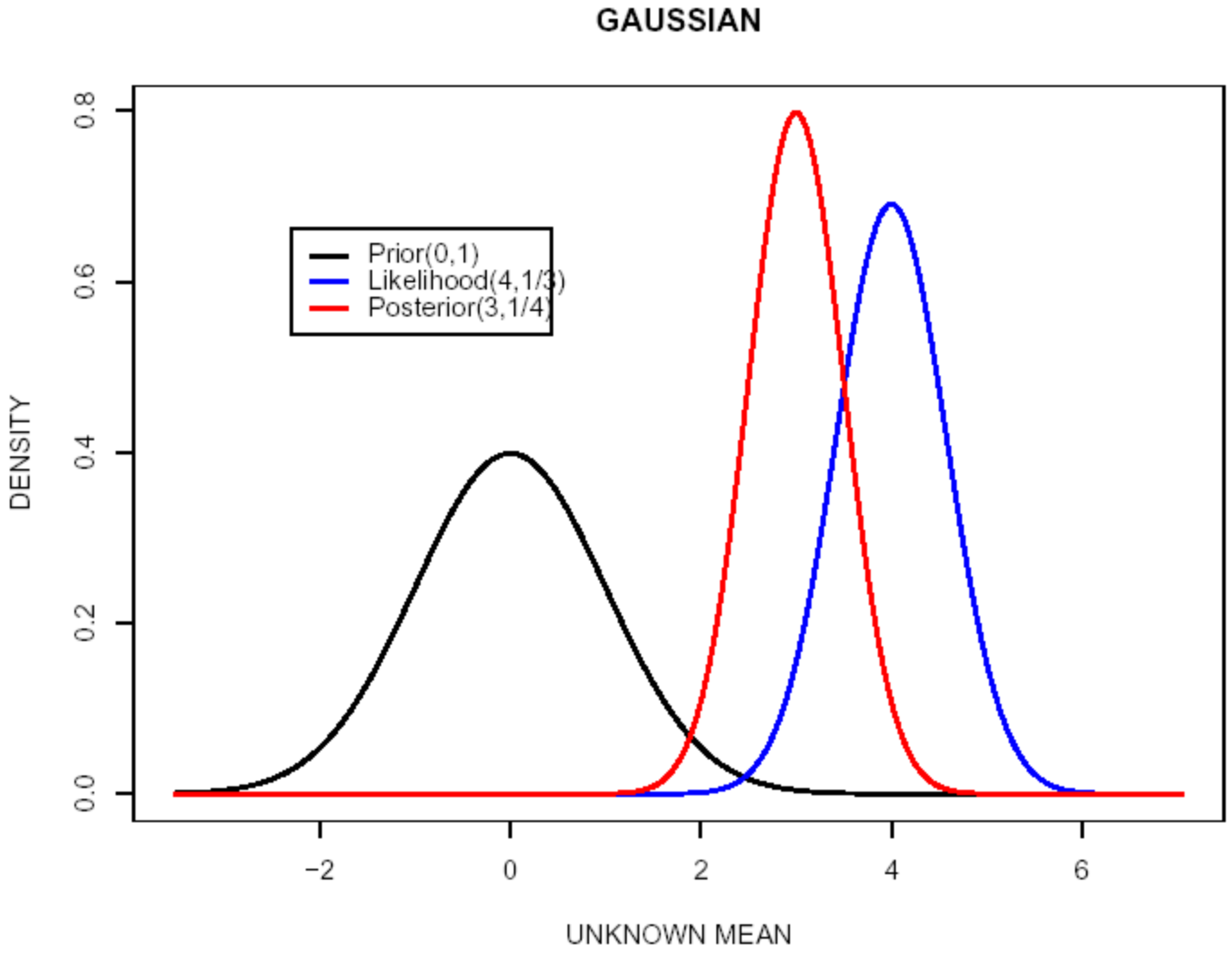
$\log(1) = 0$

$\log(XY) = \log(X) + \log(Y)$

$\log(Y/X) = \log(Y) - \log(X)$

$\log(\sqrt{X}) = 1/2 \log(X)$


GAUSSIAN
Prior(0,1)
Likelihood(4,1)
Posterior(2,.5)
DENSITY
0.0
0.1
0.2
0.3
0.4
0.5
−2
0
2
4
6
UNKNOWN MEAN


GAUSSIAN
Prior(0,1)
Likelihood(4,1/3)
Posterior(3,1/4)
DENSITY
0.0
0.2
0.4
0.6
0.8
−2
0
2
4
6
UNKNOWN MEAN

Posterior Statistics

Posterior variance estimators

$$\tilde{s}_g^2 = \frac{s_g^2 d_g + s_0^2 d_0}{d_g + d_0}$$

Moderated t-statistics

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{c_{gj}}}$$

Eliminates large t-statistics merely from very small *s*

Marginal Distributions

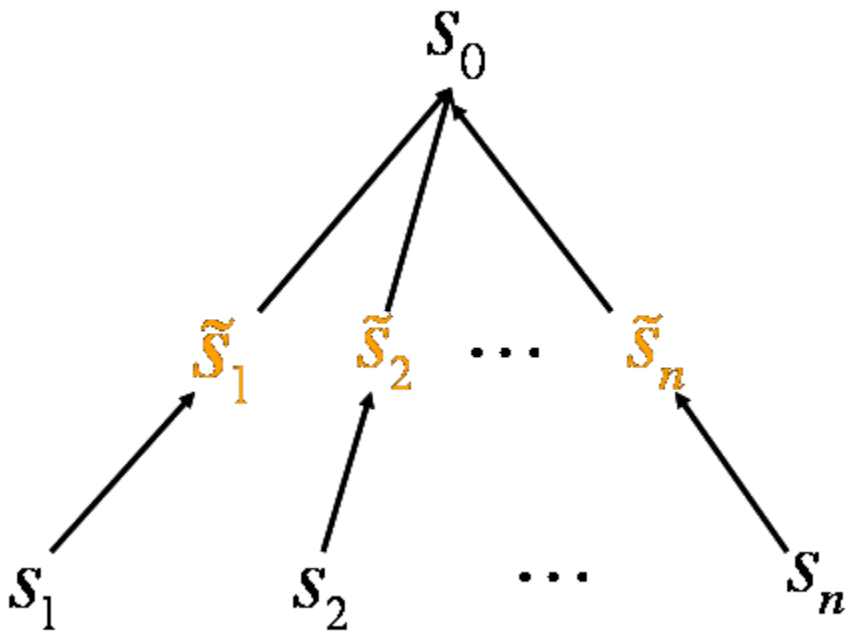
The marginal distributions of the sample variances and moderated t-statistics are mutually independent

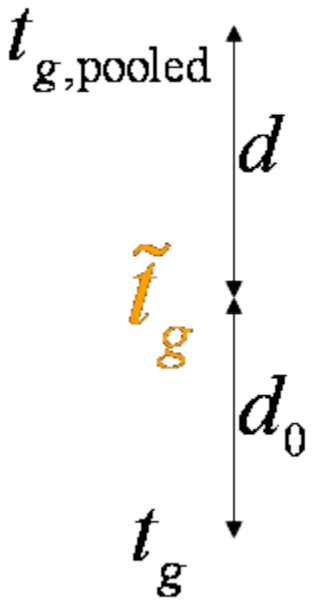
$$s_g^2 \sim s_0^2 F_{d,d_0}$$

$$\tilde{t}_g \sim \begin{cases} t_{d_0+d} & \text{with prob } 1-p \\ \sqrt{1+c_0/c}\, t_{d_0+d} & \text{with prob } p \end{cases}$$

Degrees of freedom add!

Shrinkage of Standard Deviations





The data decides whether $\tilde{t}_g$

should be closer to $t_{g,pooled}$ or to t_g

Posterior Odds

Posterior probability of differential expression for any gene is

$$\frac{p(\beta \neq 0 \mid \hat{\beta}, s^2)}{p(\beta = 0 \mid \hat{\beta}, s^2)} = \frac{p}{1-p}\left(\frac{c}{c+c_0}\right)^{1/2}\left\{\frac{\tilde{t}^2 + d + d_0}{\tilde{t}^2 \dfrac{c}{c-c_0} + d + d_0}\right\}^{\frac{1+d+d_0}{2}}$$

Monotonic function of $\tilde{t}^2$ for constant *d*

Reparametrization of Lönnstedt and Speed 2002

Multiple Hypothesis Testing

- What happens if we call all genes significant with p-values ≤ 0.05, for example?

	Called Significant	Not Called Significant	Total
Null True	*V*	$m_0 - V$	m_0
Altern.True	*S*	$m_1 - S$	m_1
Total	*R*	$m - R$	m

Error Rates

- Per comparison error rate (PCER): the expected value of the number of Type I errors over the number of hypotheses

  $PCER = E(V)/m$

- Per family error rate (PFER): the expected number of Type I errors

  $PFER = E(V)$

- Family-wise error rate: the probability of at least one Type I error

  $FEWR = Pr(V \geq 1)$

- False discovery rate (FDR) rate that false discoveries occur

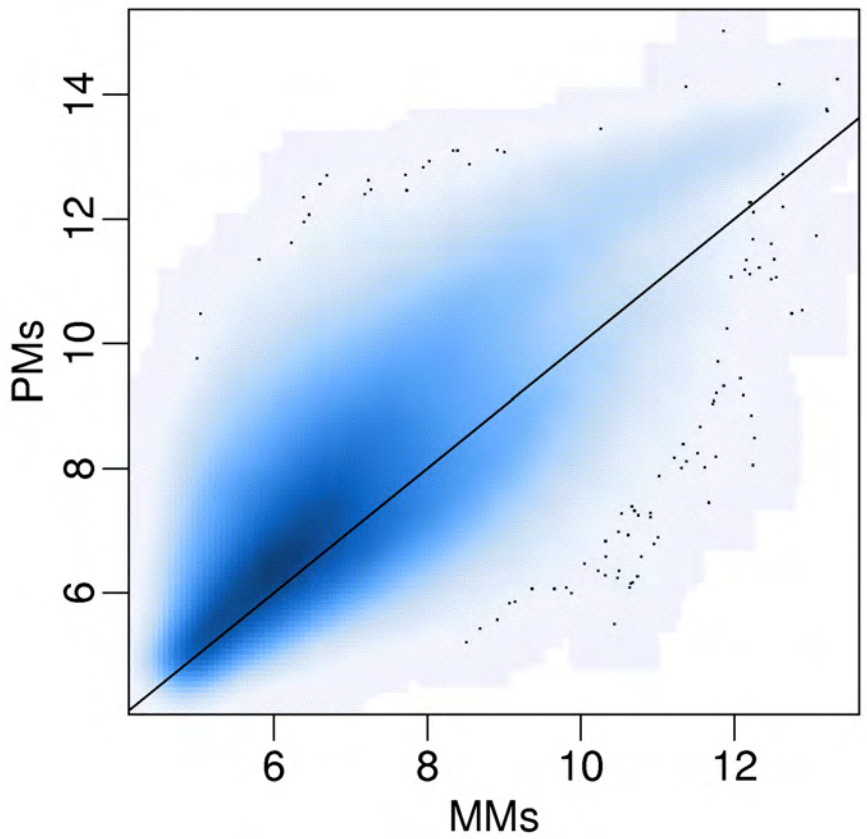
  $FDR = E(V/R; R>0) = E(V/R \mid R>0)Pr(R>0)$

- Positive false discovery rate (pFDR): rate that discoveries are false

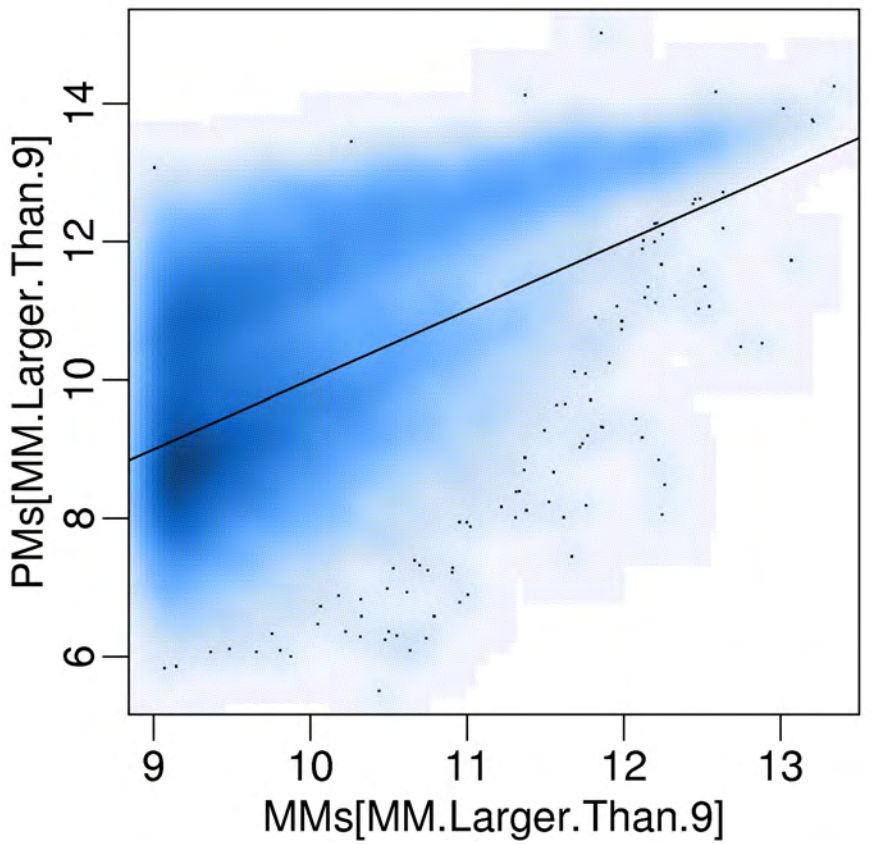
  $pFDR = E(V/R \mid R>0)$

- More later.

Why not subtract MM?



Why not subtract MM?



General Model

NSB SB

$$PM_{gij} = O_i^{PM} + \exp(h_i(\alpha_j^{PM}) + b_{gj}^{PM} + \varepsilon_{gij}^{PM}) + \exp(f_i(\alpha_j) + \theta_{gi} + \xi_{gij})$$

$$MM_{gij} = O_i^{MM} + \exp(h_i(\alpha_j^{MM}) + b_{gj}^{MM} + \varepsilon_{gij}^{MM})$$

We can calculate: $E[\theta_{gi} | PM_{gij}, MM_{gij}]$

Alternative background adjustment

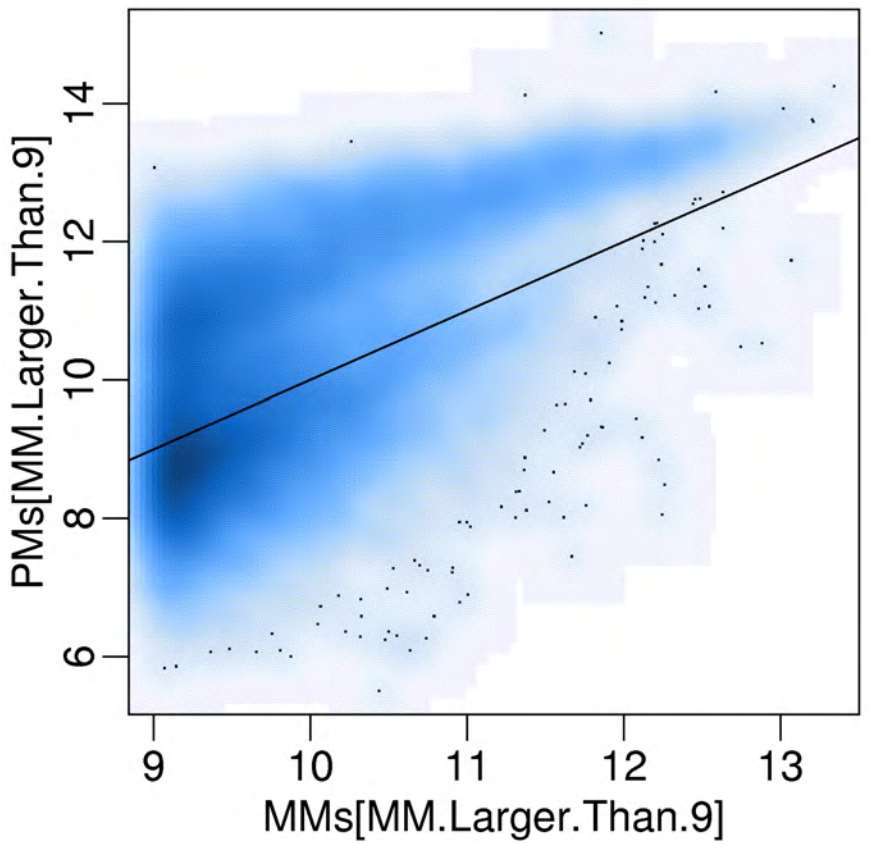
- Use this stochastic model
- Minimize the MSE:

$$E\left[\left\{\log\left(\frac{\tilde{s}}{s}\right)\right\}^2 | S > 0, PM, MM\right]$$

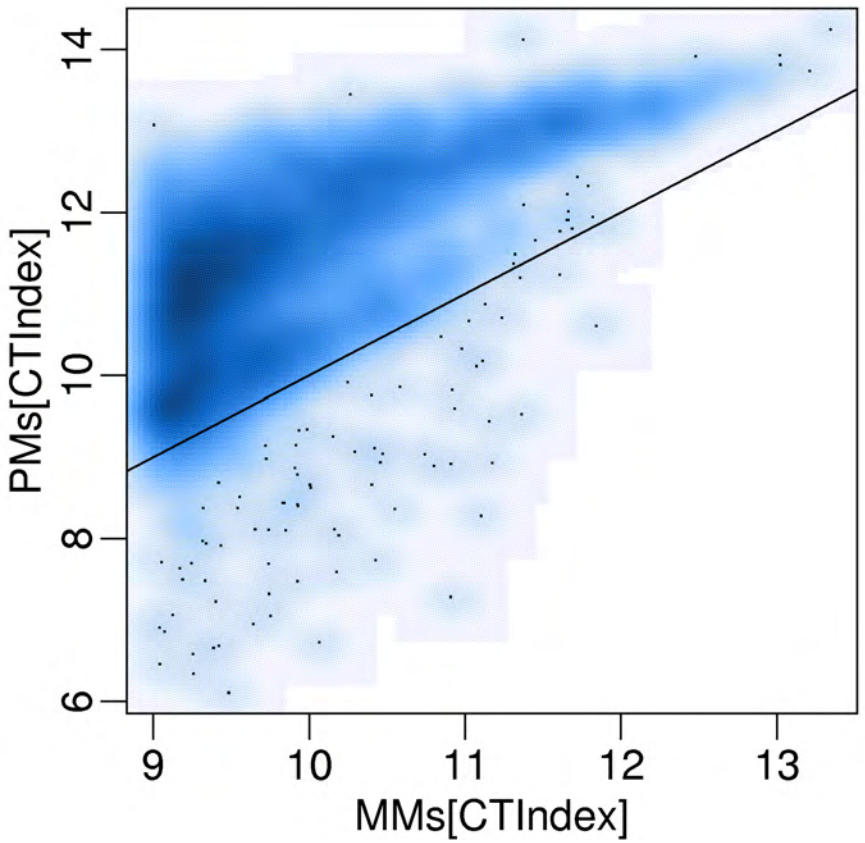
- To do this we need to specify distributions for the different components
- Notice this is probe-specific so we need to borrow strength

*These parametric distributions were chosen to provide a *closed form* solution

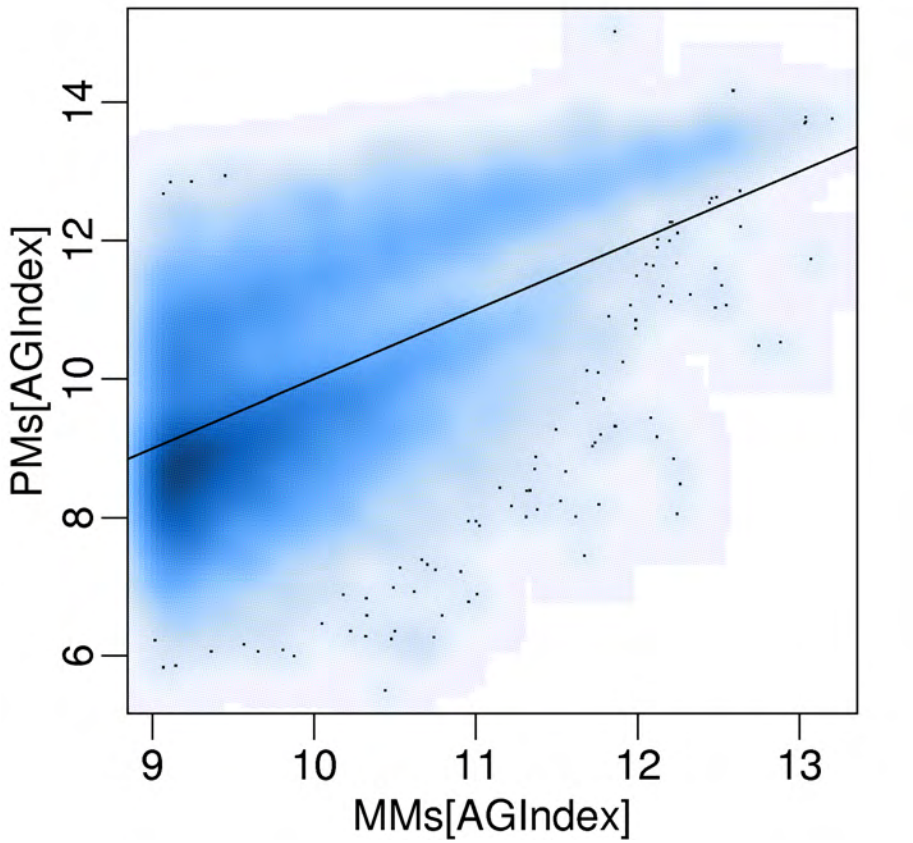
Explains Bimodality



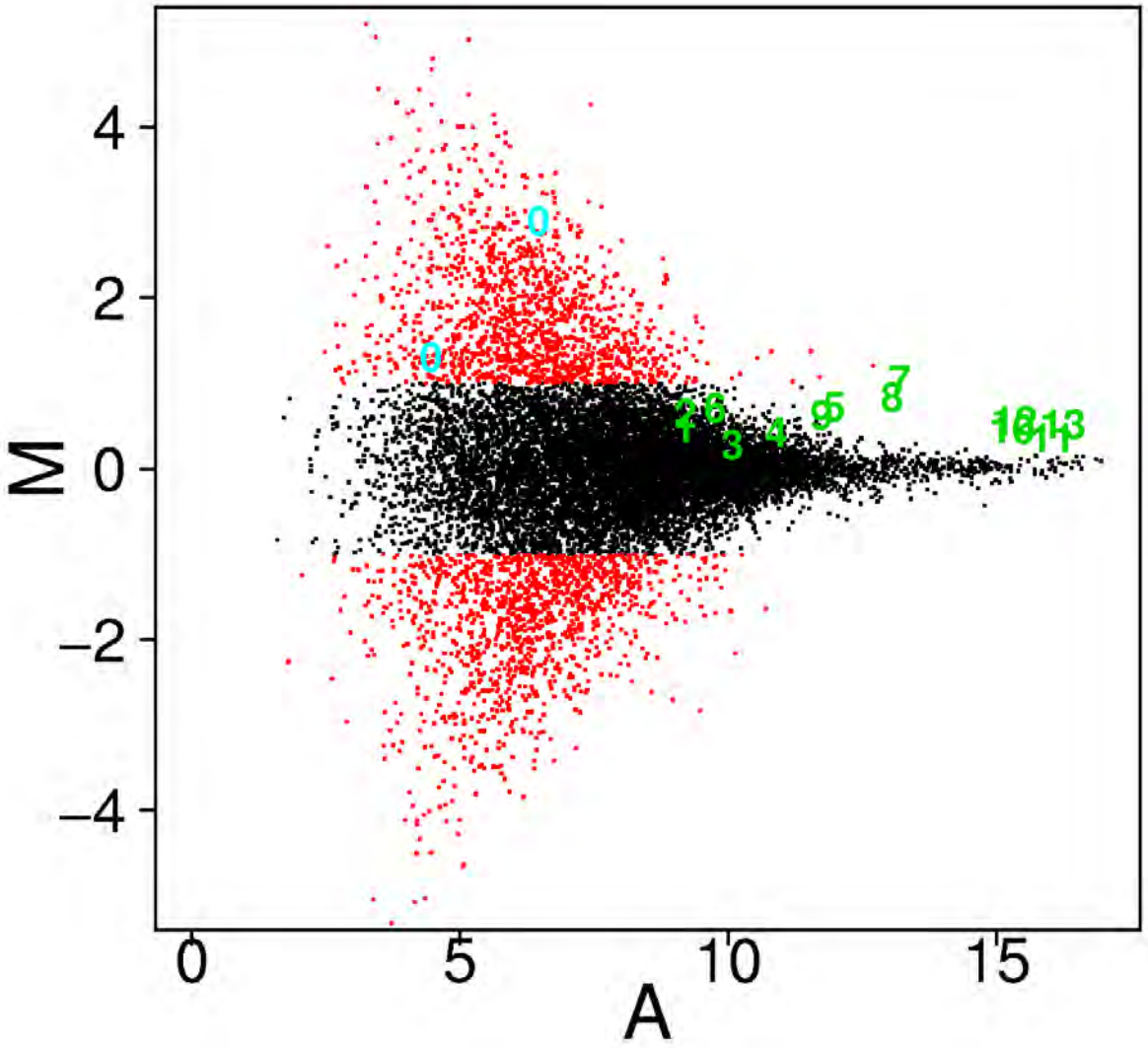
C,T in the middle



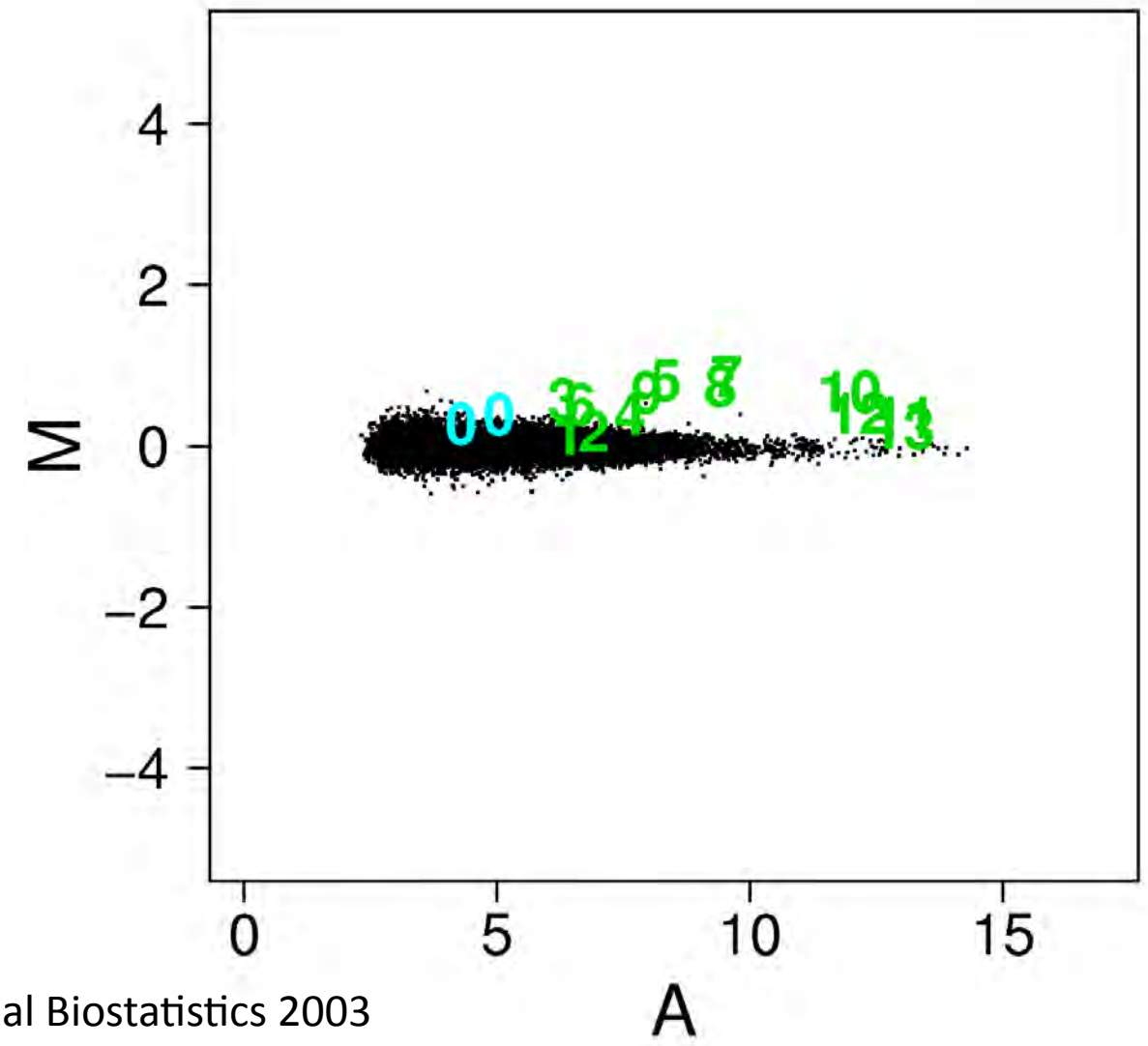
A,G in the middle



Why So Much Noise?



Signal Recovered



Irizarry et al Biostatistics 2003

Summarization (Median Polish)

$$Y_{ij} = m_i + a_j + e_{ij}$$

Y_{ij} – normalized probe value for jth probe on the ith gene chip
m_i – expression value on the ith gene chip
a_j – probe affinity effect fo the jth probe
e_{ij} – random noise

Summarization (Median Polish)

		Probe					
GeneChip		**1**	**2**	**3**	**4**	**5**	**Row Medians**
	1	18	11	8	21	4	11
	2	13	7	5	16	7	7
	3	15	6	7	16	6	7
	4	19	15	12	18	5	15

Subtract the row median from each value.

		Probe				
GeneChip		**1**	**2**	**3**	**4**	**5**
	1	7	0	-3	10	-7
	2	6	0	-2	9	0
	3	8	-1	0	9	-1
	4	4	0	-3	3	-10
Column Medians		6.5	0	-2.5	9	-4

Summarization (Median Polish)

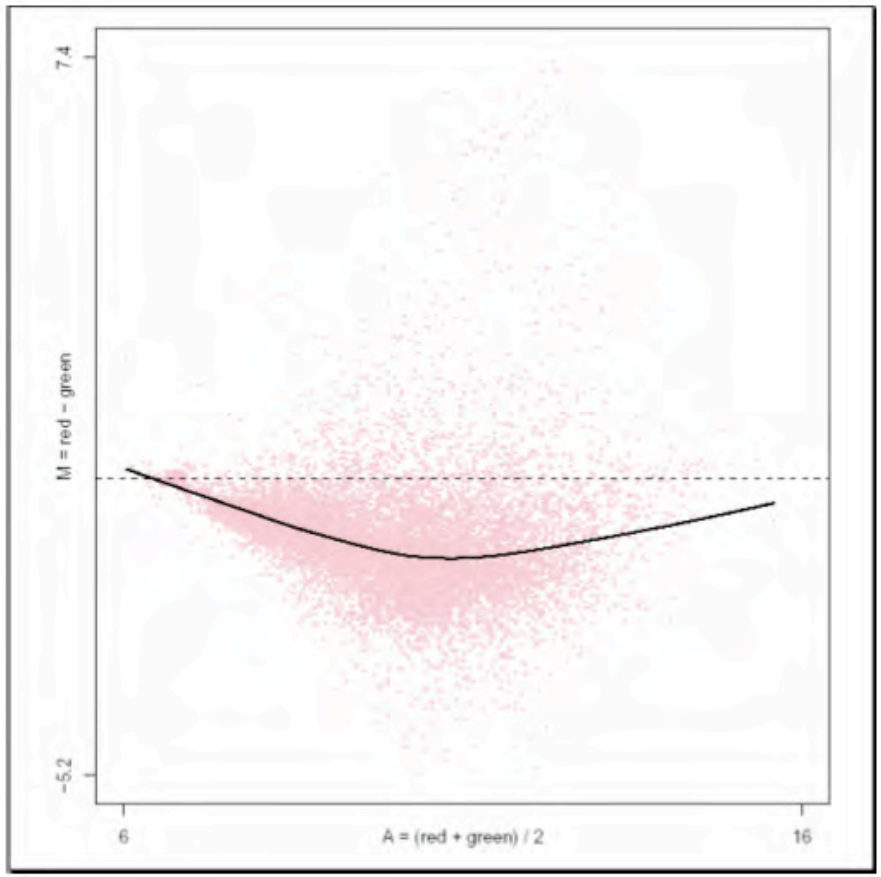
		Probe					
GeneChip		**1**	**2**	**3**	**4**	**5**	**Row Medians**
	1	0.5	0	-1.25	1.75	-2.25	0
	2	-0.5	0	-0.25	0.75	4.75	0
	3	0	-2.5	0.25	-0.75	2.25	0
	4	0	2.5	1.25	-2.75	-2.75	0

Note that row medians are all zero (above) and column medians are all zero (below).

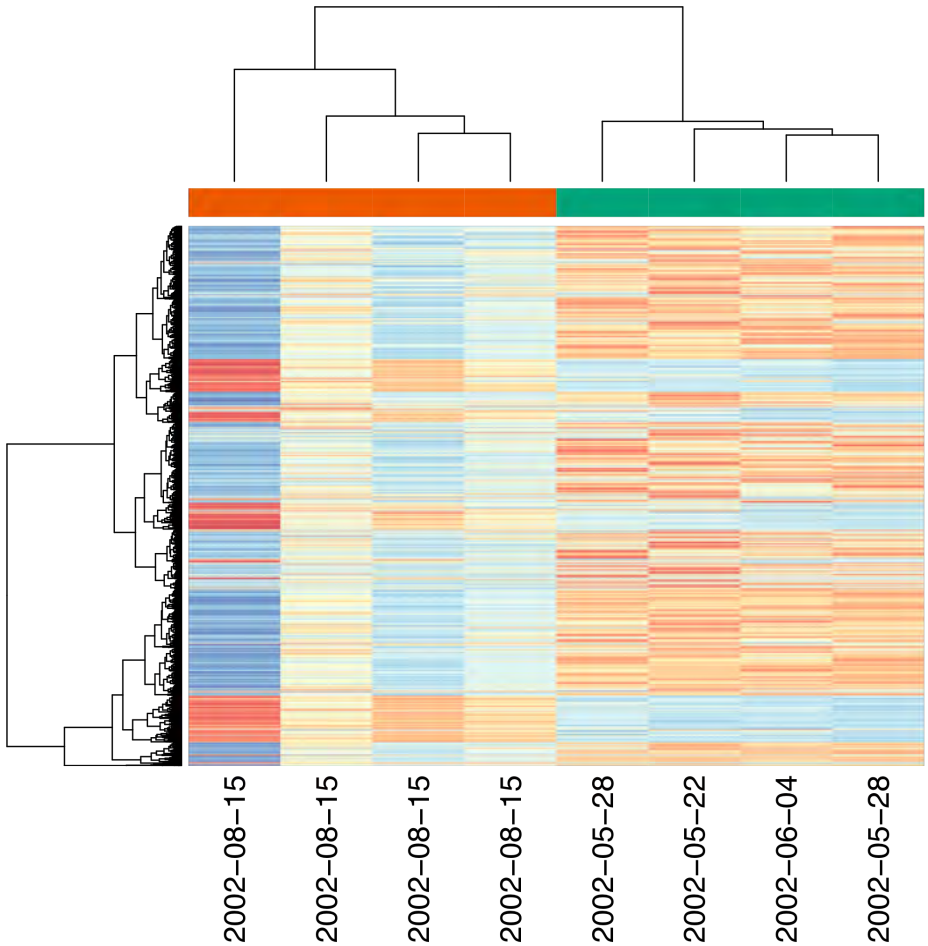
		Probe				
GeneChip		**1**	**2**	**3**	**4**	**5**
	1	0.5	0	-1.25	1.75	-2.25
	2	-0.5	0	-0.25	0.75	4.75
	3	0	-2.5	0.25	-0.75	2.25
	4	0	2.5	1.25	-2.75	-2.75
Column Medians		0	0	0	0	0

Two Channel Arrays

Intensity Dependent Effects



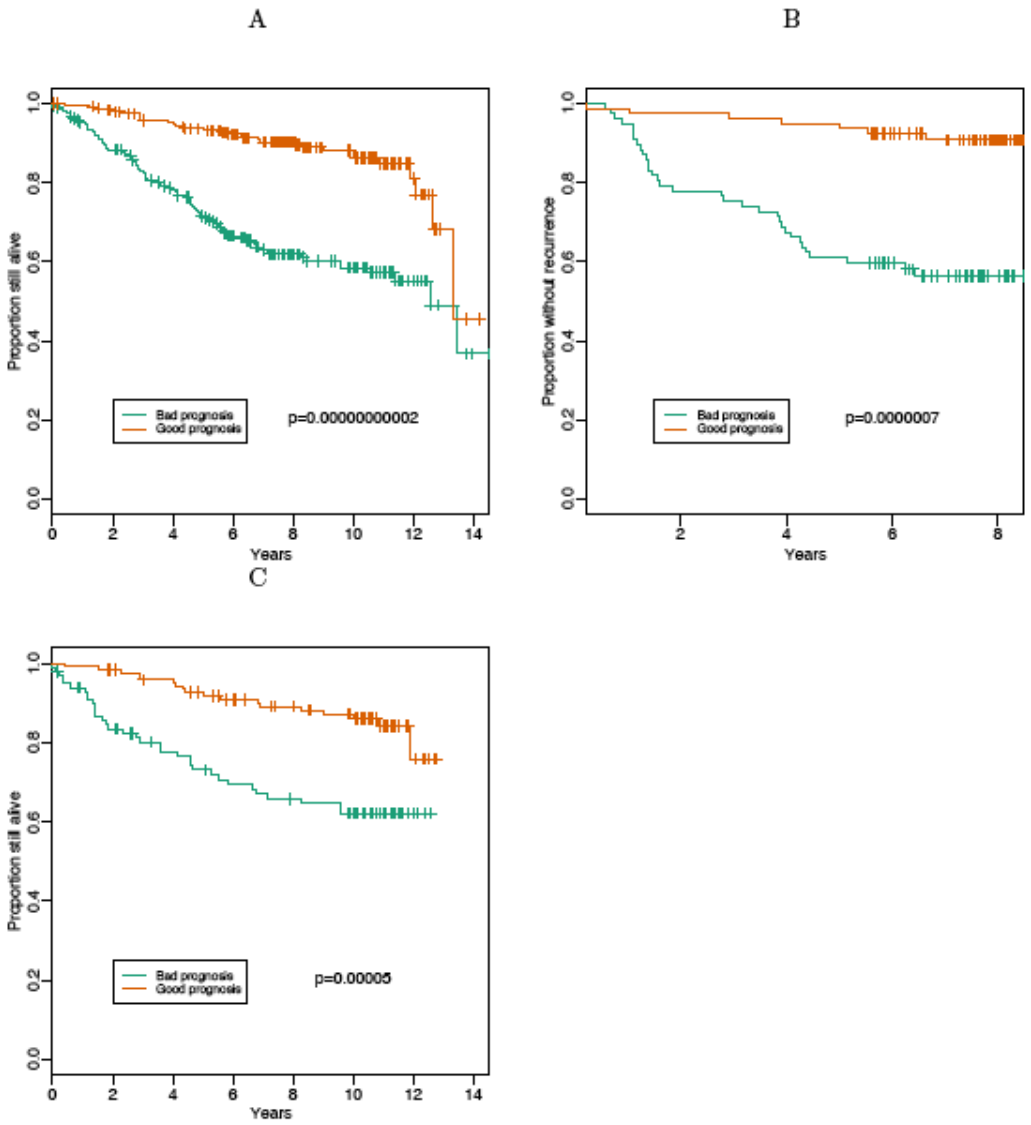
Clustering of normals



Prediction power

Samples	Comparison Type	PAM (% correct)	Bar code (% correct)
Human Normal Tissues	Different tissues	95	98
Mouse Normal Tissues	Different tissues	91	96
Alzheimer's disease	Normal versus disease	60	70
	Normal versus severe disease	83	91
Adenocarcinoma	Three different conditions	83	83
	Normal versus cancer/precursor	91	91
Bladder Cancer	Three different conditions	73	83
	Normal versus cancer	90	96
Renal Cell Carcinoma	Normal versus cancer	94	100

Applied to Breast Cancer Data

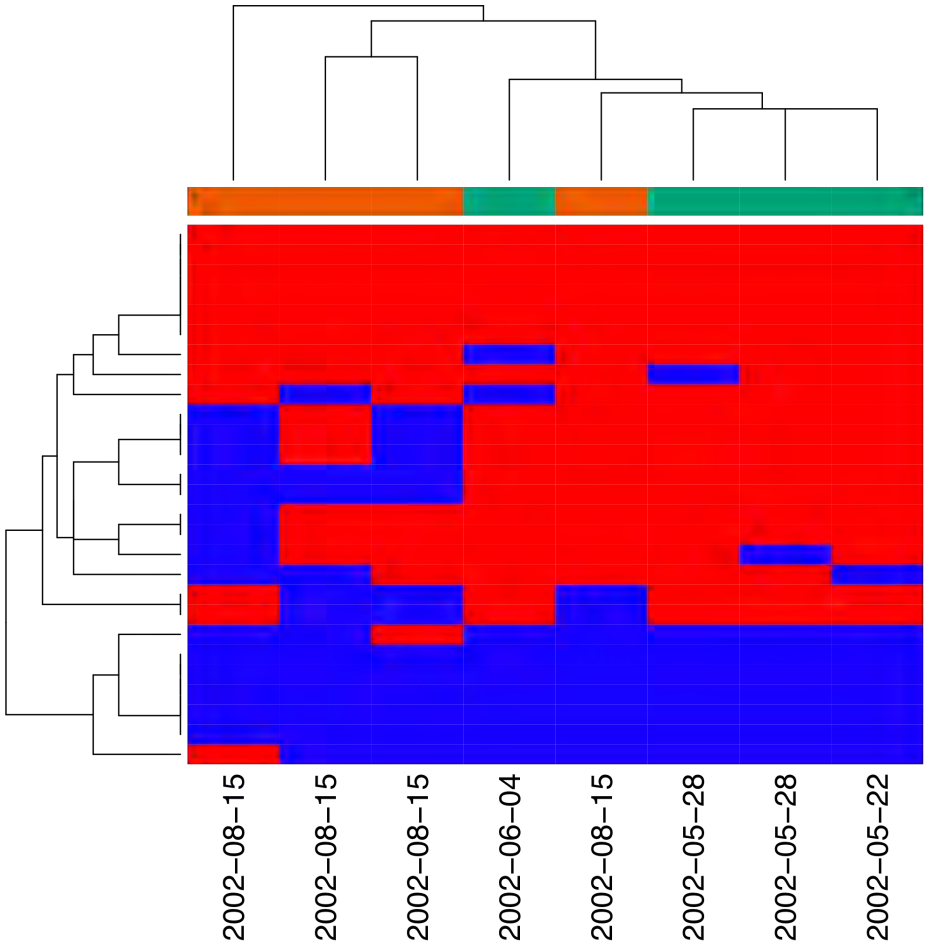


Stumped by 2 datasets

- Potti et al. (2006) Genomic signatures to guide the use of chemotherapeutics. *Nat Med*
  - Coombes, Wang, Baggerly (2007) Microarrays: retracing steps. *Nat Med*
  - A Biostatistics Paper Alleges Potential Harm To Patients In Two Duke Clinical Studies *By Paul Goldberg. The Cancer Letter, Oct 2, 2009.*

- Dyrskjot, L. et al. (2004) Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. *Cancer Res.*

With barcode not as bad



Preprocessing separately

Table 1 | Percentage accuracy comparison on independent data sets

GEO identifier	Data type	PAM (% correct)	Bar code (% correct)
GSE5388	Cortex	100	100
GSE2395	Respiratory system epithelia	0	100
GSE2665	Lymph node/tonsil	35	95
GSE1561	Breast tumor	69	100
GSE2603	Breast tumor	77	90
GSE6344	Kidney: normal versus cancer	100	100

PAM versus the bar code approach in six randomly selected data sets not included in the original database. The data described in **Supplementary Table 1** were used to train the prediction algorithms. GEO, Gene Expression Omnibus.

If you don’t account for batch your results will be wrong

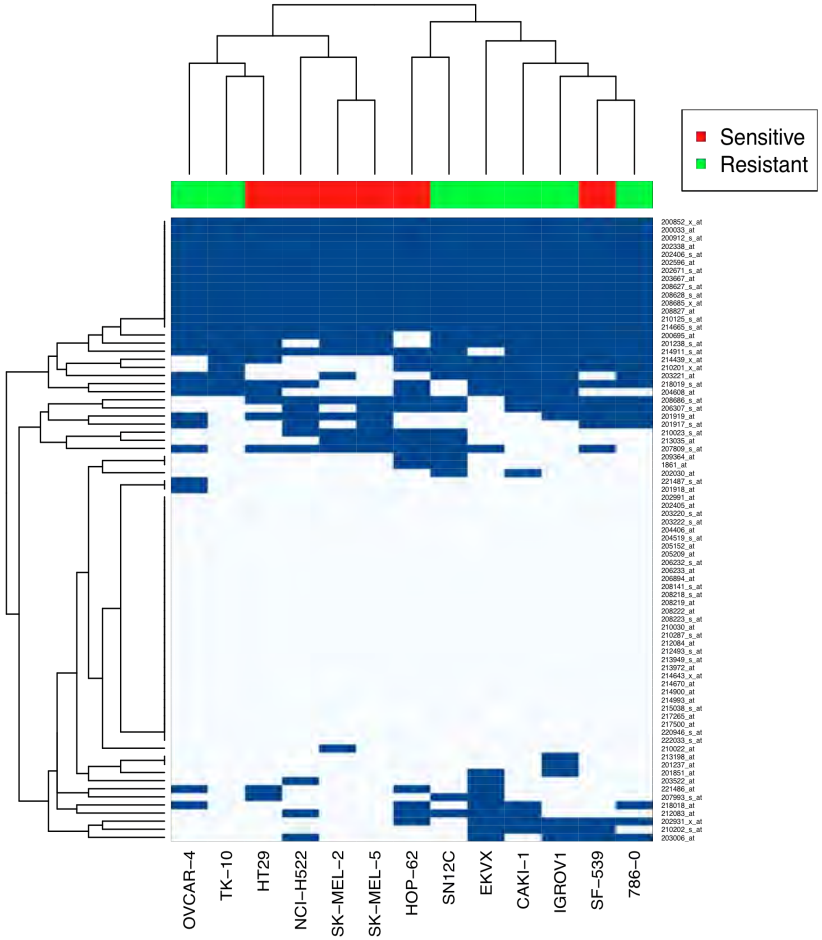
- Leek and Storey (2007) *Plos Genetics*

Acknowledgments

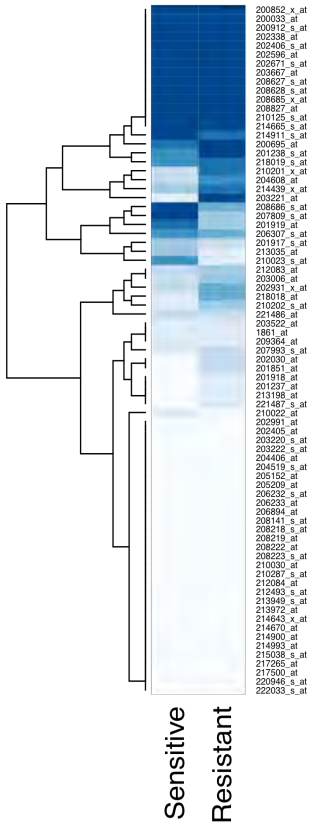
- Matthew McCall, JHU
- Michael Zilliox, Emory
- Jeff Leek, JHU
- Harris Jaffee, JHU
- GEO and Array Express

Thanks!

Potti et al Data



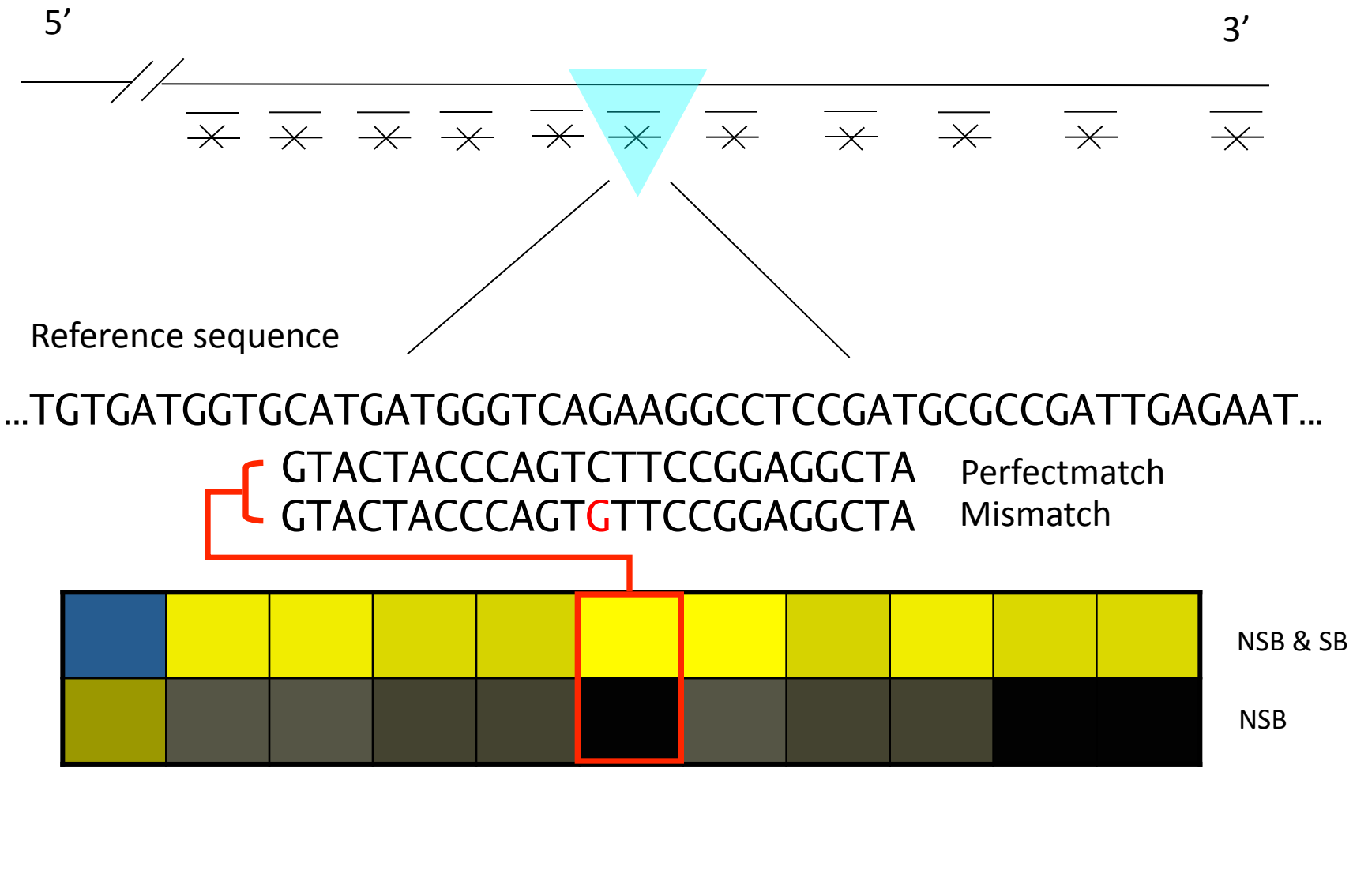
Potti et al barcodes



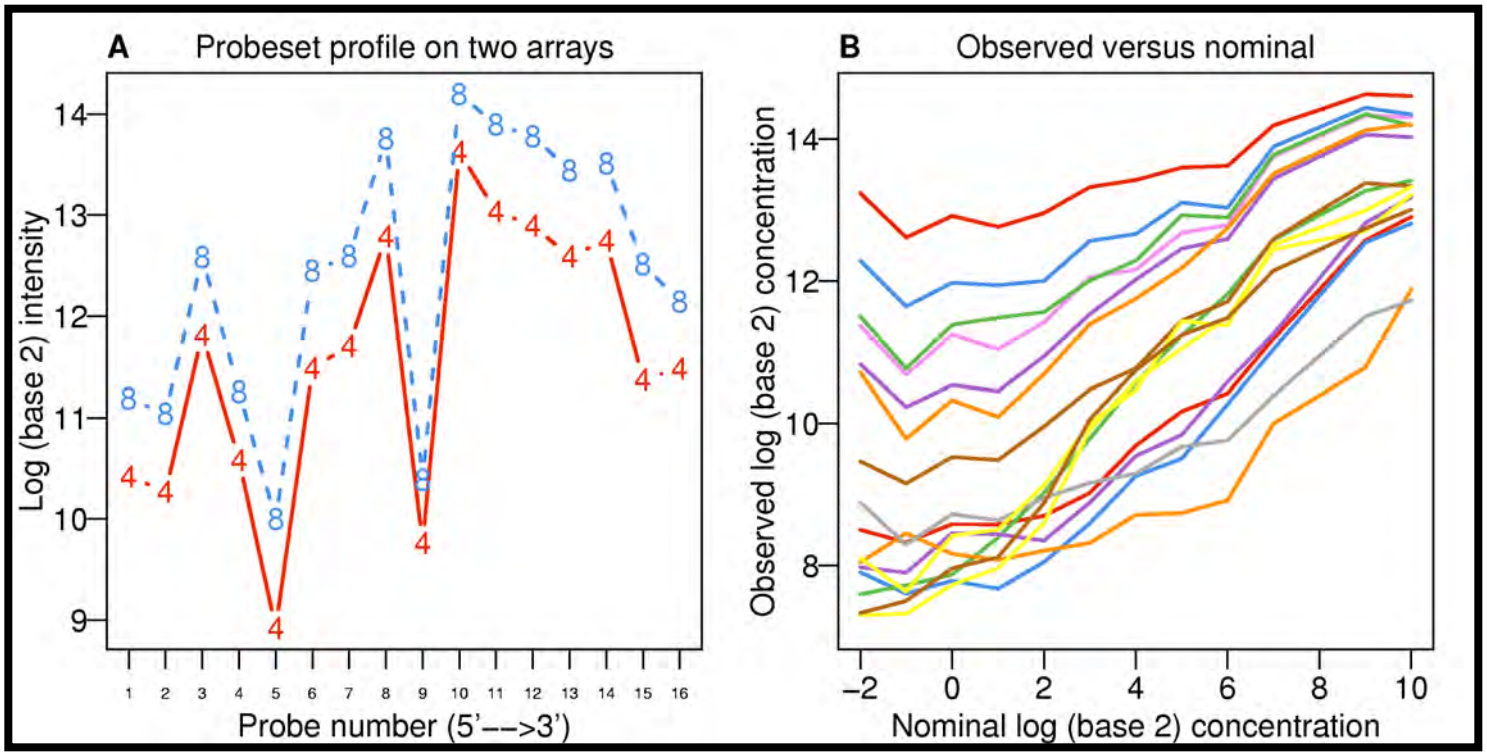
fRMA

Single Array Normalization

Affymetrix GeneChip Design



Probe effect



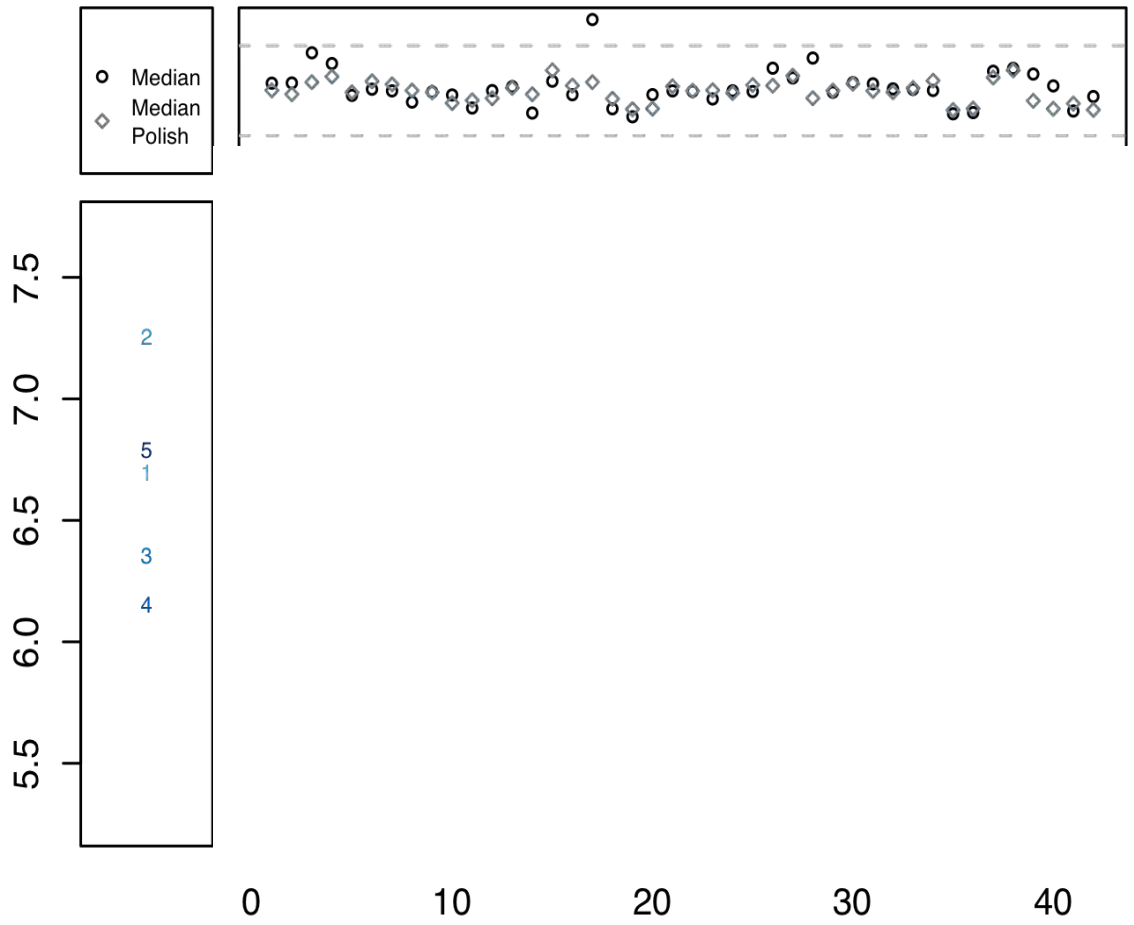
RMA model

- For each gene:

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

- i is array/sample
- j is probe
- We need to estimate a_i robustly

Advantage of multi-array model



Basic Idea

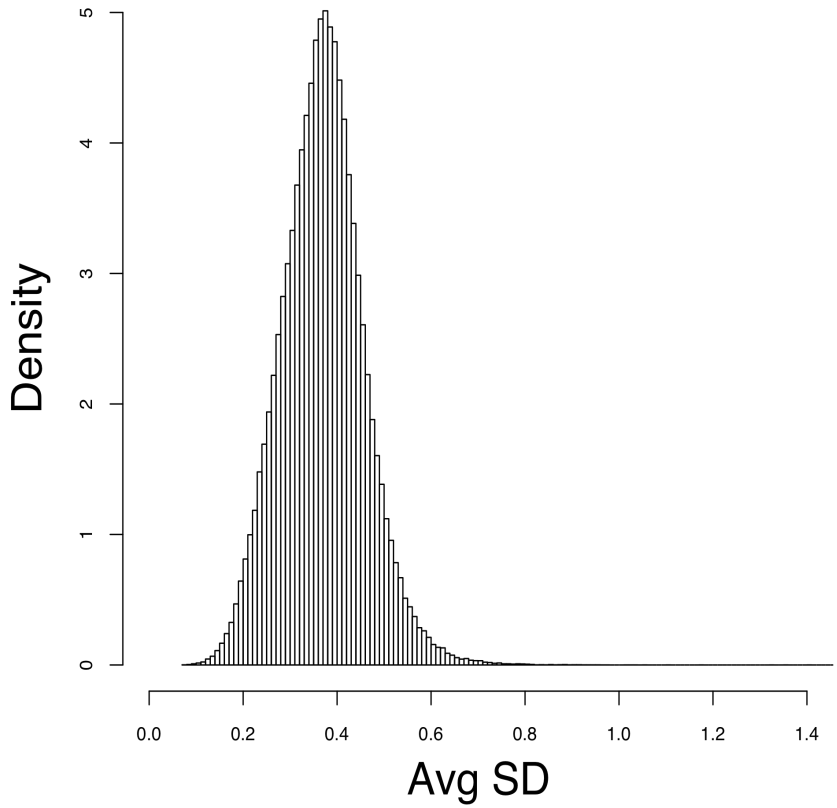
- Create large database with many tissues from many labs
- Background correct
- Quantile normalize and keep the reference distribution
- Fit probe level model and keep probe effect estimates

With new array

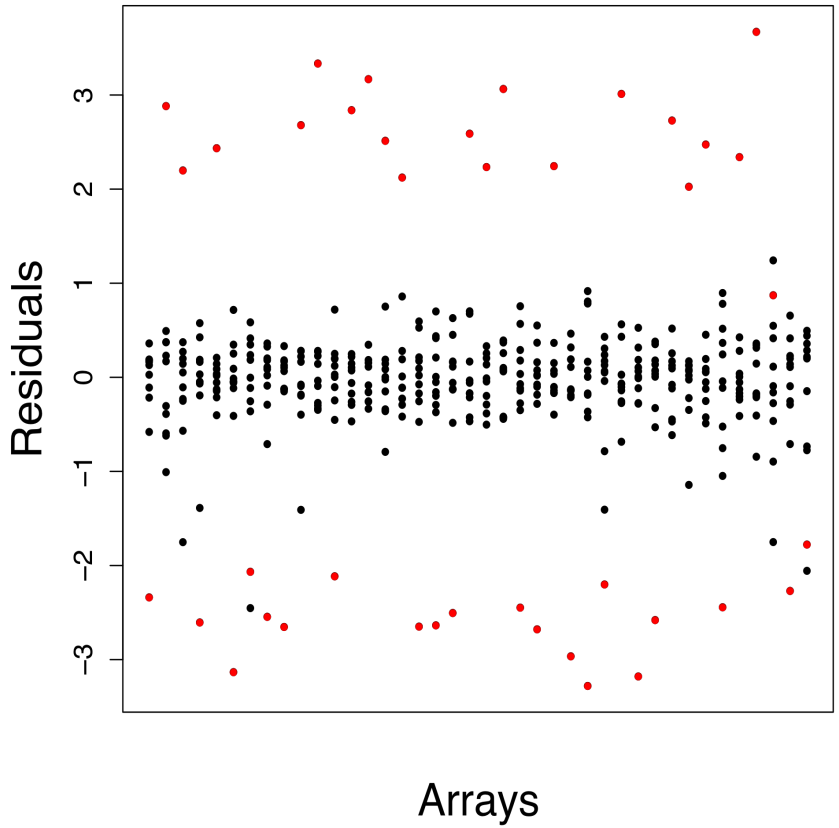
- Background correct
- Normalize to reference distribution
- Subtract saved probe effect
- Take robust median

- But there is more

SD of probe across arrays

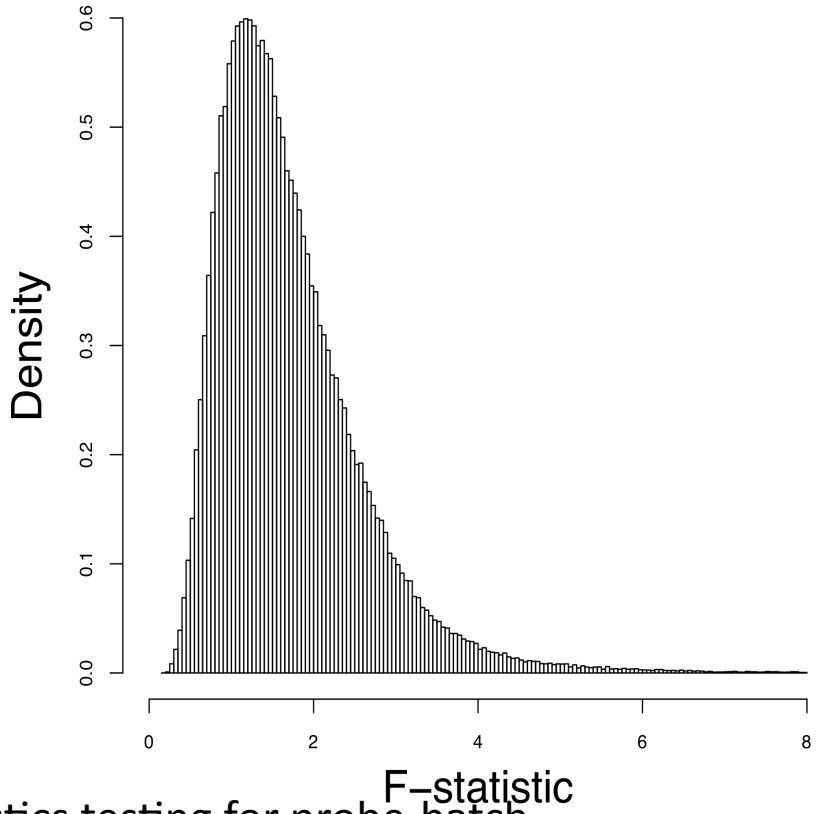


RMA residuals



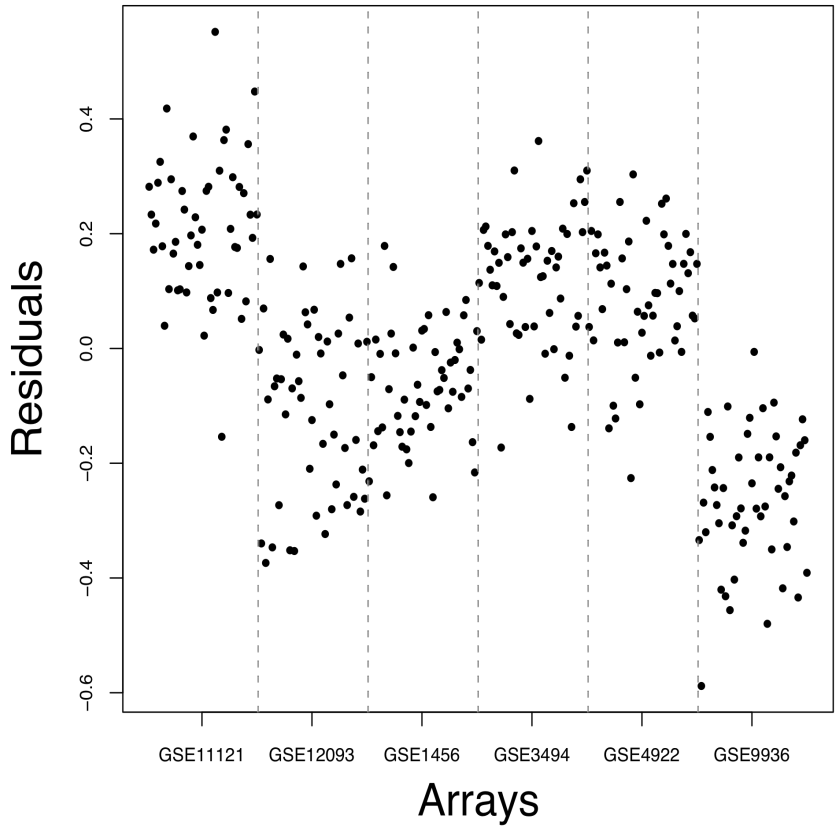
Red points come from highly variable probe

Bigger problem

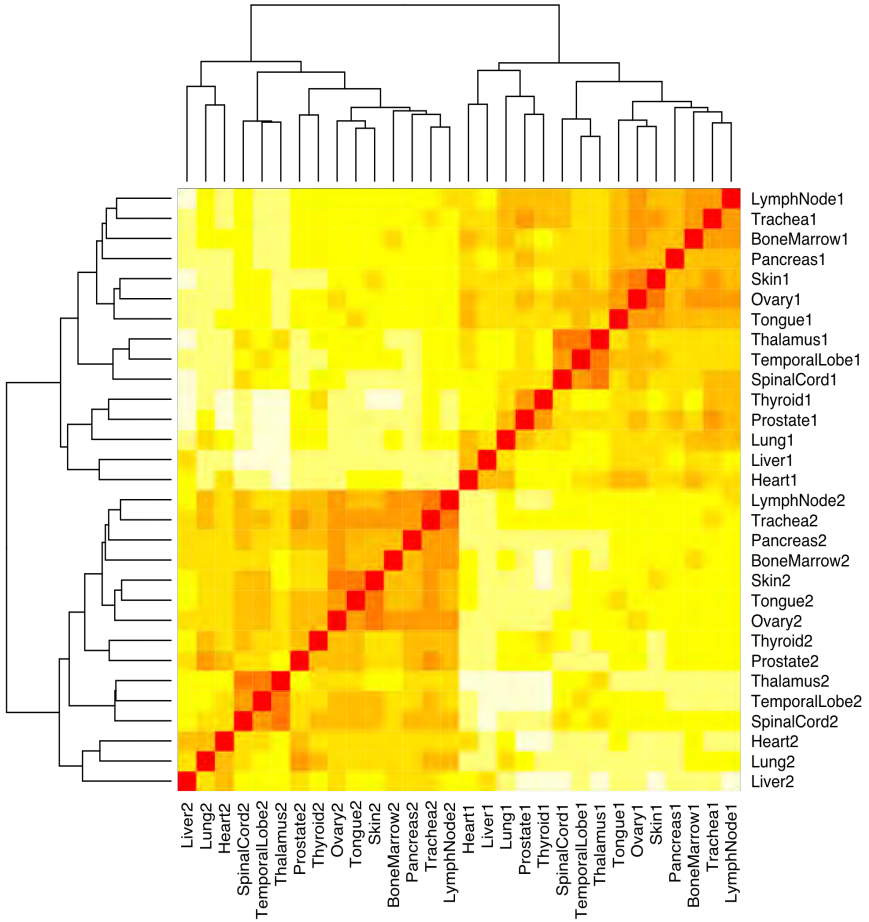


These are f-statistics testing for probe-batch interaction

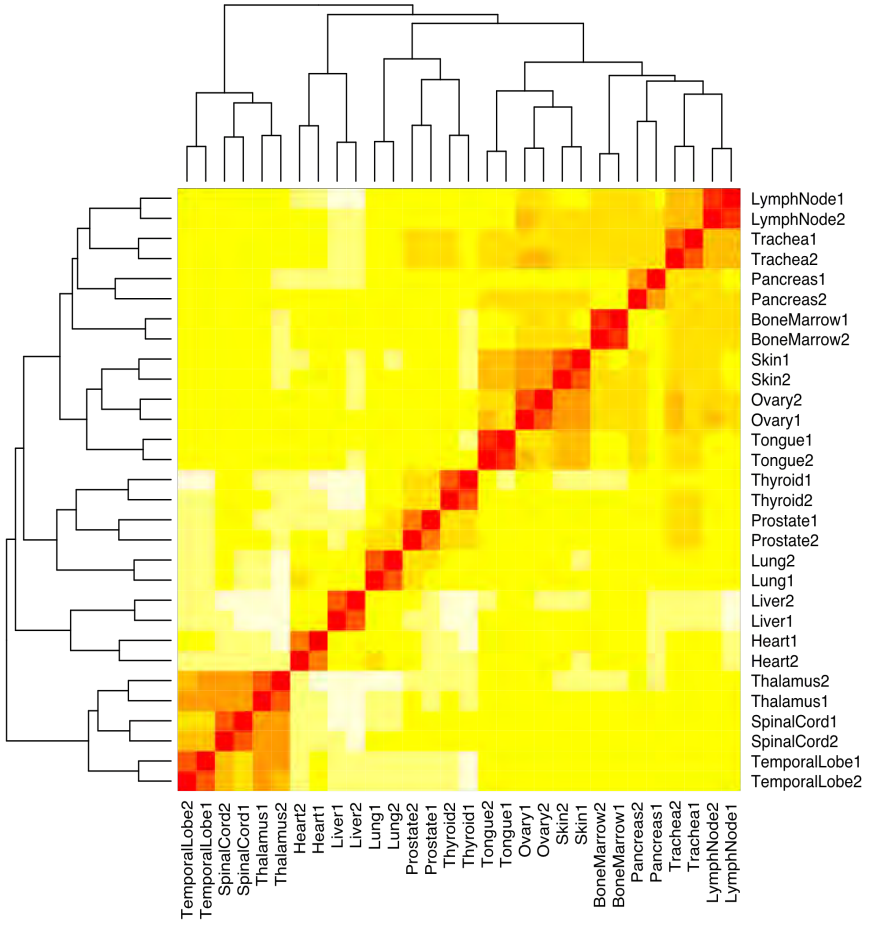
Probe sensitive to batch



RMA normalized in batches



fRMA


LymphNode1
LymphNode2
Trachea1
Trachea2
Pancreas1
Pancreas2
BoneMarrow1
BoneMarrow2
Skin1
Skin2
Ovary2
Ovary1
Tongue1
Tongue2
Thyroid1
Thyroid2
Prostate1
Prostate2
Lung2
Lung1
Liver2
Liver1
Heart1
Heart2
Thalamus2
Thalamus1
SpinalCord1
SpinalCord2
TemporalLobe1
TemporalLobe2
TemporalLobe2
TemporalLobe1
SpinalCord2
SpinalCord1
Thalamus1
Thalamus2
Heart2
Heart1
Liver1
Liver2
Lung1
Lung2
Prostate2
Prostate1
Thyroid2
Thyroid1
Tongue2
Tongue1
Ovary1
Ovary2
Skin2
Skin1
BoneMarrow2
BoneMarrow1
Pancreas2
Pancreas1
Trachea2
Trachea1
LymphNode2
LymphNode1