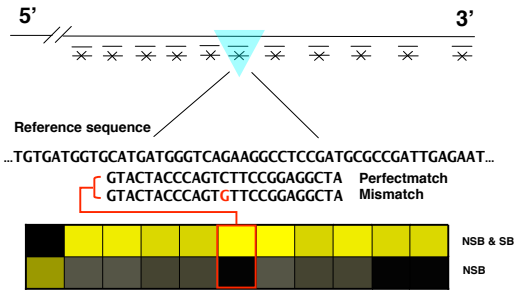
Feature Level Data

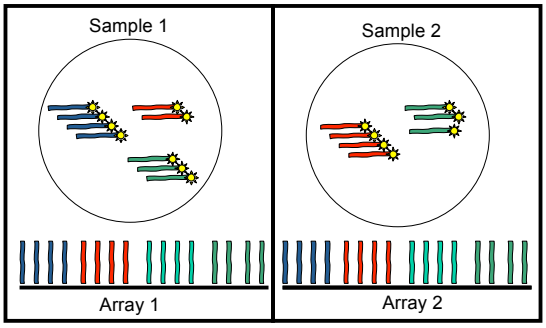
Outline

- Affymetrix GeneChip arrays
- Two color platforms

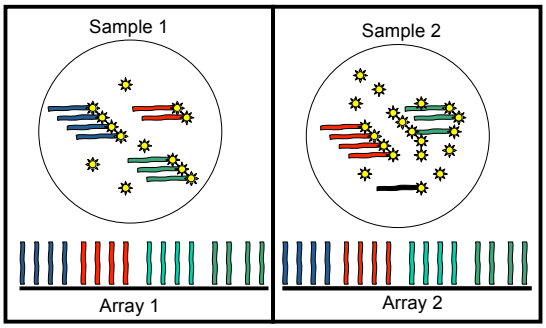
Affymetrix GeneChip Design



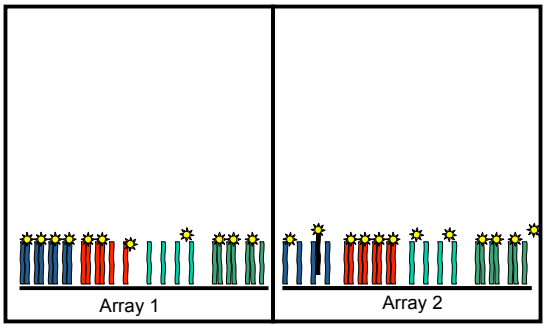
Before Hybridization



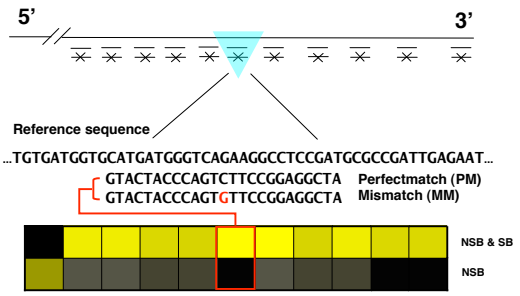
More Realistic



Non-specific Hybridization



Affymetrix GeneChip Design



GeneChip Feature Level Data

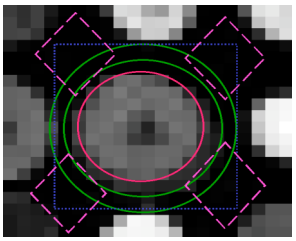
- MM features used to measure optical noise and non-specific binding directly
- More than 10,000 probesets
- Each probeset represented by 11-20 feature Note 1: Position of features are haphazardly distributed about the array. Note 2: There are between 20-100 chip types
- So we have PM_{gij}, MM_{gij} (*g* is gene, *i* is array and *j* is feature)
- A default summary is the avg of the PM-MM

Two color platforms

- Common to have just one feature per gene
- Typically, longer molecules are used so non-specific binding not so much of a worry
- Optical noise still a concern
- After spots are identified, a measure of local background is obtained from area around spot

Local background

---- GenePix
---- QuantArray
---- ScanAnalyze



GenePix does something different these days

Two color feature level data

- **Red and Green foreground and and background obtained from each feature**
- **We have Rf_{gij}, Gf_{gij}, Rb_{gij}, Gb_{gij} (*g* is gene, *i* is array and *j* is replicate)**
-
- **A default summary statistic is the log-ratio: (Rf-Rb) / (Gf - Gb)**

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!

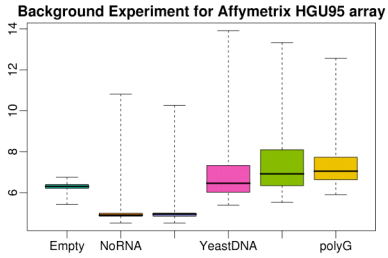
Background Effects Experiments

Learn about optical effect and NSB

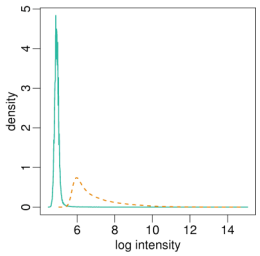
	label	sample type
Empty	0	empty
NoRNA	1	no RNA
NoLabel	0	human
YeastDNA	1	yeast genomic DNA
polyC	1	poly C
polyG	1	poly G

The Background Effects

Background Effect

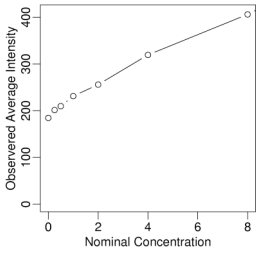


Background Effect

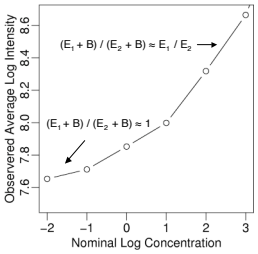


This are the no-label and Yeast DNA chips

Why Adjust for Background?

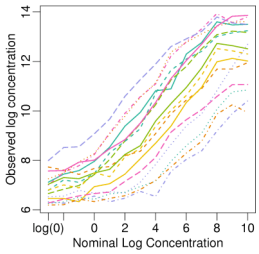


Why Adjust for Background?

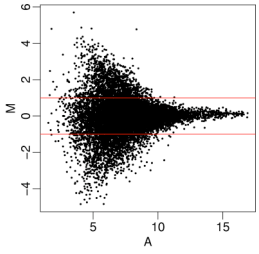


Notice local slope decrease as the nominal concentration becomes small

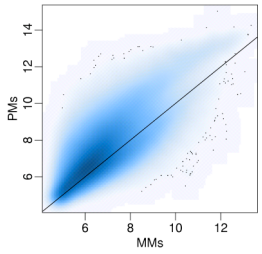
Probe-specific NSB



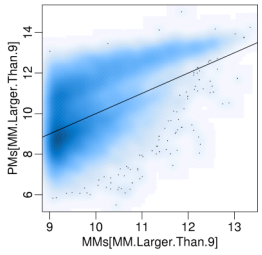
Why not subtract MM,BG?



Why not subtract MM?



Why not subtract MM?



Solutions

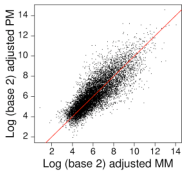
Direct Measurement Strategy

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**

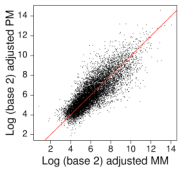


Stochastic Model

Better to assume:

$PM = B_{PM} + S$
$MM = B_{MM}$
$Cor[\log(B_{PM}), \log(B_{MM})] = 0.7$ → $Var[\log(PM - MM^*)] \sim 1/S^2$

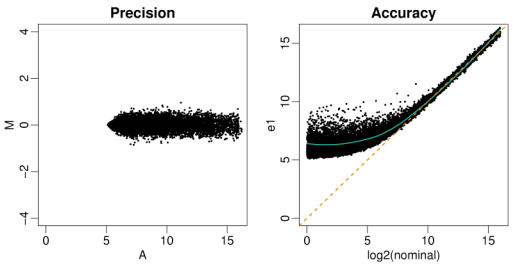
Alternative solution:
E[S I PM]



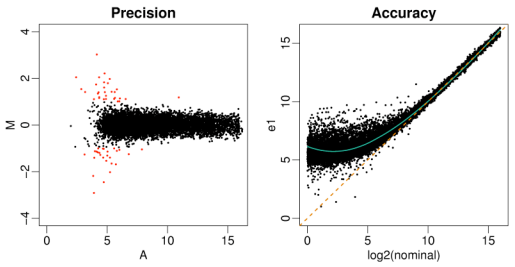
Simulation

- **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.**

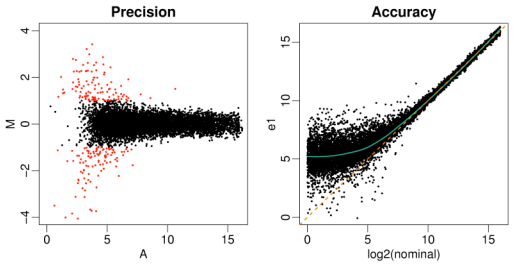
k=0


Precision
Accuracy
M
A
e1
log2(nominal)

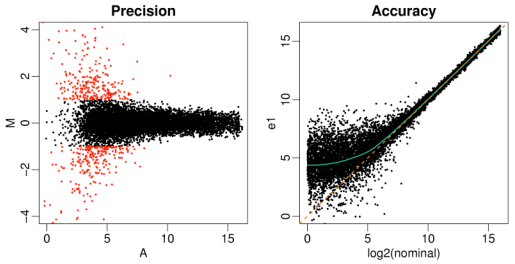
k=1/4


Precision
Accuracy
M
A
e1
log2(nominal)

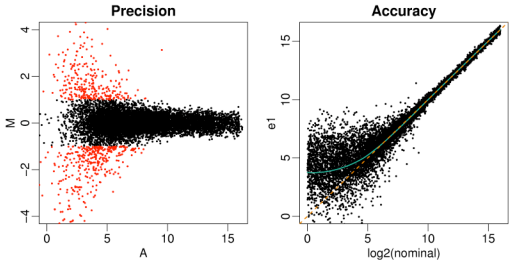
k=1/2


Precision
Accuracy
M
A
e1
log2(nominal)

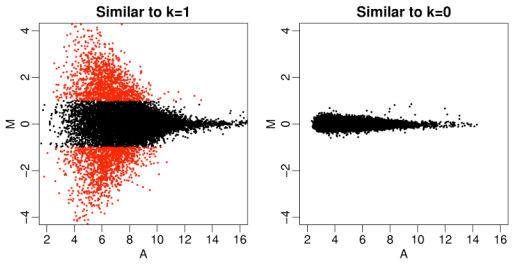
k=3/4



k=1



Real Data



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

The Basic Idea

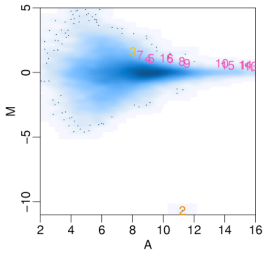
PM=B+S

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

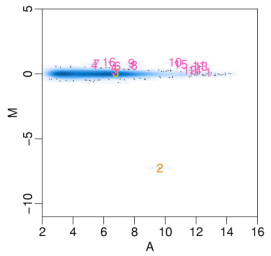
$$\hat{S} = \mathrm{E}[S \mid PM]$$

  - No MM
  - Borrowing strength across probes

MAS 5.0



RMA



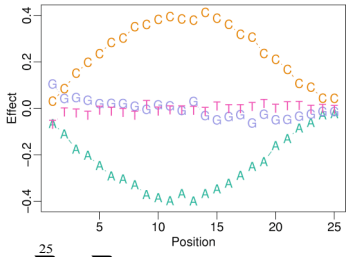
Notice improved precision but worst accuracy

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



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

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

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

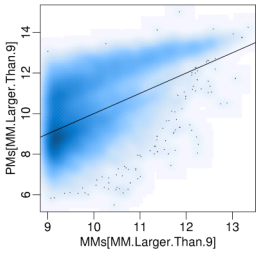
- **Use this stochastic model**
- **Minimize the MSE:**

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

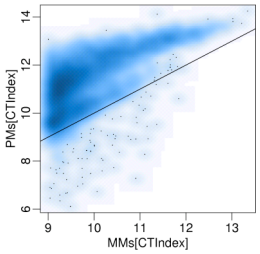
- **To do this we need to specify distributions for the different components**
- **Notice this is probe-specific so we need to borrow strength**

These parametric distributions were chosen to provide a closed form solution

Explains Bimodality



C,T in the middle



A,G in the middle