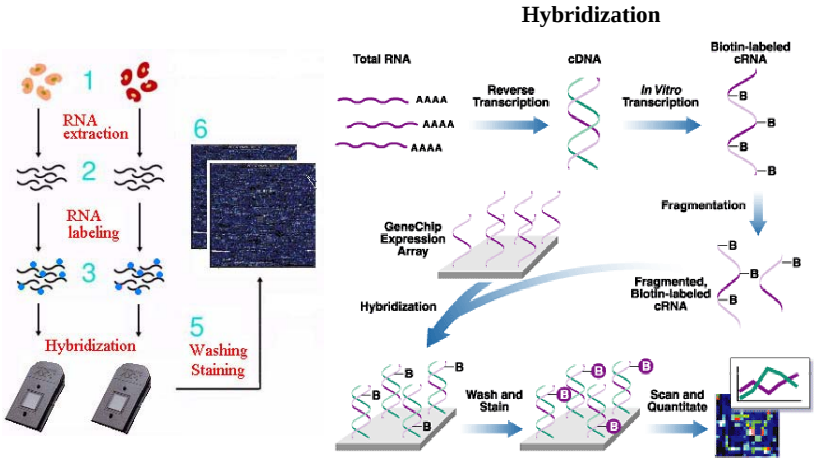
Statistical analysis and software for Affymetrix GeneChip arrays and some recent advances

George C. Tseng
Dept of Biostatistics / Human Genetics
University of Pittsburgh
12/15/04

Outline

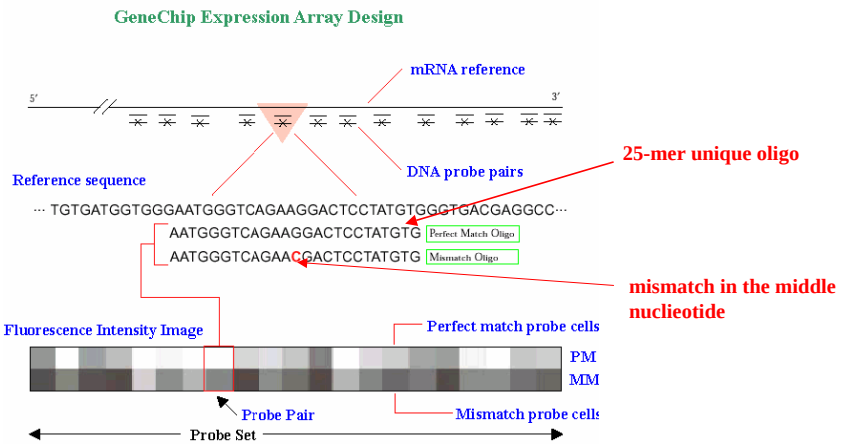
1. **GeneChip introduction**
   - **Design & manufacturing**
   - **Chip advances**
2. **Probe level analysis**
   - **Background adjustment**
     - PM-MM
     - PM only
     - RMA
     - GC-RMA
   - **Normalization**
     - Constant
     - Rank invariant
     - loess
     - Quantile normalization
   - **Expression measure**
     - MAS 4.0
     - LI-Wong (dChip)
     - MAS 5.0
     - RMA
3. **A simple case study by Bioconductor**

Overview of the technology



from Affymetrix Inc.

Array Design

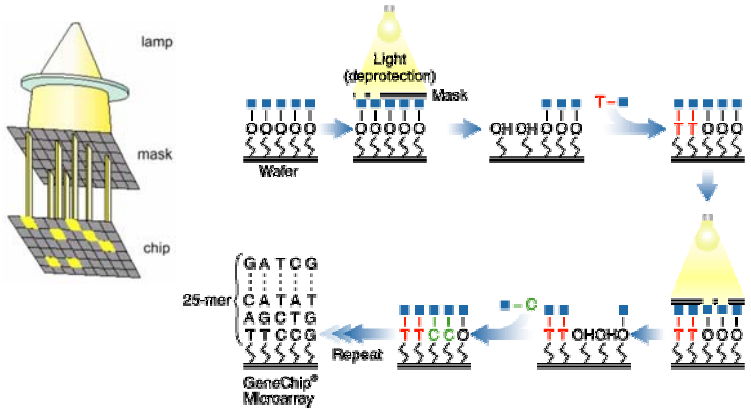


multiple probes (11~16) for each gene

from Affymetrix Inc.

Array Manufacturing

Technology adapted from semiconductor industry.
(photolithography and combinatorial chemistry)



Needs at most 4×25=100 masking and coupling.

from Affymetrix Inc.

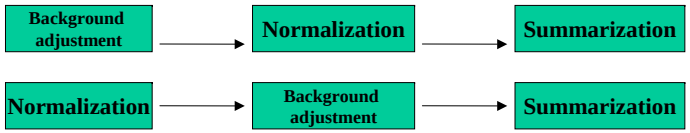
Chip Advances

	HG-U95	HG-U133 Set	HG-U133 Plus 2.0 Array
sequence source	Build 95 UniGene database (Oct, 2, 1999??)	Build 133 UniGene database (April, 20, 2001)	Build 133 UniGene database (April, 20, 2001)
Probe uniqueness	21/25 bases	Two 8-mers including at least one 12-mer	Two 8-mers including at least one 12-mer
# of probes	~16	11	11
# of arrays	5	2	**1**
# of transcripts	~54000 genes **HG-U95Av2: ~12000** **HG-U95B-E: ~44000 EST**	~33,000 genes	~38500 genes
Feature size	20 µm	18 µm	**11 µm**

Chip Advances

- Few years ago, U95 set had 5 arrays. Normally only U95Av2 is used.
- Improved probe selection algorithm to avoid non-specific binding.
  ⇒ Decreased # of probes in each probe set (20 => 11)
- Smaller probe size
  20 µm => 11 µm
- More genes on each array and less cost
  (Only one array for HG-U133 Plus)

Array Probe Level Analysis



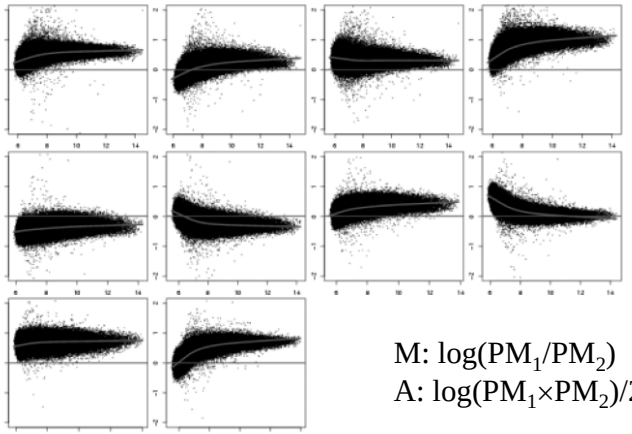
- Give an expression measure for each probe set on each array
- The result will greatly affect subsequent analysis (e.g. clustering and classification). If not modeled properly,
  => "Garbage in, garbage out"

We will leave the discussion of "backgound adjustment" to the last because there're more new exciting & technical advances.

Normalization

The need for normalization:

M-A plot



M: $\log(PM_1/PM_2)$
A: $\log(PM_1 \times PM_2)/2$

Normalization

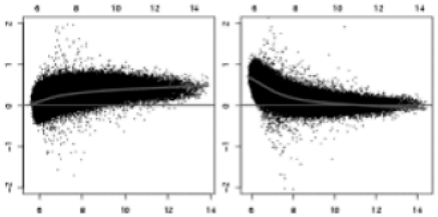
Constant scaling

- Distributions on each array are scaled to have identical mean.
- Applied in MAS 4.0 and MAS 5.0 but they perform the scaling after computing expression measure.

Normalization

Loess (Terry Speed)

- Using all genes to fit a non-linear normalization curve at the M-A plot scale.
- Perform normalization between arrays pairwisely.
- Has been extended to perform normalization globally without selecting a baseline array but then is time-consuming.



Normalization

Invariant set (dChip)

- Select a baseline array (default is the one with median average intensity).
- For each “treatment” array, identify a set of genes that have ranks conserved between the baseline and treatment array. This set of rank-invariant genes are considered non-differentially expressed genes.
- Each array is normalized against the baseline array by fitting a non-linear normalization curve of invariant-gene set.

Invariant set (dChip)

Advantage:

More robust than fitting with all genes as in loess. Especially when expression distribution in the arrays are very different.

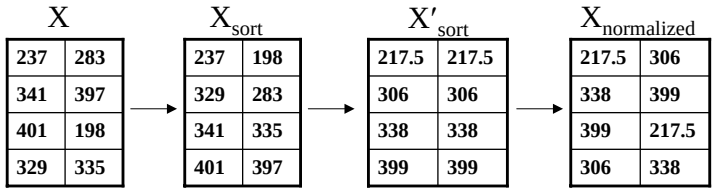
Disadvantage:

The selection of baseline array is important.

Normalization

Quantile normalization (RMA)

1. Given n array of length p, form X of dimension p×n where each array is a column.
2. Sort each column of X to give X_{sort}.
3. Take the means across rows of X_{sort} and assign this mean to each element in the row to get X'_{sort}.
4. Get $X_{normalized}$ by rearranging each column of X'_{sort} to have the same ordering as original X.



Normalization

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

A careful comparison of different normalization methods and concluded to settle with quantile normalization

Summarize Expression Index

- **MAS 4.0**

For each probe set, (I: # of arrays, J: # of probes)

PM_{ij}-MM_{ij}= $\theta_i + \varepsilon_{ij}$, i=1,…,I, j=1,…,J

θ_i estimated by average difference

1. Negative expression
2. Noisy for low expressed genes
3. Not account for probe affinity
4. Average without taking log

❖ **dChip (DNA chips)**

For each probe set, (I: # of arrays, J: # of probes)

$PM_{ij}=\upsilon_j + \theta_i\alpha_j + \theta_i\phi_j + \varepsilon''_{ij}$

$MM_{ij}=\upsilon_j + \theta_i\alpha_j + \varepsilon'_{ij}$

$PM_{ij} - MM_{ij}= \theta_i\phi_j + \varepsilon_{ij}$, i=1,…,I, j=1,…,J

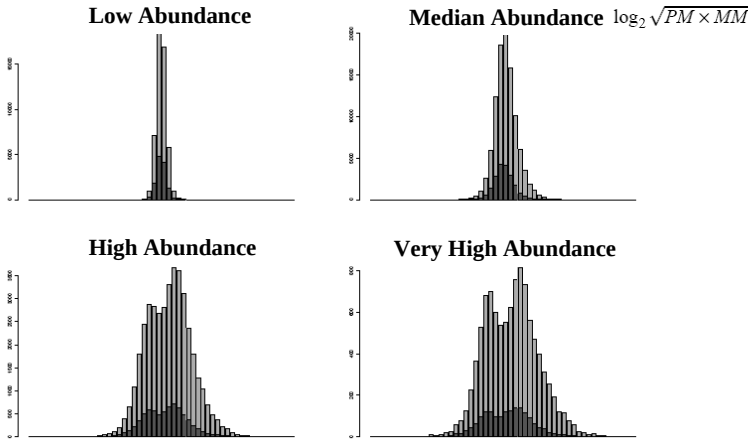
$\Sigma\phi_j = J$, $\varepsilon_{ij} \sim N(0, \sigma^2)$

1. Account for probe affinity effect, ϕ_j.
2. Outlier detection through multi-chip analysis
3. Recommended for more than 10 arrays

Multiplicative model: $PM_{ij} - MM_{ij}= \theta_i\phi_j + \varepsilon_{ij}$ (better)

Additive model: $PM_{ij} - MM_{ij}= \theta_i + \phi_j + \varepsilon_{ij}$

Histograms of log ratio $\log_2(PM/MM)$



❖ **MAS 5.0**

For each probe set, (I: # of arrays, J: # of probes)

$\log(PM_{ij}-CT_{ij})=\log(\theta_i)+\varepsilon_{ij}$, i=1,…,I, j=1,…,J

$CT_{ij}=MM_{ij}$ if $MM_{ij}<PM_{ij}$

less than PM_{ij} if $MM_{ij}\geq PM_{ij}$

θ_i estimated by a robust average (Tukey biweight).

1. No more negative expression
2. Taking log adjusts for dependence of variance on the mean.

❖ **RMA (Robust Multi-array Analysis)**

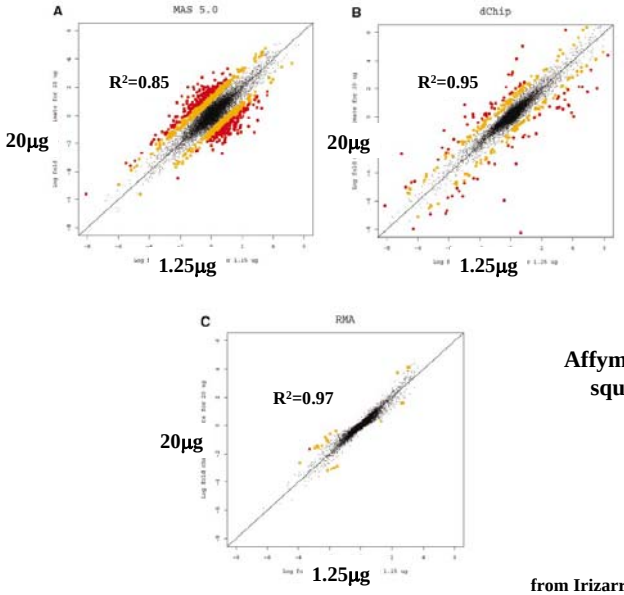
For each probe set, (I: # of arrays, J: # of probes)

$\log(T(PM_{ij}))= \theta_i + \phi_j + \varepsilon_{ij}$, i=1,…,I, j=1,…,J

T is the transformation for background correction and normalization

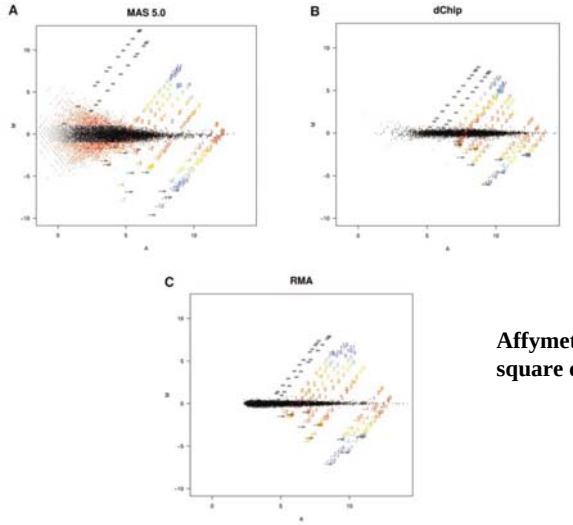
$\varepsilon_{ij} \sim N(0, \sigma^2)$

1. Log-scale additive model
2. Suggest not to use MM
3. Fit the linear model robustly (median polish)



Affymetrix Latin square data

from Irizarry et al. (NAR, 2003)



Affymetrix Latin square data

from Irizarry et al. (NAR, 2003)

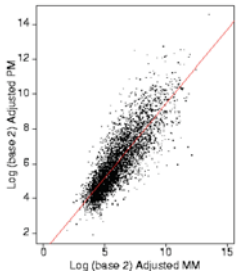
Background Adjustment

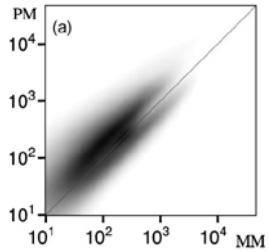
- **Direct subtraction: PM-MM**
  MAS4.0, dChip, MAS5.0

  Assume the following deterministic model:
  PM=O+N+S **(O: optical noise, N: non-specifi binding)**
  MM=O+N

  => PM-MM=S>0

  Is it true?





MM does not measure background noise of PM

- Yeast sample hybridized to human chip
- If MM measures non-specific binding of PM well, PM≅MM.
- R^2 only 0.5.

Many MM>PM

- 86 HG-U95A human chips, human blood extracts
- Two fork phenomenon at high abundance
- 1/3 of probes have MM>PM

Reasons MM should not be used:

1. MM contain non-specific binding information but also include signal information and noise
2. The non-specific binding mechanism not well-studied.
3. MM is costly (take up half space of the array)

❖ **Ignore MM**

dChip has an option for PM-only model

❖ **RMA background correction**

$PM_{ijg}=B_{ijg}+S_{ijg}$

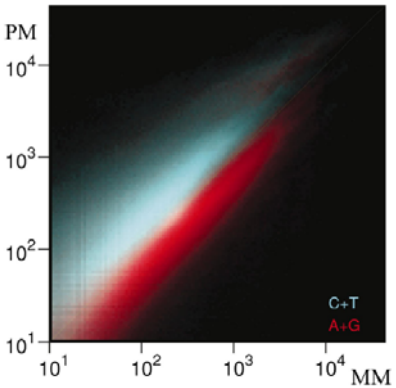
$B_{ijg}\sim N(\mu_i, \sigma_i^2)$, $S_{ijg}\sim Exp(\beta_{ig})$, B & S independent

Take $E(S_{ijg}|PM_{ijg})$ as the background adjusted PM.

1. S is exponential distributed to avoid negative signal.
2. The model doesn't fit data perfectly but seems work well in practice.

Is it possible to
- **Include MM in background adjustment?**
- **Increase signal while not sacrifice much on noise?**

Consider sequence information



AATGGGTCAGAAGGACTCCTATGTG Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTG Mismatch Oligo

Naef & Magnasco, 2003

1. 95% of (MM>PM) have purine (A, G) in the middle base.
2. In the current protocol, only pyrimidines (C, T) have biotin-labeled florescence.

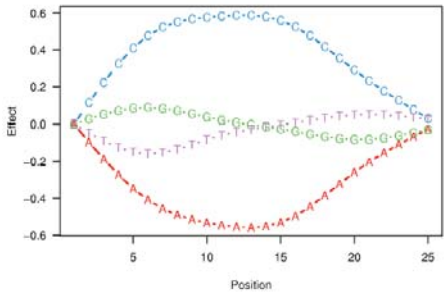
Fit a simple linear model:

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

$\mu_{j,k}$ represents the contribution to affinity of base j in position k.

using a spline with 5 degrees of freedom

1. C > G ≅ T > A
2. Boundary effect



$\hat{\mu}_{j,k}$

Some chemical explanation of the result:

PM	C	G	T	A
MM	G	C	A	T
labeling	Yes (+)	No (-)	Yes (+)	No(-)
Labeling impedes binding	Yes (-)	No	Yes (-)	No
Hydrogen bonds	3 (+)	3 (+)	2	2
Sequence specific brightness	High	average	average	Low

❖ GC-RMA

$$PM = O_{PM} + N_{PM} + S$$

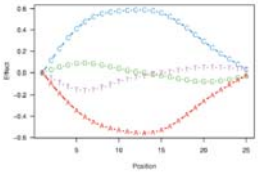
$$MM = O_{MM} + N_{MM} + \phi S$$

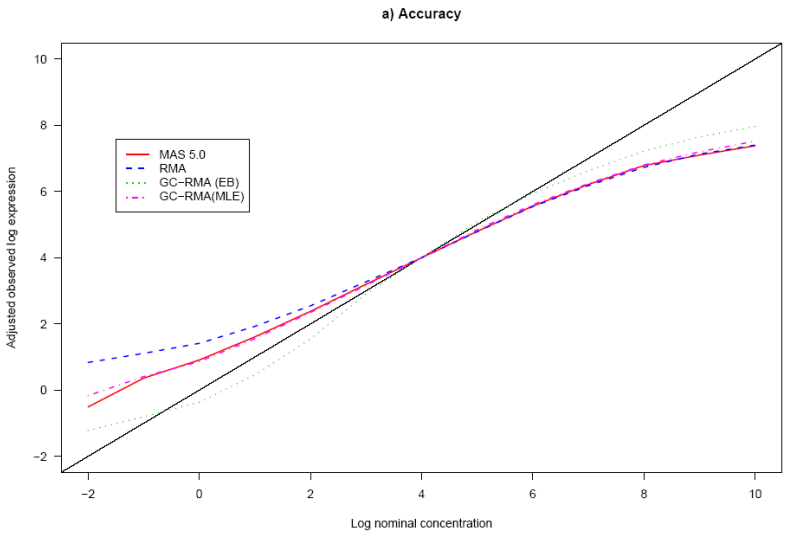
O: optical noise, log-normal dist.
N: non-specific binding

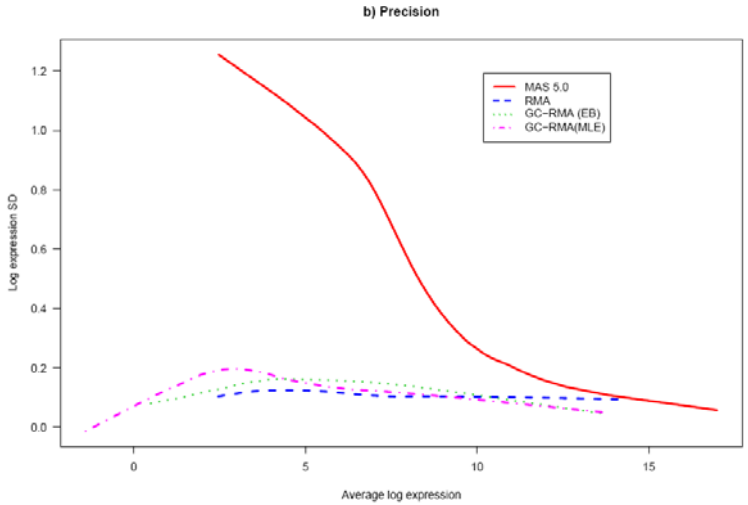
$$\begin{pmatrix} \log(N_{PM}) \\ \log(N_{MM}) \end{pmatrix} = N\left(\begin{bmatrix} \mu_{PM} \\ \mu_{MM} \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

$$\begin{pmatrix} \mu_{PM} \\ \mu_{MM} \end{pmatrix} = h\begin{pmatrix} \alpha_{PM} \\ \alpha_{MM} \end{pmatrix}$$

h: a smooth (almost linear) function.
α: the sequence information weight computed form the simple linear model.







GC-RMA

Method	0.50/0.25	1/0.5	2/1	4/2	8/4	16/8	32/16	64/32	128/64	256/128	512/256	1024/512
MAS 5.0	0.87	0.55	0.7	0.77	0.82	0.8	0.8	0.76	0.65	0.57	0.31	0.28
RMA	0.28	0.3	0.51	0.62	0.73	0.73	0.79	0.76	0.63	0.55	0.4	0.27
EB	0.42	0.44	0.83	1.08	1.37	1.09	1.09	0.85	0.68	0.6	0.43	0.32
MLE	0.58	0.45	0.7	0.79	0.82	0.84	0.85	0.76	0.64	0.56	0.4	0.32
PM-only EB	0.29	0.41	0.8	1.08	1.27	1.24	1.25	0.87	0.71	0.58	0.39	0.37

	Fee	GUI	Flexibility to programming and mining		Audience
MAS 4.0	Commercial	Yes	No	Average Difference	Manufacturer default
dChip	Free	Yes	Some extra tools	Li-Wong model	Biologists
MAS 5.0	Commercial	Yes	No	Robust average of log difference	Manufacturer default
RMAExpress	Free	Yes	No	RMA	Biologists
Bioconductor	Free	Some	Best	All of above	Statistician, programmer
ArrayAssist	Commercial ($1500)	Yes	No	RMA, GC-RMA	Biologists

Probe level analysis in Bioconductor

(affy package)

Background Methods	Normalization Methods	PM correction Methods	Summarization Methods
none rma/rma2 mas	quantiles loess contrasts constant invariantset qspline	mas pmonly subtractmm	avgdiff liwong mas medianpolish playerout

A Simple Case Study

Latin Square Data

59 HG-U95A arrays

14 spike-in genes in 14 experimental groups

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

M, N, O, P are replicates and Q, R, S, T another replicates

http://www.affymetrix.com/analysis/download_center2.affx

A Simple Case Study

- Take the following to replicate groups.

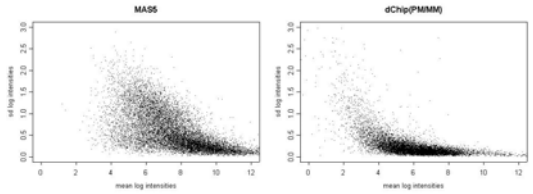
M	1521m99hpp_av06.CEL	1521q99hpp_av06.CEL	Q
N	1521n99hpp_av06.CEL	1521r99hpp_av06.CEL	R
O	1521o99hpp_av06.CEL	1521s99hpp_av06.CEL	S
P	1521p99hpp_av06.CEL	1521t99hpp_av06.CEL	T

- Use Bioconducotr to perform a simple evaluation of different probe analysis algorithms.
- Note: This is only a simple demonstration. The evaluation result in this presentation is not conclusive.

A Simple Case Study

- **Average log intensities vs SD log intensities. (the first four replicates: M, N, O, P)**
- **RMA looks different. Further investigation found RMA and GC-RMA had taken log transformation while others didn't.**

MAS5.0



dChip (PM/MM)

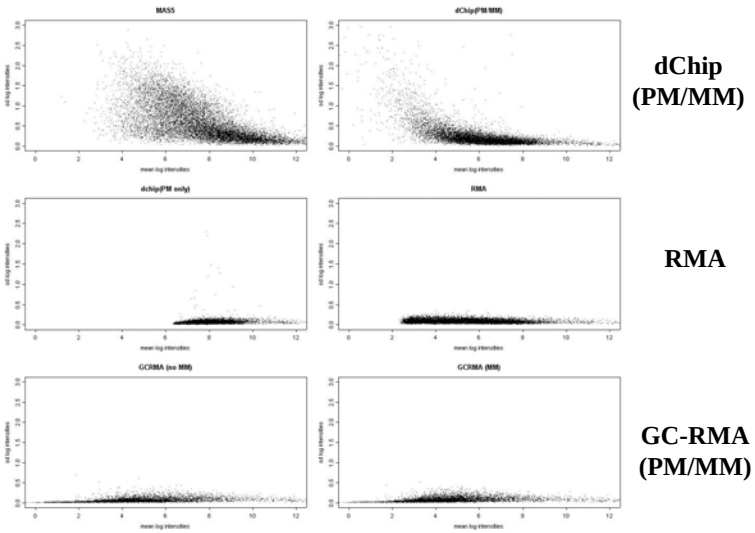
dChip (PM only)

RMA

A Simple Case Study

Average log intensities vs SD log intensities. (M, N, O, P)

MAS5.0



dChip (PM/MM)

dChip (PM only)

RMA

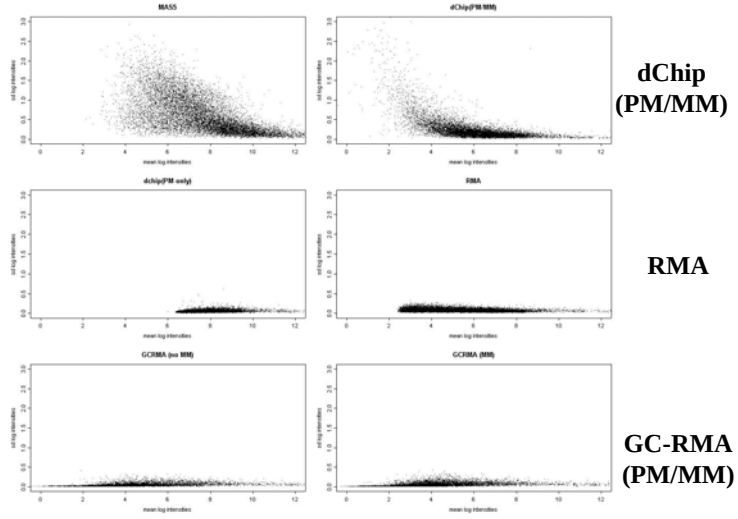
GC-RMA (PM only)

GC-RMA (PM/MM)

A Simple Case Study

Average log intensities vs SD log intensities. (Q, R, S, T)

MAS5.0



dChip (PM/MM)

dChip (PM only)

RMA

GC-RMA (PM only)

GC-RMA (PM/MM)

A Simple Case Study

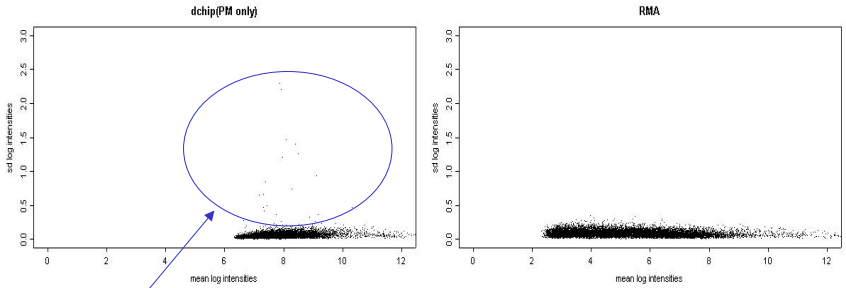
Average pair-wise correlations between replicates

	M, N, O, P	Q, R, S, T
MAS5	0.8930	0.9002
dChip (PM/MM)	0.9604	0.9621
dChip (PM-only)	0.9940	0.9966
RMA	0.9978	0.9978
GC-RMA(PM/MM)	0.9988	0.9990
GC-RMA(PM-only)	0.9993	0.9994

Replicate correlation performance:
GCRMA(PM-only)>GC-RMA(PM/MM)>RMA>dChip(PM-only)>>dChip(PM/MM)>>MAS5

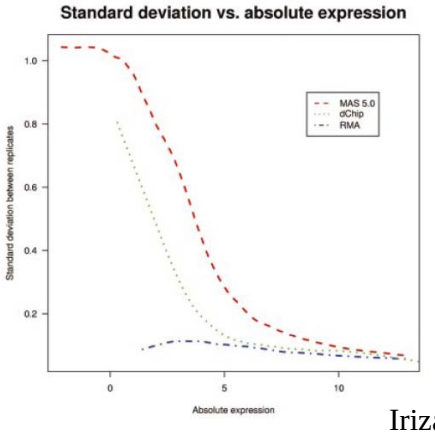
A Simple Case Study

- **RMA greatly improves dChip(PM/MM) but dChip(PM-only) model generally seems a little better than RMA.**
- **Average replicate correlations of RMA (0.9978) is a little better than dChip(PM only) (0.9940 & 0.9966)**
- **dChip(PM only) suffers from a number of outlying genes in the model.**



Outlying genes that do not fit Li-Wong model

Figure reported in the RMA paper. Seems RMA was only compared to dChip(PM/MM) model. Didn't mention how dChip PM-only model compares to RMA.



A different data set from GeneLogic was used.

Irizarry et al. 2003 *NAR*

Affycomp

- A package included in Bioconductor)
- Computes a number of measures to compare different probe level analysis methods.

Cope, LM, Irizarry, RA, Jaffee, H, Wu, Z, Speed, TP (2004) A Benchmark for Affymetrix GeneChip Expression Measures. *Bioinformatics* 20: 323-331.

Conclusion:

1. Technological advances have been made to have smaller probe size and better sequence selection algorithms to reduce # of probes in a probe set. This will enable more biologically meaningful genes on a slide and reduce the cost.

2. Recent analysis advances have been focused on understanding and modelling hybridization mechanisms. This will allow a better use of MM probes or eventually suggest to remove MMs from the array.

References:

Affymetrix GeneChip:
-- Lockhart et al. Expression monitoring by hybridization to high-density oligonucleotide arrays (PNAS 2001)

Normalization:
-- 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

dChip:
-- Li and Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection (PNAS 2001)

RMA:
-- Irizarry, RA, Bolstad BM, Collin, F, Cope, LM, Hobbs, B, and, Speed, TP (2003) Summaries of Affymetrix GeneChip Probe Level Data. *Nucleic Acids Research*. Vol. 31, No. 4 e15

Background and MM exploration:
-- Felix Naef, Daniel A. Lim, Nila Patil, and Marcelo Magnasco, *DNA hybridization to mismatched templates: a chip study*, Phys. Rev. E 65, 040902 (2002)
-- Felix Naef and Marcelo Magnasco, *Solving the riddle of the bright mismatches: the physics of hybridization*, Phys. Rev. E 68, 011906 (2003).

GC-RMA
-- Wu, Zhijin, Irizarry, RA, Gentleman, R, Martinez Murillo, F, Spencer, F (2003) A Model Based Background Adjustement for Oligonucleotide Expression Arrays. To appear in JASA.

Bioconductor
-- Gautier, L, Cope, LM, Bolstad, BM, and Irizarry, RA (2002) affy - A package for the analysis of Afftmetrix GeneChip data at the probe level.. Bioinformatics.

Performance evaluation
-- Cope, LM, Irizarry, RA, Jaffee, H, Wu, Z, Speed, TP (2004) A Benchmark for Affymetrix GeneChip Expression Measures. *Bioinformatics* 20: 323-331.