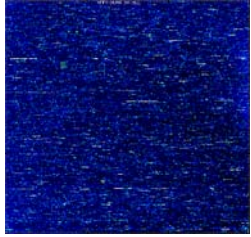


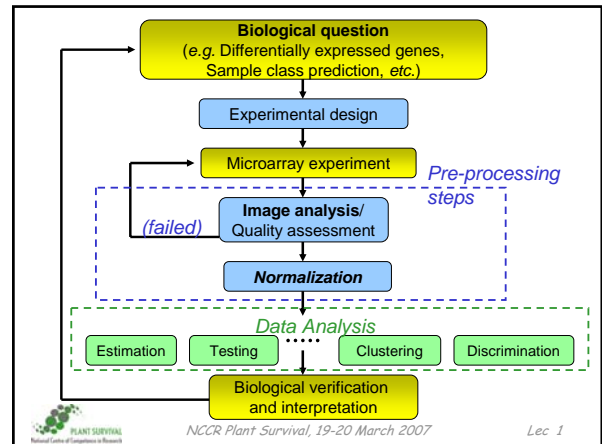
Statistics for Affymetrix GeneChips

Affymetrix signal quantification;
Introduction to quality assessment

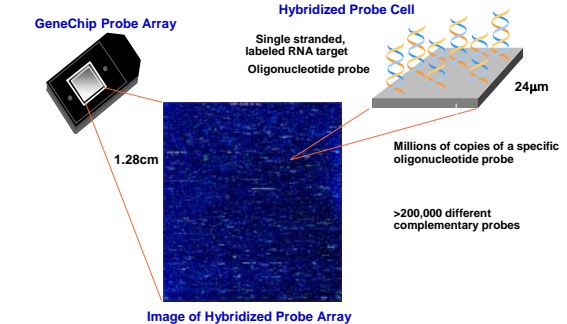


<http://www.isrec.isb-sib.ch/~darlene/NCCR-PS/>

NCCR Plant Survival, 19-20 March 2007 Lec 1



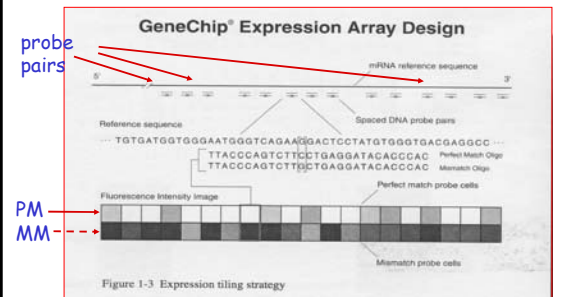
Affymetrix GeneChip Probe Arrays



Compliments of D. Gerhold

NCCR Plant Survival, 19-20 March 2007 Lec 1

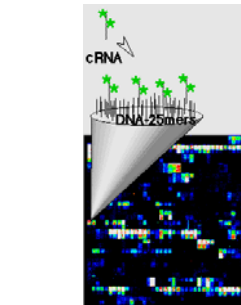
Array design



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

NCCR Plant Survival, 19-20 March 2007 Lec 1

Image analysis



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

NCCR Plant Survival, 19-20 March 2007 Lec 1

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)*
- Not the same as M-values for cDNA microarrays, but analogous

NCCR Plant Survival, 19-20 March 2007 Lec 1

Some possible problems

- Some probe pairs may hybridize better than the rest
- Removing the middle base might not make a difference for some probes
- Some MMs are PMs for some other gene
- There is a need for normalization



NCCR Plant Survival, 19-20 March 2007

Lec 1

Expression Measures

- There are many possibilities for getting an expression measure
- 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.*)



NCCR Plant Survival, 19-20 March 2007

Lec 1

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



NCCR Plant Survival, 19-20 March 2007

Lec 1

Why ignore MM values?

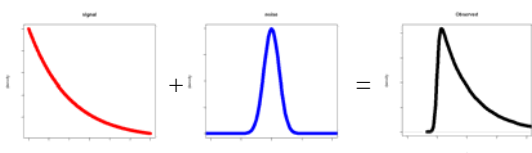
- MM probes are meant to measure background (noise), but the MM values have information about *both* signal and noise
- Using it without adding more noise is challenging and is a topic of current research (*gcrma*)
- It should be possible to improve the bg correction using MM, without having the noise level increase greatly



NCCR Plant Survival, 19-20 March 2007

Lec 1

Background model pictorially



Signal + Noise = Observed

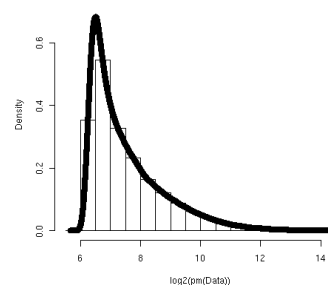


NCCR Plant Survival, 19-20 March 2007

Lec 1

PM data on \log_2 scale

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



NCCR Plant Survival, 19-20 March 2007

Lec 1

Quantile normalization

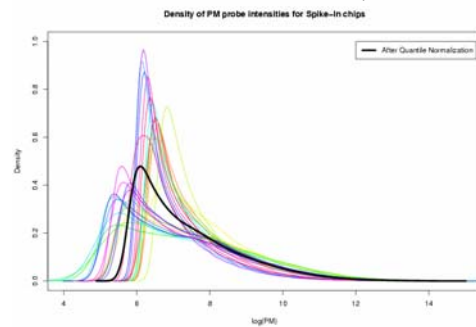
- The purpose of *normalization* is to remove artifactual differences between arrays (e.g., differences in total intensity)
- Quantile normalization makes the distribution of probe intensities *the same for every chip*
- The normalization distribution is chosen by *averaging each quantile* across chips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



NCCR Plant Survival, 19-20 March 2007

Lec 1

Quantile normalization: pictorially

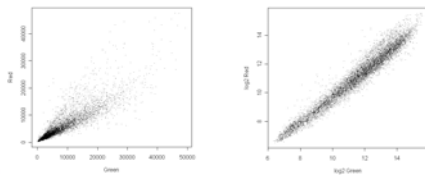


NCCR Plant Survival, 19-20 March 2007

Lec 1

Scatterplots

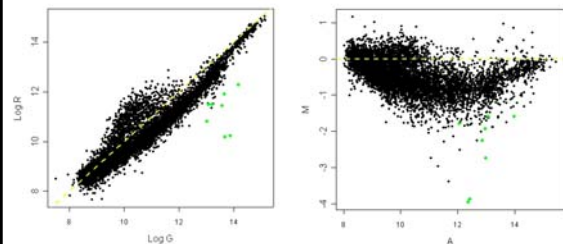
- MA plot (also sometimes called MVA plot, RI plot, SI plot):
 - A = average of two measurements
 - M = difference of the measurements
- Why we take logs:



NCCR Plant Survival, 19-20 March 2007

Lec 1

Scatterplots: always log*, always rotate



Chip 2 vs Chip 1

$M = \text{Chip 2} - \text{Chip 1}$ vs
 $A = (\text{Chip 2} + \text{Chip 1})/2$

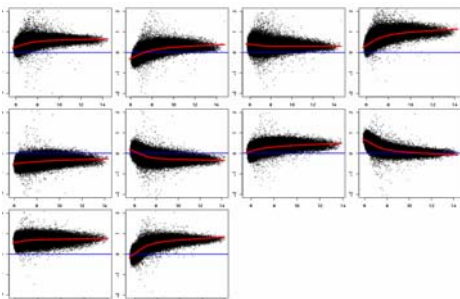
* Other transformations may provide improvement



NCCR Plant Survival, 19-20 March 2007

Lec 1

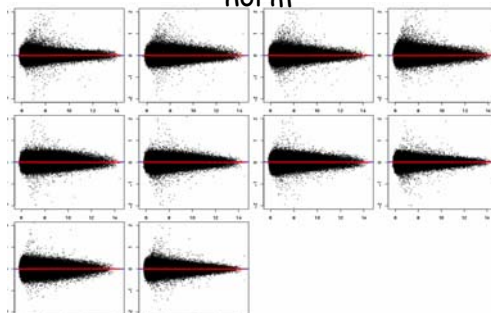
MA plots of chip pairs: before norm



NCCR Plant Survival, 19-20 March 2007

Lec 1

MA plots of chip pairs: after quantile norm



NCCR Plant Survival, 19-20 March 2007

Lec 1

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?*
To get even more out of the multi-chip analysis: robust summaries can improve over the standard ones by down-weighting outliers



NCCR Plant Survival, 19-20 March 2007

Lec 1

Robust Multi-chip Analysis

- Base analysis on the linear model embodying the parallel behavior:
 $\log_2 n(\text{PM}_{ij} - *BG) = \text{chip effect}_i + \text{probe effect}_j + \epsilon_{ij}$
where i labels chips and j labels probes ('genes')
- Current implementation estimates using *median polish* (it's faster than IRLS)



NCCR Plant Survival, 19-20 March 2007

Lec 1

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*



NCCR Plant Survival, 19-20 March 2007

Lec 1

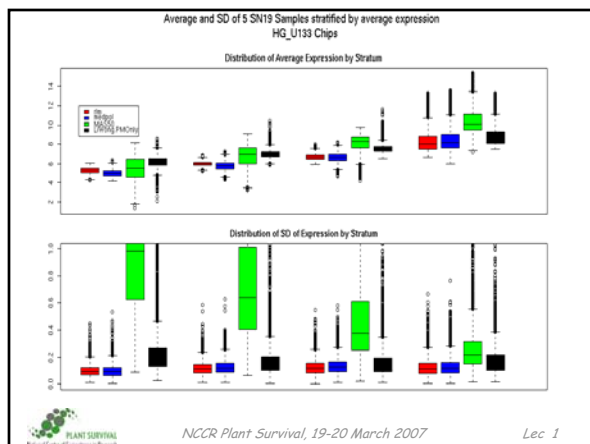
Comparisons

- Trade-offs, of either:
 - Bias/variance (accuracy/precision), or
 - False positives/true positives
- GeneLogic and Affymetrix have carried out purposeful experiments where the *truth* is 'known' so that such quantities can be assessed

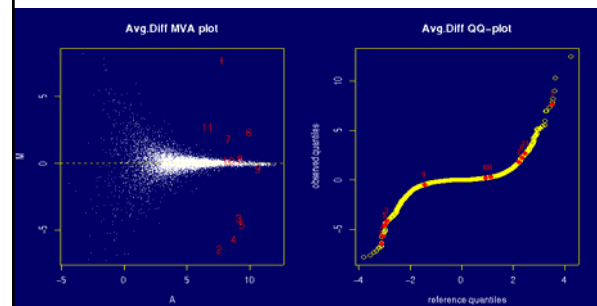


NCCR Plant Survival, 19-20 March 2007

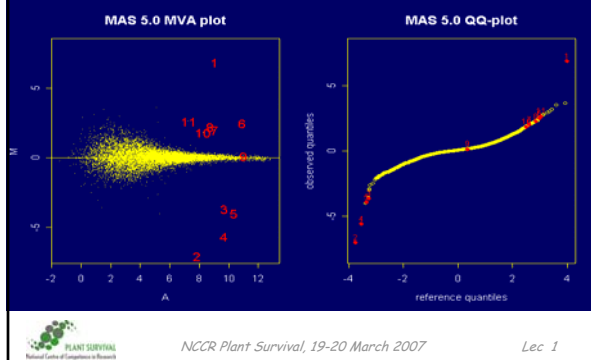
Lec 1



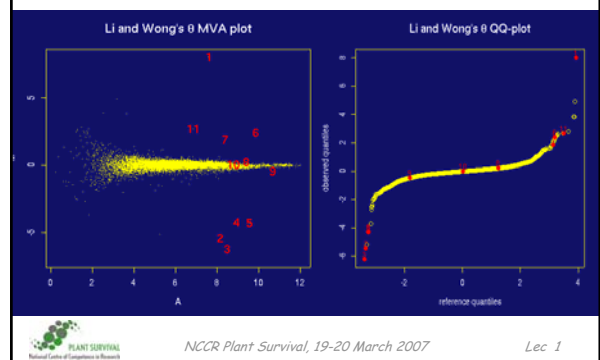
Differential Expression: AvDiff



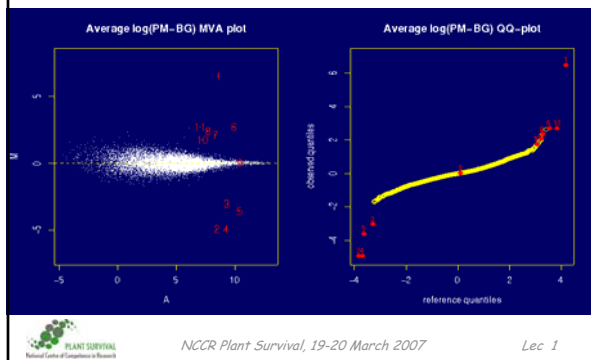
Differential expression: MAS 5.0



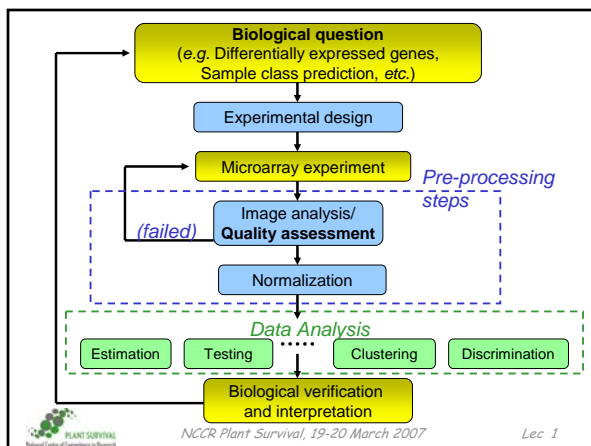
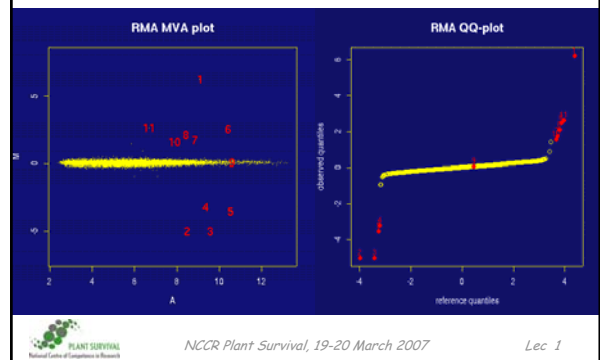
Differential expression: Li-Wong



Differential expression: log(PM-BG)



Differential expression: RMA



Affymetrix recommended QC

- Sample prep QC
 - bioanalyzer profiles
- Data QC
 - preliminary checks: inspect image, oligo b2, grid alignment
 - rpt file

Data quality metrics in rpt file

- Control spikes: BioB, BioC, BioD, cre
- Internal control genes: actin, gapdh
- % Present call
- Scaling Factor (if scaling)
- Noise (RawQ)
- Background



NCCR Plant Survival, 19-20 March 2007

Lec 1

Oligo B2 Performance



NCCR Plant Survival, 19-20 March 2007

Lec 1

Control Spikes

Spike Controls:

Probe Set	Sig(5')	Det(5')	Sig(M')	Det(M')	Sig(3')	Det(3')
Sig(all)	Sig(3'/5')					
BIOC	60.8 M	63.7 P	63.9 A	62.81 1.05		
BIOC	134.7 P		75.1 P	104.91 0.56		
BIOCN	105.0 P		677.7 P	391.35 6.46		
CREK	907.2 P		1486.7 P	1196.97 1.64		
DAPX	14.6 A	8.5 A	1.8 A	8.30 0.12		
LYSX	1.4 A	8.4 A	11.0 A	6.92 8.09		
PHEX	3.7 A	1.8 A	5.3 A	3.60 1.46		
TRFX	1.4 A	4.0 A	3.3 A	2.91 2.39		
TRFX	4.2 A	4.3 A	1.7 A	3.42 0.40		

- BioB should be P ~ 70% of the time
- BioC, BioD, cre should always be P



NCCR Plant Survival, 19-20 March 2007

Lec 1

Internal control genes

Housekeeping Controls:

Probe Set	Sig(5')	Det(5')	Sig(M')	Det(M')	Sig(3')	Det(3')
Sig(all)	Sig(3'/5')					
HIMISCF3A/M97935	26.4 P	149.6 M	272.6 P	149.54	10.31	
HIMRGE/M10098	3.1 A	5.0 A	10.7 A	6.26 3.49		
HIMGAPDH/M33197	3300.4 P	3005.6 P	3221.6 P	3175.87	0.98	
HEPCD7/X00351	7532.9 P	8839.1 P	6645.4 P	7672.49	0.88	
MZ7830	65.3 P	35.7 A	144.4 A	81.81 2.21		

- actin, gapdh should have all P
- 3'/5' ratio < 3



NCCR Plant Survival, 19-20 March 2007

Lec 1

% Present

Total Probe Sets: 22283
Number Present: 9235 41.4%
 Number Absent: 12666 56.8%
 Number Marginal: 382 1.7%

Average Signal (P): 413.4
 Average Signal (A): 28.8
 Average Signal (M): 87.6
 Average Signal (All): 189.2

- % P ~ 30 - 50%
- 'good indicator of assay performance'
- similar values across replicates (also SF, RawQ)



NCCR Plant Survival, 19-20 March 2007

Lec 1

Background

Background:
 Avg: 83.50 Std: 2.02 Min: 77.40 Max: 89.30

Noise:
 Avg: 4.46 Std: 0.28 Min: 3.60 Max: 5.40

Corner+
 Avg: 112 Count: 32

Corner-
 Avg: 8894 Count: 32

Central-
 Avg: 7568 Count: 9

- Should be under 100
- similar values across replicates



NCCR Plant Survival, 19-20 March 2007

Lec 1

Problems with these measures

- Relate to the experimental process, and *not* directly to the end result (a measure of gene expression)
- *Single chip* measures, which do not put each chip in the context of the others
- dChip makes 'outlier' calls, but algorithm also has some troubling aspects
- *by-products of the RMA calculation can also provide quality information*



(BREAK)

