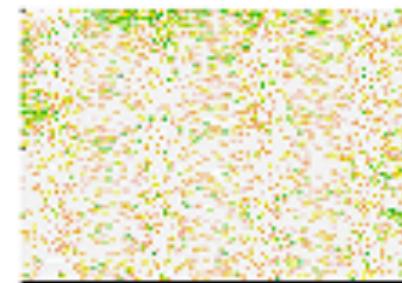
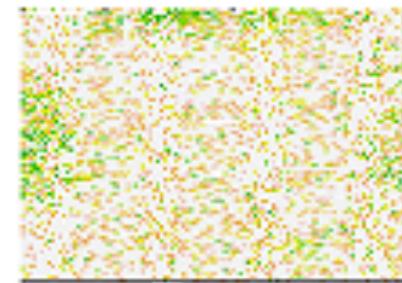
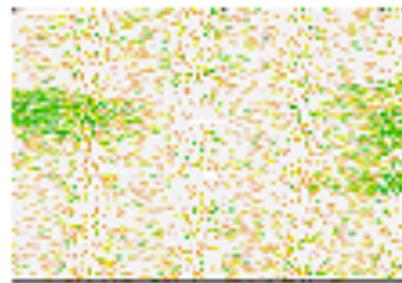


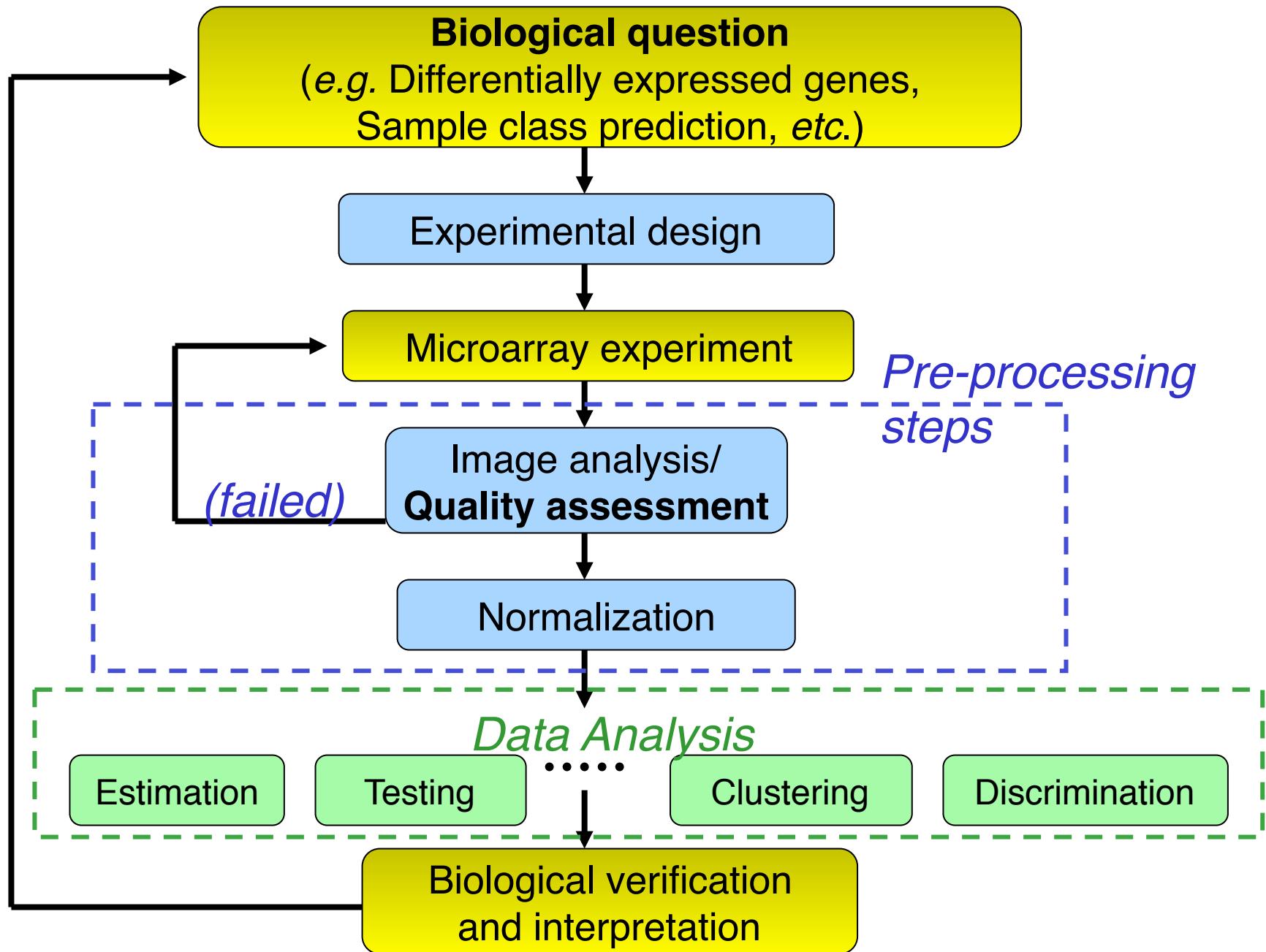
Statistics for Genomic Data Analysis

Affymetrix QA/QC ; Robust regression



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





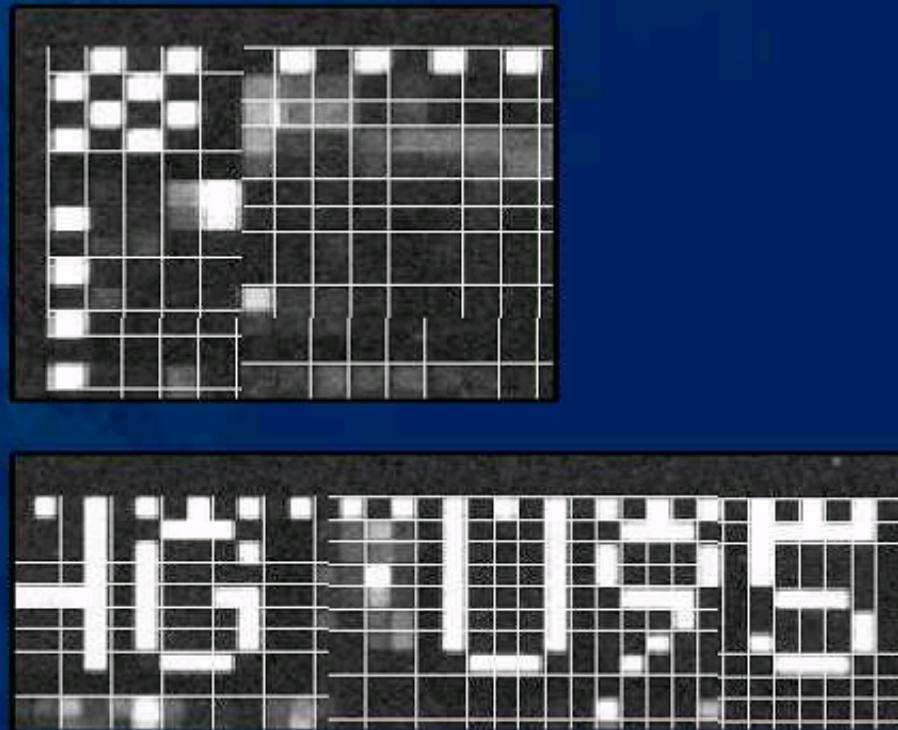
Affymetrix recommended QC

- Sample prep QC
 - pre-hyb QC
 - bioanalyzer profiles
 - preempt hybing poor quality
- Data QC
 - post-hyb QC
 - visual inspection of image, oligo b2, grid alignment
 - metrics in rpt file



Oligo B2 Performance

Oligo B2 Performance & Grid Alignment



AFFYMETRIX 



Spike-ins and controls

- *Unlabelled poly-A controls* : dap, lys, phe, thr, tryp ; used to monitor wet lab work
- *Hybridization controls* : bioB, bioC, bioD, cre
- *Housekeeping/control genes* : actin, gapdh
 - 3' to 5' signal intensity ratios of control probe sets



Control Spikes

Spike Controls:

Probe Set	Sig(5')		Det(5')		Sig(M')		Det(M')		Sig(3')		Det(3')	
	Sig(all)	Sig(3'/5')										
BIOB	60.8	M	63.7	P	63.9	A	62.81	1.05				
BIOC	134.7	P			75.1	P	104.91	0.56				
BIODN	105.0	P			677.7	P	391.35	6.46				
CREX	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				
THRX	1.4	A	4.0	A	3.3	A	2.91	2.39				
TRPNX	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



Internal control genes

Housekeeping Controls:

Probe Set	Sig(5')	Det(5')	Sig(M')	Det(M')	Sig(3')	Det(3')
Sig(all)	Sig(3'/5')					
HUMSGE3A/M97935	26.4 P	149.6 M	272.6 P	149.54	10.31	
HUMREFM10098	3.1 A	5.0 A	10.7 A	6.26	3.49	
HMGAT1/HMB3197	3300.4	P	3005.6	P	3221.6	P
HSAC07/X00361	7532.9	P	8839.1	P	6645.4	P
M27830	65.3 P	35.7 A	144.4 A	81.81	2.21	

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



Quality metrics in Affy rpt file

- % Present call: *20-50%* ; consistency
- Scaling Factor:
 - Target/(2% trimmed mean signal values) ; consistency
- P/A calls, SF : measure how much is PM > MM
- Background: *under 100* ; consistency
 - Average signal in lowest 2%
- Noise (RawQ): *1.5-3* is ok
 - Pixel-to-pixel variation among probe cells used to calculate the background



MAS 5 algorithms

- Present calls : p -value from Wilcoxon signed rank test based on $R_i = (PM_i - MM_i) / (PM_i + MM_i)$
 - H : median $(R_i - \tau) = 0$ vs. A : median $(R_i - \tau) > 0$
 - τ small ($= 0.015$)
 - P = 'present': $p < 0.04$; A = 'absent': $p \geq 0.06$;
 M = 'marginal': $0.04 < p < 0.06$
- Signal : $\log_2(S) = \sum_i w_i \log_2 (PM_i - MM_i^*)$,
with w_i Tukey biweight from initial fit
- Tukey biweight: $w_i = (1 - (r_i/c^2)^2)^2$ if $|r_i| \leq c$;
= 0 otherwise

% 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 ~ 20 - 50%
- ‘good indicator of assay performance’
- similar values across replicates (also SF, RawQ)



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



Problems with these measures

- Relate to the experimental process, *not* directly to the end result (gene expression)
- Quality of spike-in data may not be representative of whole chip quality
- In general, thought, inferences (DE, clustering, etc.) are based on ME
- *Single chip* measures, which do not put each chip in the context of the others
- *By-products of RMA calculation (robust regression) can also provide quality info*



What is ‘quality’?

- It is useful to distinguish between the various facets of the general term ‘*quality*’
- In chronological order:
 - condition of the starting RNA (*RNA integrity*)
 - caliber of the experimental process and resulting hybridization (*noise*)
 - acceptability of the resulting expression measures:
 - *array adjustment*
 - *outlier identification*



New quality measures - RMA-QC

- Aims:
 - To use QA/QC measures directly based on expression summaries and that can be used in a routine way
 - To examine whether chips are different in a way that affects expression summaries
- Focus on *weights* and *residuals* from fits in probe intensity models



RMA - Additive model for gene expression based on probe intensity data

- *Probe-level model* for gene expression:

$$\log_2(\text{PM}^*_{ij}) = c_i + p_j + \varepsilon_{ij}$$

- c_i = \log_2 scale expression level for chip j
- p_j = probe affinity effect
- ε_{ij} = iid error term

- For *identifiability*, fit with constraint $\sum_j p_j = 0$
- Model fit (separately) for *each probe set*



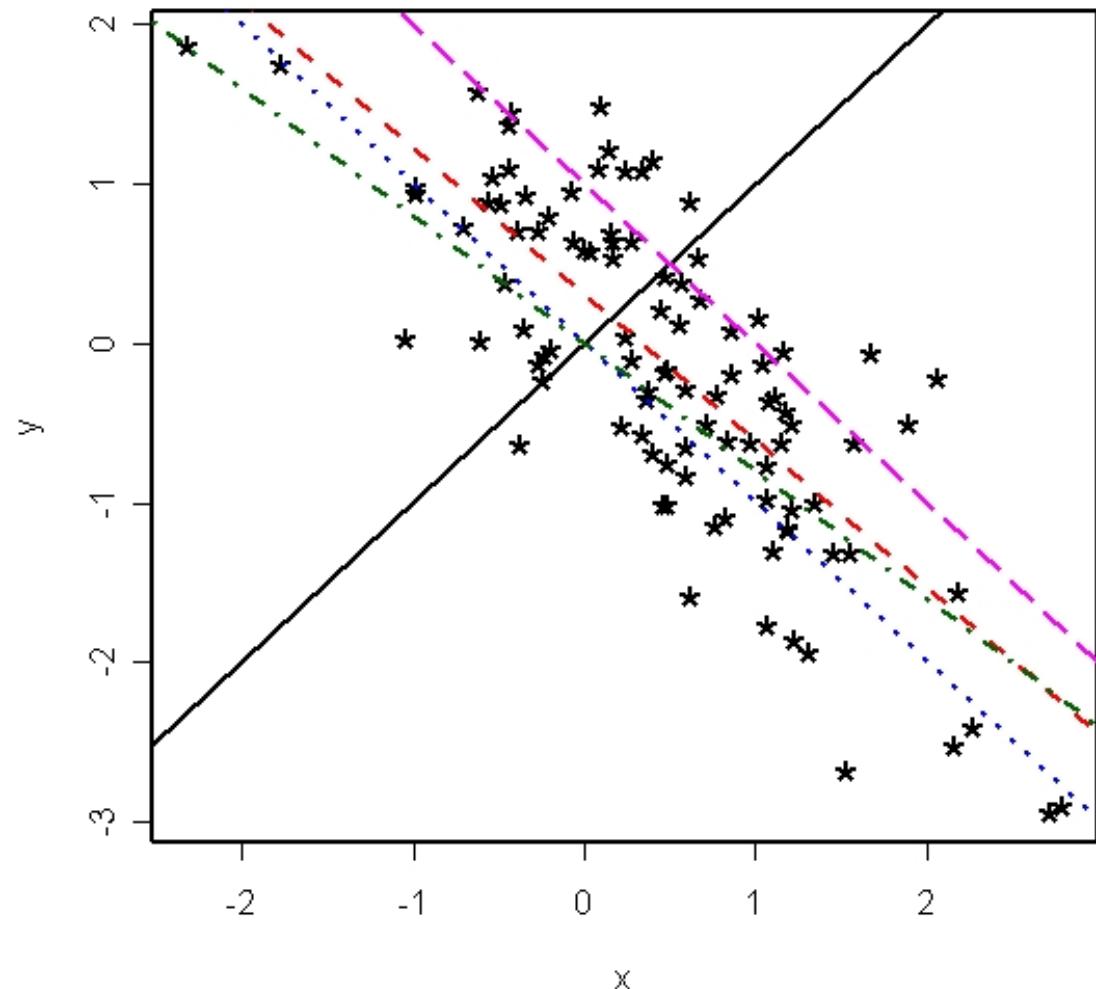
RMA: Summary

- Chips analysed in *sets* (e.g. an entire experiment)
- *Use only PM*, ignore MM
- *Background* correct PM on raw intensity scale
- *Quantile Normalization* of $\log_2(\text{PM}-\text{BG})$
- Assume additive model (on \log_2 scale) for each probeset: $\log_2 \text{normalized}(\text{PM}_{ij} - \text{BG}) = c_i + p_j + e_{ij}$
- Parameters c_i provide measure of gene expression for each chip
- Estimate parameters using a *robust* method
 - median polish - quick
 - robust linear model - yields quality diagnostics



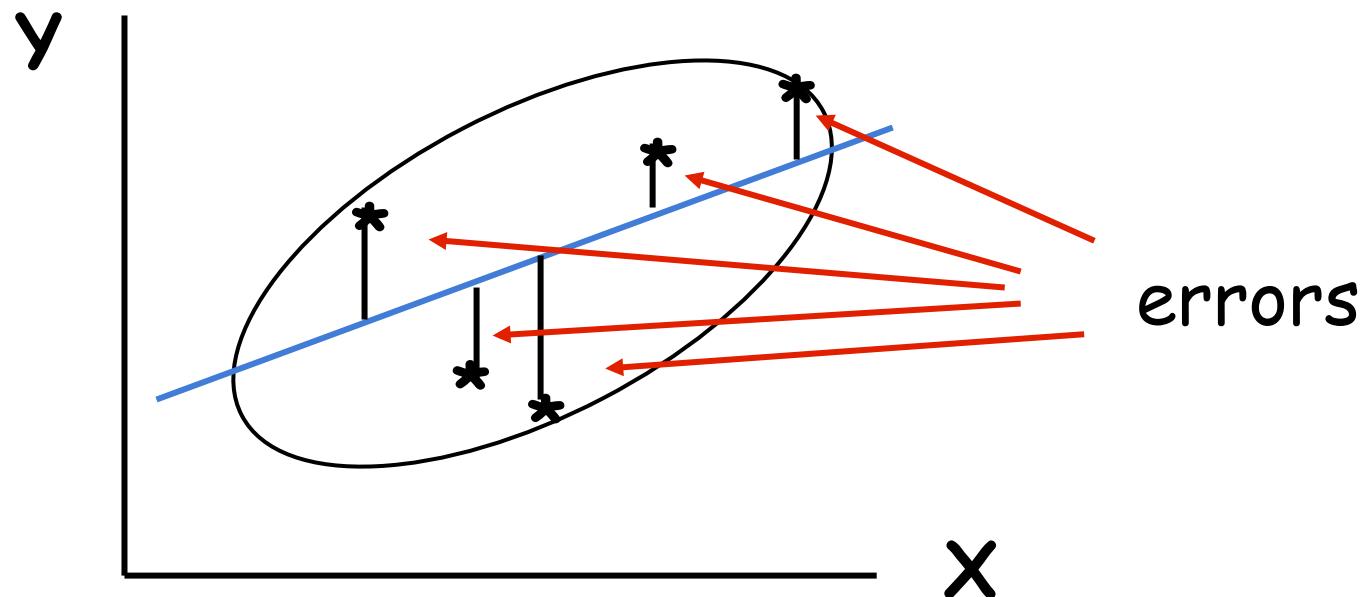
Simple linear modeling: which line?

- There are *many possible lines* that could be drawn through the cloud of points in the scatterplot ...
- How to choose?



Least Squares

- **Q:** Where does the regression equation come from?
A: It is the line that is ‘best’ in the sense that it *minimizes* the sum of the *squared* errors (residuals) in the vertical (*y*) direction



What is robustness?

- The term *robustness* is used to mean several possible things:
 - Lack of sensitivity to *distributional assumptions* (especially normality)
 - Lack of sensitivity to *outliers*
 - Small sets of the data *don't have a strong influence*



Why robust (vs. LS)?

- Want fitting procedure to produce good estimates in the presence of various types of outliers:
 - *probe outliers*: e.g. probes that ‘don’t work’
 - *chip outliers*: chips that are unusual
 - Image artifacts
- Want procedure to assess quality
- Distinguish between approach based on *outlier identification/exclusion* and approach based on *modeling / quality weights*



Median polish algorithm

y_{11}	L	y_{1J}	0
M	O	M	M
y_{I1}	L	y_{IJ}	0
0	L	0	0

Sweep Columns

Iterate

Sweep Rows

Imposes
Constraints

$$\text{median}_i e_{ij} = \text{median}_j e_{ij} = 0$$

$\hat{\varepsilon}_{11}$	L	$\hat{\varepsilon}_{1J}$	$\hat{\alpha}_1$
M	O	M	M
$\hat{\varepsilon}_{I1}$	L	$\hat{\varepsilon}_{IJ}$	$\hat{\alpha}_I$

$\hat{\beta}_1$	L	$\hat{\beta}_J$	\hat{m}
-----------------	---	-----------------	-----------



Median polish - example

	1	1	2	3	3	0
	2	4	5	7	5	0
	3	3	6	6	7	0
probe	2	3	5	6	5	array

	-1	-2	-3	-3	-2	-2
	0	1	0	1	0	0
	1	0	1	0	2	1
probe						array

	1	0	-1	-1	0	
	0	1	0	1	0	
	0	-1	0	-1	1	
probe	0	0	0	-1	0	array



(BREAK)



Robust regression

- Idea: *downweight* observations that produce large residuals
- More *computationally intensive* than least squares regression (which gives equal weight to each observation)
- Use *maximum likelihood* if can assume specific error distribution
- When not, use *M-estimators*



Robust regression in microarray analysis

- There are many ways that robust regression can be/is used in analysis of microarray data
- We will use it in two ways:
 - for *quantifying gene expression* measured with Affymetrix GeneChips (like we saw with RMA)
 - for *assessing quality* of Affymetrix GeneChip gene expression measures (coming up next)



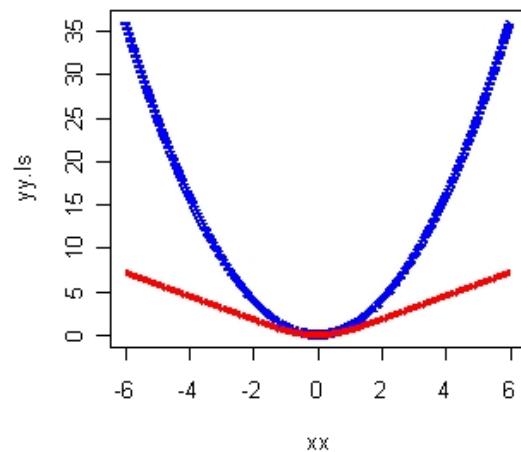
Loss, weight functions

- Least squares: ‘lose’ square of vertical error
- Here, squared error = *loss function*
- Each observation has *equal weight*
- Problem: *outliers* can have strong effect on estimates (slope, intercept of line; model parameters more generally)
- Solution: could use *other loss/weight functions*

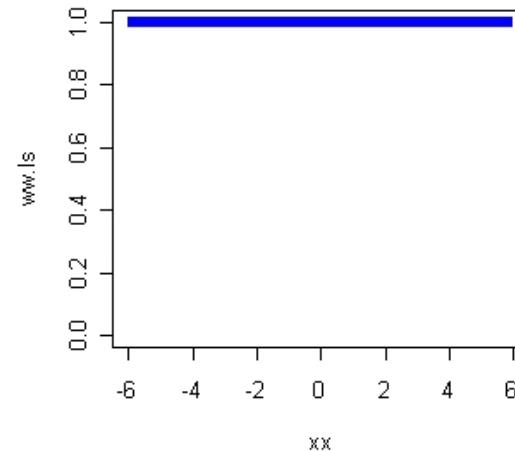


Examples of Loss, Weight Functions

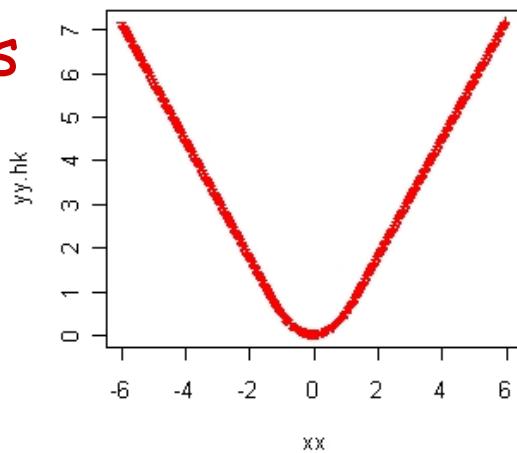
Squared
error loss



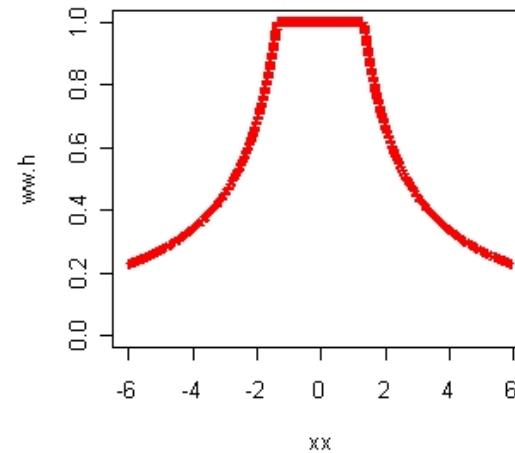
Equal
weight



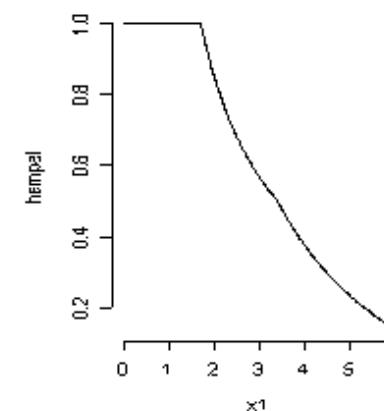
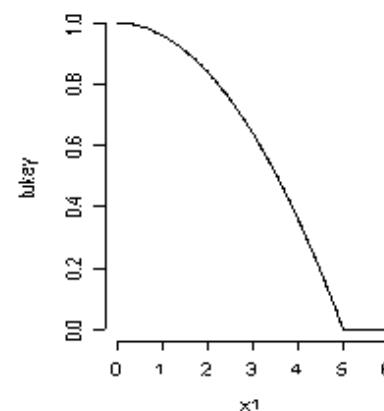
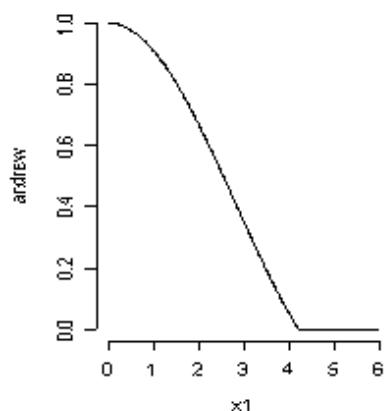
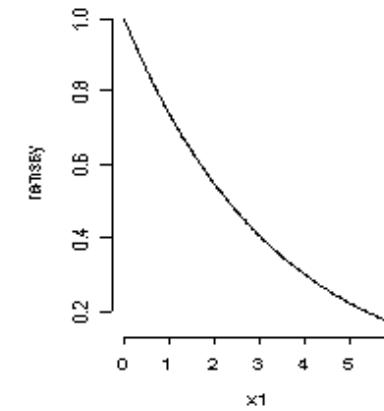
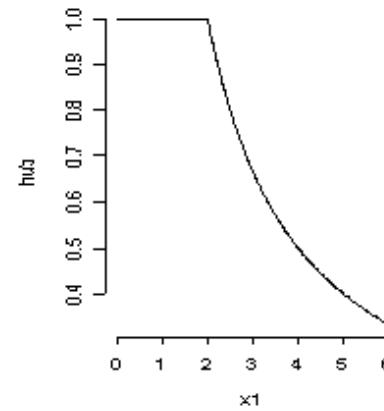
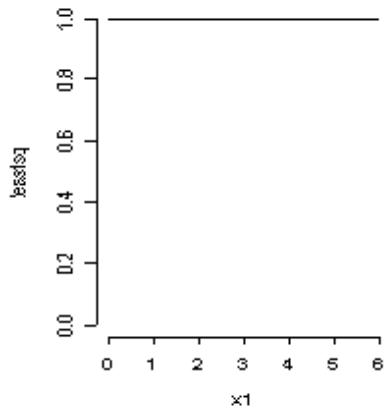
Huber loss



Huber
weights



More weight functions



Robust regression estimation

- Robust procedures *perform well* under a range of possible models
- Facilitates *outlier detection*
- Good estimates even if some bad data points
- Can identify `bad' probe behavior:
 - some probes may cross hybridize to non-target fragments
 - some may not bind at all to target fragment



M-estimators

- ‘Maximum likelihood type’ estimators
- Assume independent errors with distribution $f(\varepsilon)$
- Parameter estimates solutions to

$$\min_{p_i, c_j} \sum_{i,j} \rho\left(\frac{Y_{ij} - p_i - c_j}{\hat{\sigma}}\right) = \min_{p_i, c_j} \sum_{i,j} \rho(u_{ij})$$

- $\rho(x)$ is a (bounded for robustness) positive, symmetric function increasing more slowly than x
- $\hat{\sigma}$ is an estimate of scale (eg. MAD)
- eg, $\rho(u) = u^2$ corresponds to minimizing the sum of squares



M-estimation procedure

- To minimize $\sum_i \rho\{(\mathbf{y}_i - \mathbf{x}_i' \boldsymbol{\beta})/s\}$ wrt the β 's, take derivatives and equate to 0 ('normal equations')
- Resulting equations *do not have an explicit solution* in general
- Solve by *iteratively reweighted least squares* (IRLS) with weights
$$w_{ij} = \rho'(u_{ij})/u_{ij} = \psi(u_{ij})$$
- Acts like *automatic outlier rejector*, since large residual values lead to very *small weights*



IRLS algorithm

- *Weights* at each iteration are calculated by applying the loss function to the residuals obtained from the previous iteration
- The weight function gives *lower weight* to points that do not fit well ('outliers')
- The results are *less sensitive* to outliers in the data (compared to OLS)



Robust fit by IRLS for each probe set

- Use *Huber* loss function ρ :
 - $\rho(e) = e^2/2$ for $|e| \leq k$; $k|e| - k^2/2$ for $|e| > k$
- Starting with robust (or LS) fit, at each iteration:
 - $r_{ij} = Y_{ij} - \text{current est}(p_j) - \text{current est}(c_i)$
 - $S = \text{mad}(r_{ij}) \cdot c$ - robust est. of scale of σ
 - $u_{ij} = r_{ij}/S$ - rescaled residuals
 - $w_{ij} = \psi(|u_{ij}|)/|u_{ij}|$ - weights used in next fit
(for Huber loss, $w = 1$ if $|u| \leq k$; $k/|u|$ if $|u| > k$)
- Next step estimates obtained by (weighted) LS



Quality Assessment using PLM

- PLM = Probe Level Model
- PLM quantities useful for assessing *chip quality* (*expression measure*)
 - Weights
 - Residuals
 - Standard Errors (NUSE)
- Expression values relative to (virtual) ‘median’ chip
(RLE = Relative Log Expression)



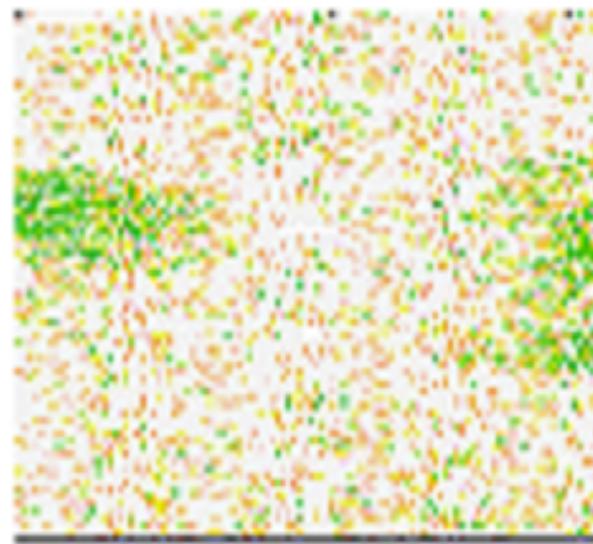
Role of model components in QA/QC

- Residuals, weights - now >200K per array
 - *summarize* to produce a chip index of quality
 - view as chip *image*, analyse spatial patterns.
 - scale of residuals for probe set models can be *compared* between experiments
- Chip effects > 20K (probe sets) per array
 - can examine distribution of relative expressions across arrays
- Probe effects > 200K per model (HG_U133A)
 - can be compared across fitting sets



Chip weight pseudo-images

- Image indicates the (robust regression) *weight* associated with the probe
- Areas of *low weight* (outliers) are greener, *high weights* are light gray
- ‘More color’ \Leftrightarrow ‘worse chip’ (more of an outlier)

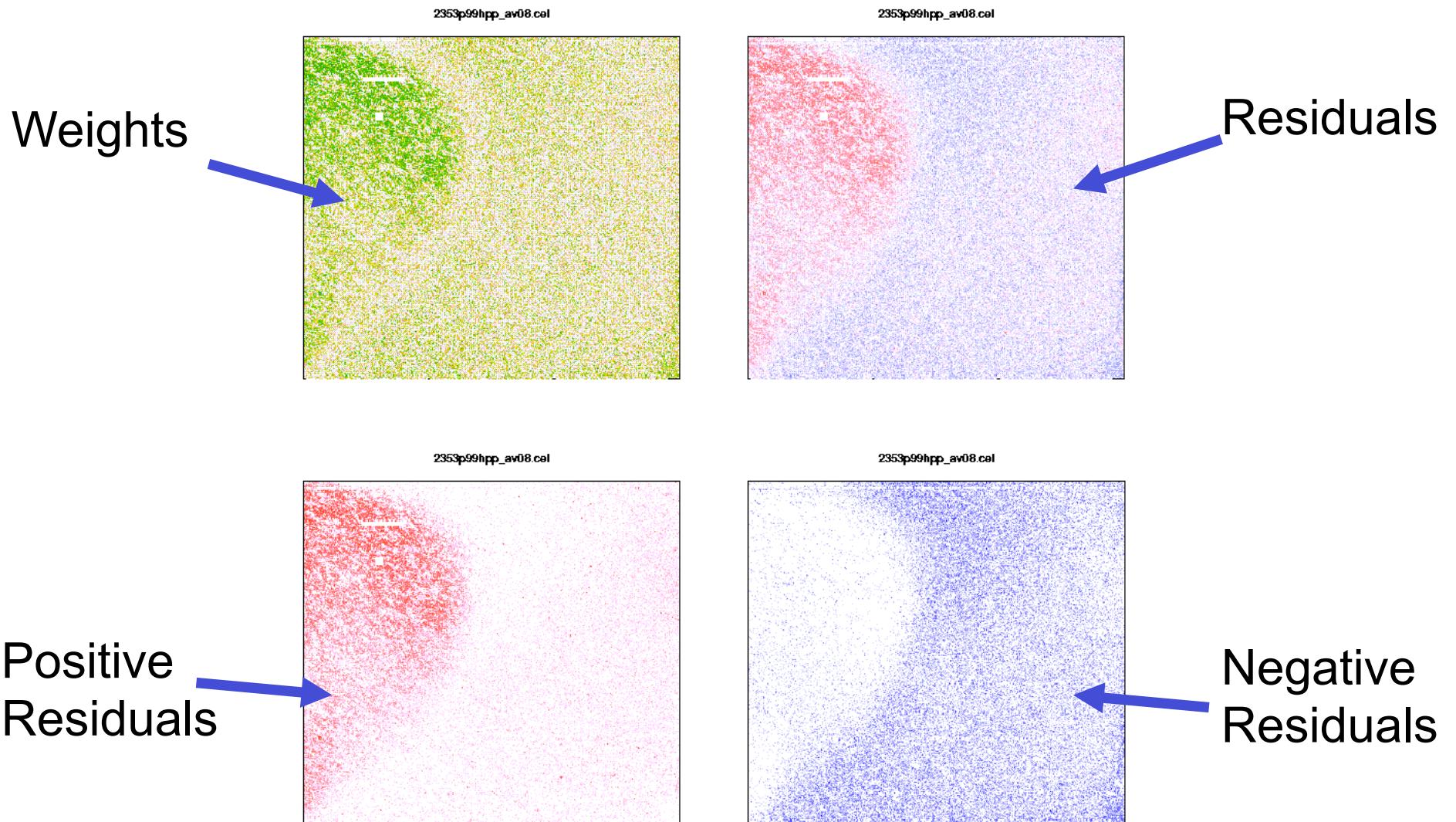


Using residuals from the fitting

- Many types of *problems* will be reflected by *inflated residuals* from the fits to the probe + chip effect models
- *Summarizing the residuals* on a chip can provide good discrimination among chips producing data of varying quality



Pseudo-chip images



Chip index of relative quality

- We assess gene expression index (eg, RMA value) variability for gene (probe set) k ($=1, \dots, G$ genes) by its *unscaled SE* (j indexes probes):

$$\text{unscaled SE}(\hat{c}_{ki}) = 1/\sqrt{\sum_j w_{kij}}$$

- We then *normalize* by dividing by the *median* unscaled SE over the chip set (i):

$$\text{NUSE}(\hat{c}_{ki}) = \frac{1/\sqrt{\sum_j w_{kij}}}{\text{median}_i(1/\sqrt{\sum_j w_{kij}})}$$

NUSE

- NUSE = ‘Normalized Unscaled SE’ - estimate SE(expression estimates), summarize at the chip level
- Each chip will have a NUSE for each probe set, which can be summarized by the *median*
- This provides one useful *summary* of the residual variability, and can be used to judge *quality* relative to other chips
- Median NUSE fluctuates around 1
- High values (> 1.05) indicate ‘worse’ chips (unusual / outliers)



RLE

- How much are robust summaries affected?
- Can gauge reproducibility of expression measures by summarizing the distribution (across genes) of *relative log expressions*
- $RLE_i = RMA_i - \text{reference expression}_i$ ($i = 1, \dots, p$)
- For reference expression, can use *median expression value* for that gene in a set of chips
- This provides one useful summary of the residuals, and can be used to judge quality relative to other chips



RLE summaries

- IQR(RLE) measures variability
- Includes *Noise + DE* in biological replicates
- When biological replicates are similar (eg. RNA from same tissue type), can typically detect *processing effects* with IQR(RLE)
- Median(RLE) should be close to zero if
 # up-regulated genes \approx down-regulated genes
- Can combine $IQR(RLE) + |\text{Median}(RLE)|$ to give measure of chip expression measurement error



Example: HD

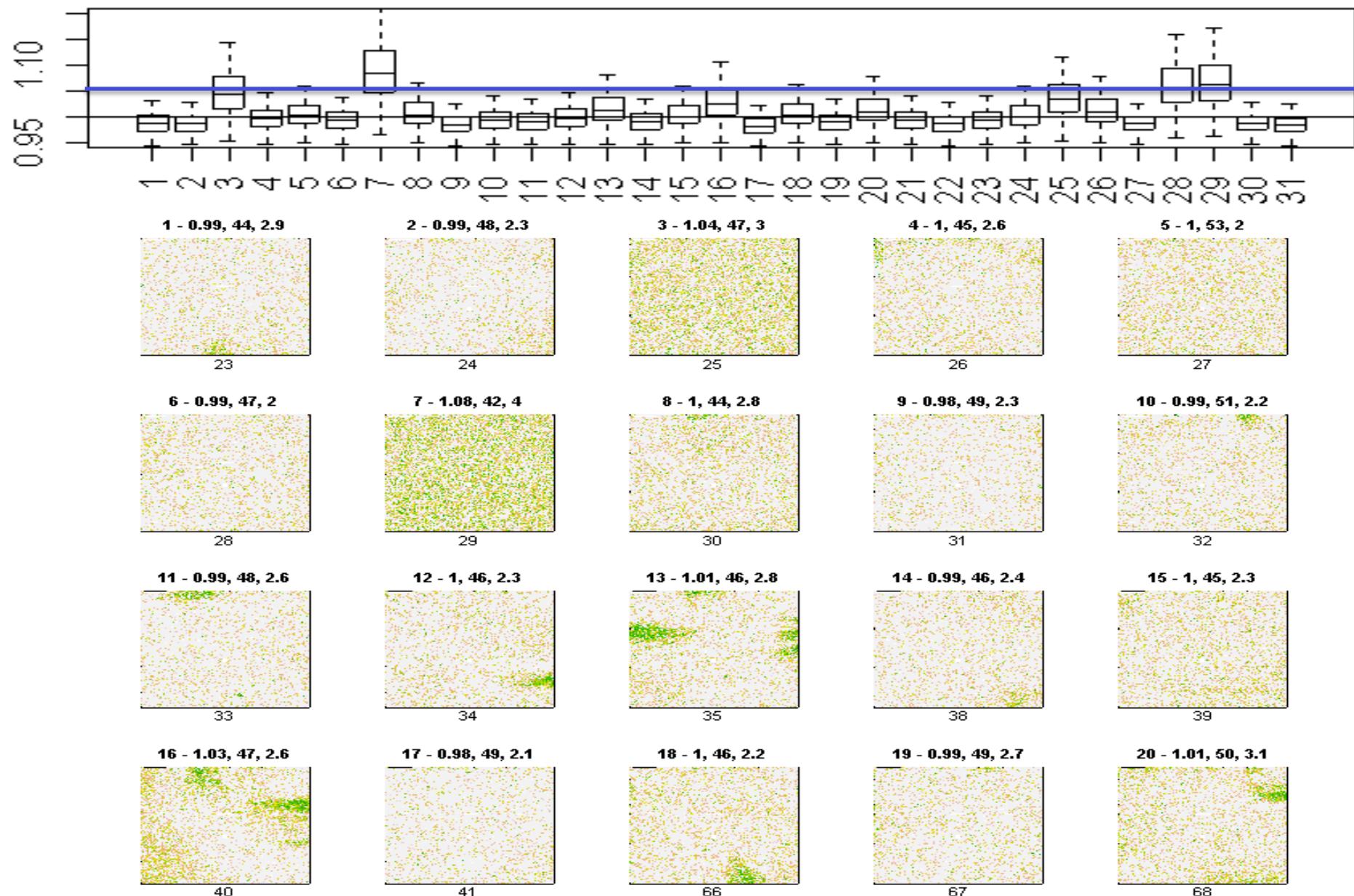
- About 70 individuals, U133A,B chips on each of 3 tissues
- Fitted RMA models
- Displays: NUSE plot, chip pseudo-image of residual weights

Title = Chip Number - Median NUSE, %P, SF

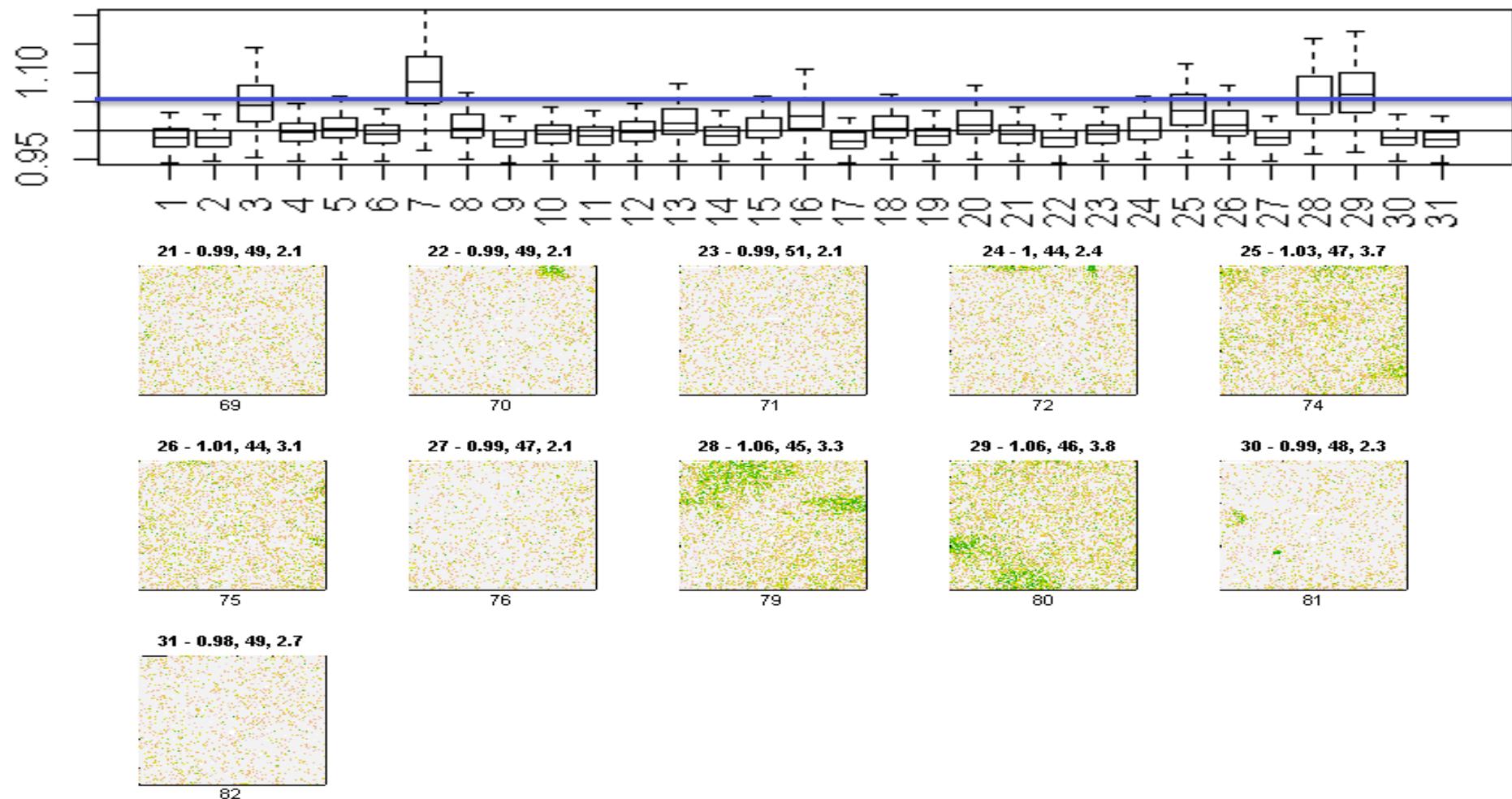
Subtitle = ChipId

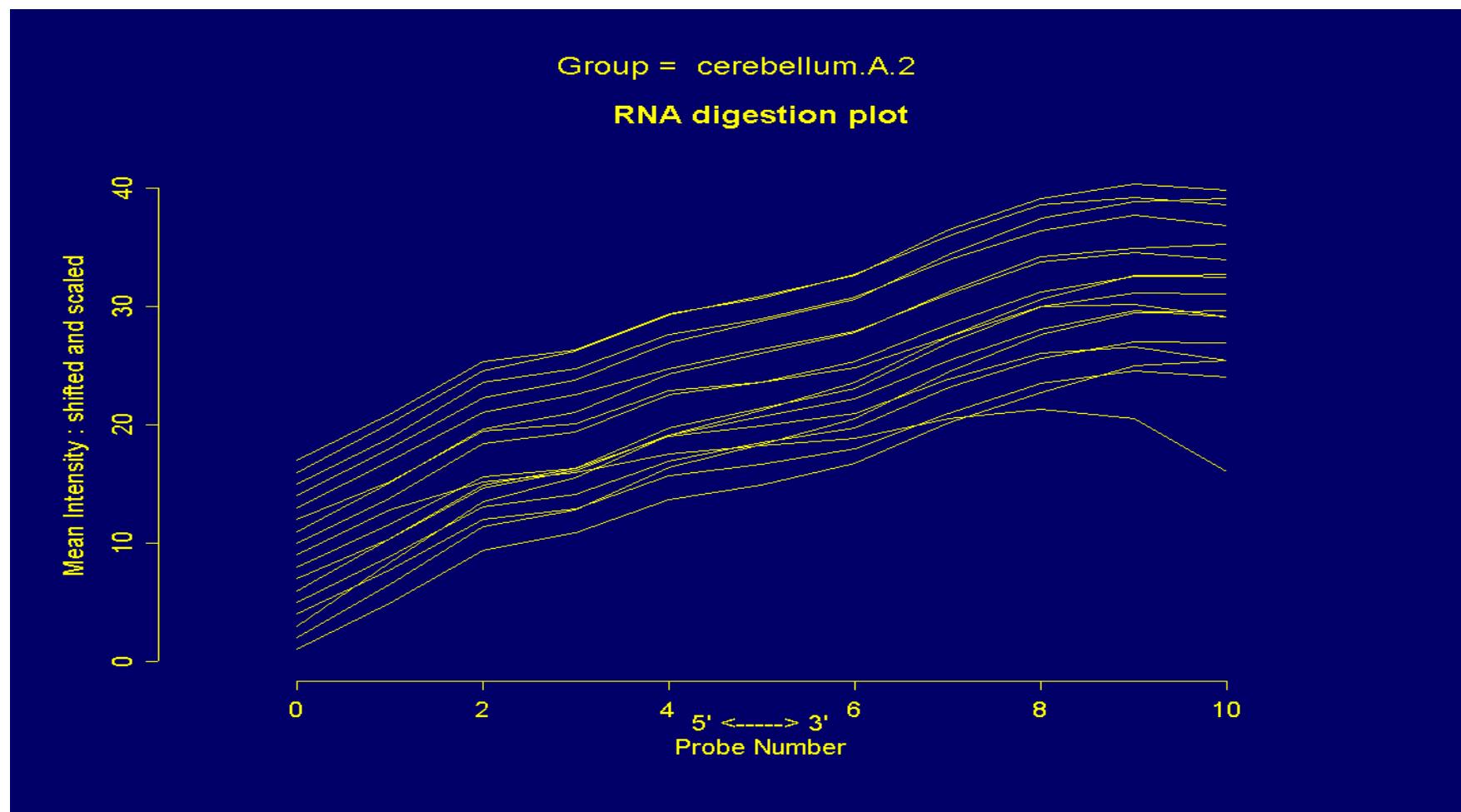
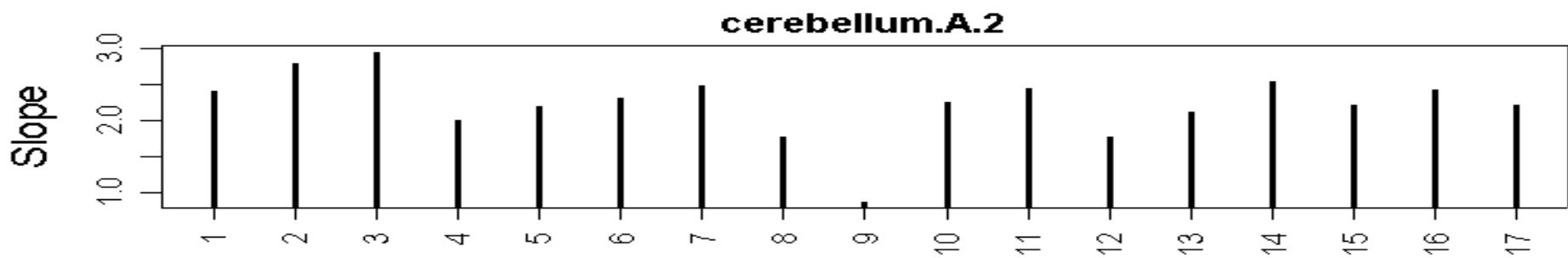


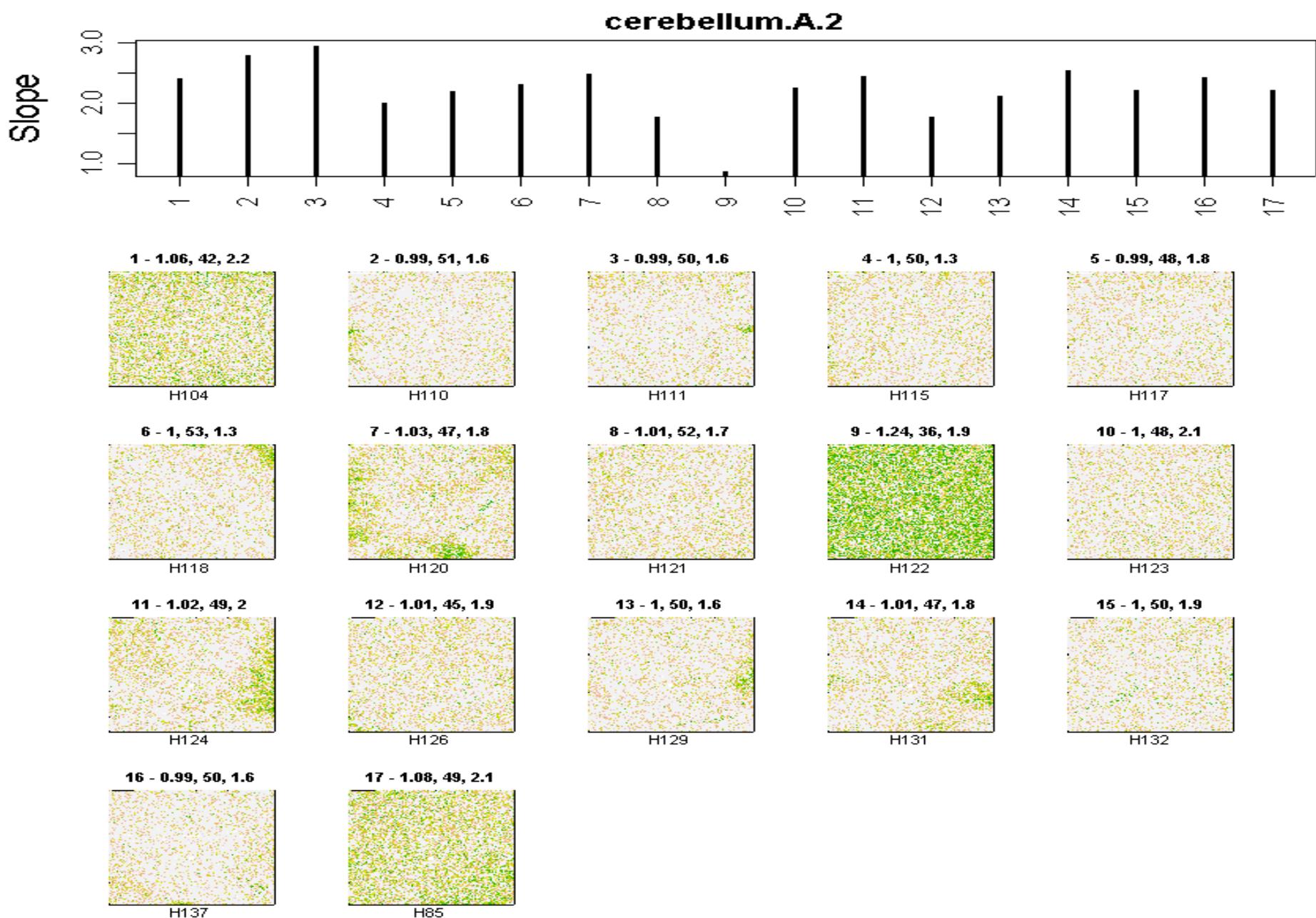
F. cerebellum.A.1 P. cerebellum.A.1 - Boxplots of NUSE values



F. cerebellum.A.1 P. cerebellum.A.1 - Boxplots of NUSE values







Measuring quality

- Different measures view quality from different (but overlapping) perspectives
- Affymetrix measures (.rpt file) are most prominent in the *noise* and *integrity* aspects, but also touch on *array adjustment*
- RMA-QC measures dominate in *outlier identification*, but also include *array adjustment*



Conclusions

- PLM-based quality assessment appears to show good sensitivity to chip problems that impact *measures of expression*
- Provides useful basis for chip quality, inclusion/exclusion decisions
- RMA-QC measures implemented in the **affyPLM** package (BioConductor)
- **affyPLM** documentation gives more details of estimation procedure
- <http://plmimagegallery.bmbolstad.com/>



Exploratory data analysis/quality assessment

- PM signal intensity:
 - pseudo-images
 - histograms
 - boxplots
 - pairwise scatterplots (MA version)
- Pseudo-images of *weights* and *residuals*
- Boxplots of *NUSE values*
- Boxplots of *RLE values*
- Boxplots of normalized signal values (RMA)

