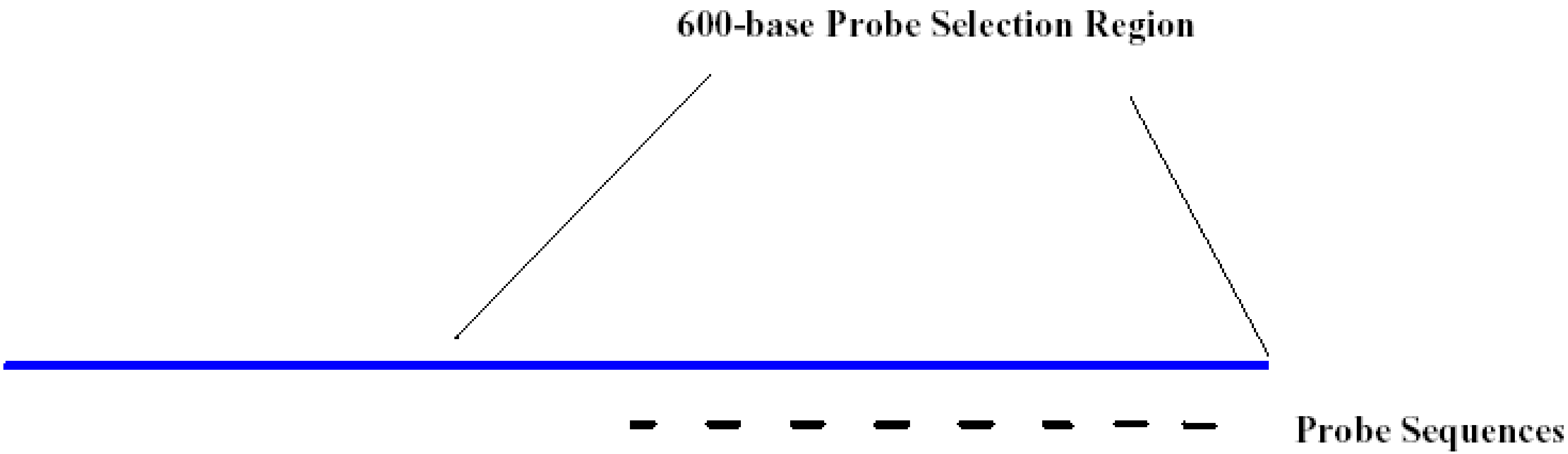
Assessing gene expression quality in Affymetrix microarrays

Outline

- The Affymetrix platform for gene expression analysis
- Affymetrix recommended QA procedures
- The RMA model for probe intensity data
- Application of the fitted RMA model to quality assessment

The Affymetrix platform for gene expression analysis

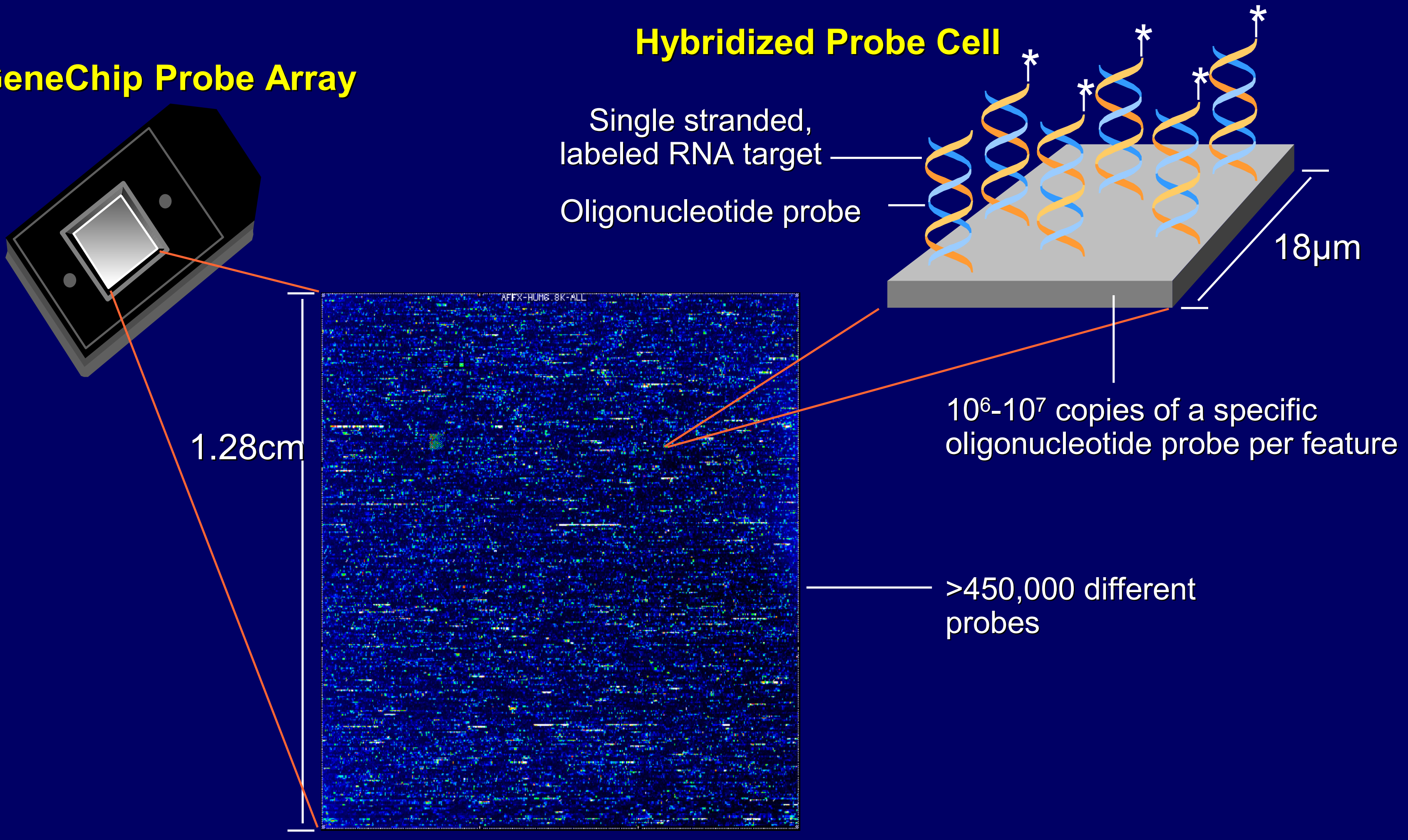
Probe selection



Probes are 25-mers selected from a target mRNA sequence.

5-50K target fragments are interrogated by probe sets of 11-20 probes. Affymetrix uses PM and MM probes

Oligonucleotide Arrays



GeneChip Probe Array

Hybridized Probe Cell

Single stranded, labeled RNA target

Oligonucleotide probe

18µm

1.28cm

10^6-10^7 copies of a specific oligonucleotide probe per feature

>450,000 different probes

Image of Hybridized Probe Array

Compliments of D. Gerhold

Obtaining the data

- RNA samples are prepared, labeled, hybridized with arrays, arrrays are scanned and the resulting image analyzed to produce an intensity value for each probe cell (>100 processing steps)
- Probe cells come in (PM, MM) pairs, 11-20 per probe set representing each target fragment (5-50K)
- Of interest is to analyze probe cell intensities to answer questions about the sources of RNA – *detection of mRNA*, *differential expression assessment*, *gene expression measurement*

Affymetrix recommended QA procedures

Pre-hybe RNA quality assessment

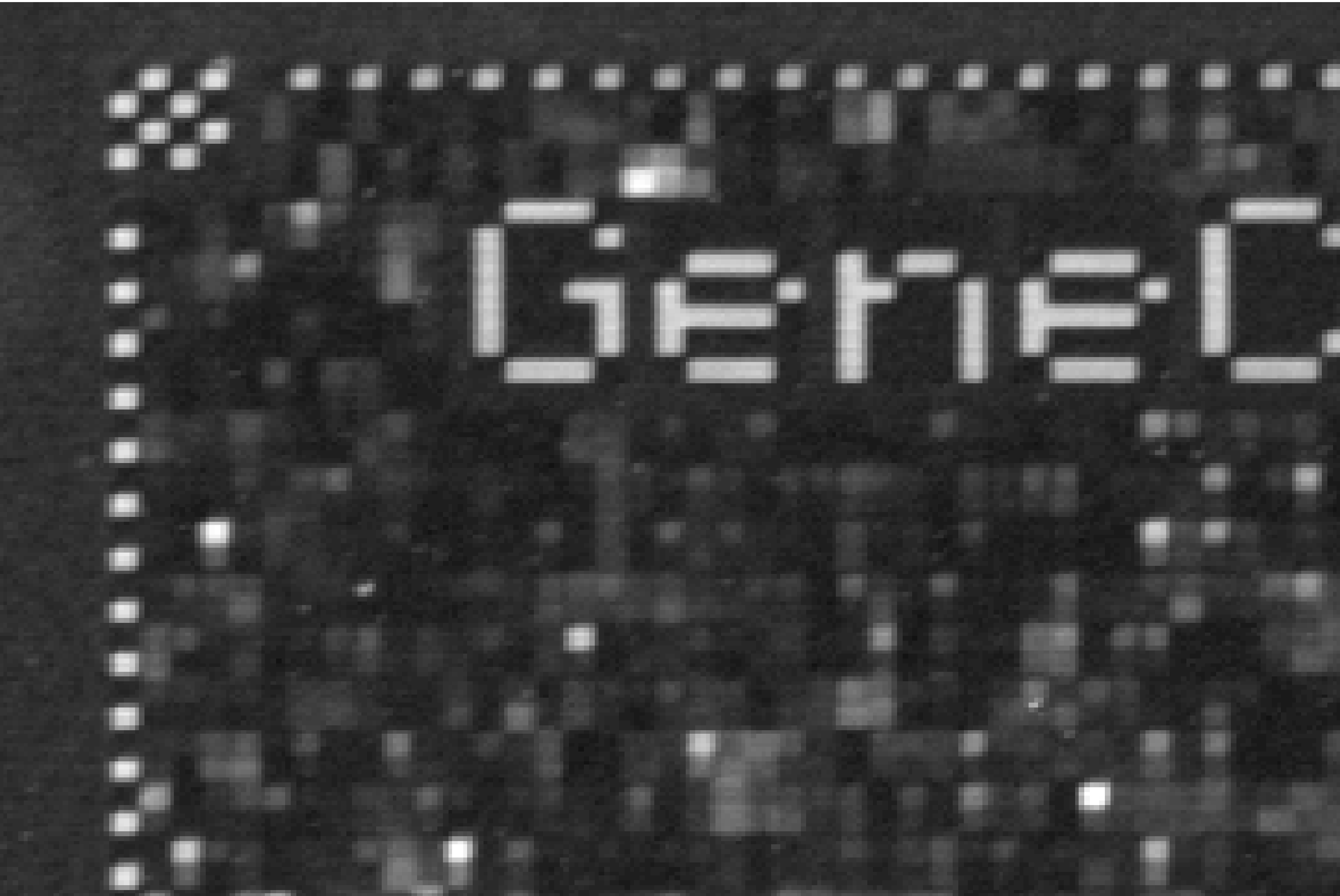
Look at gel patterns and RNA quantification to determine hybe mix quality.

QA at this stage is typically meant to preempt putting poor quality RNA on a chip, but loss of valuable samples may also be an issue.

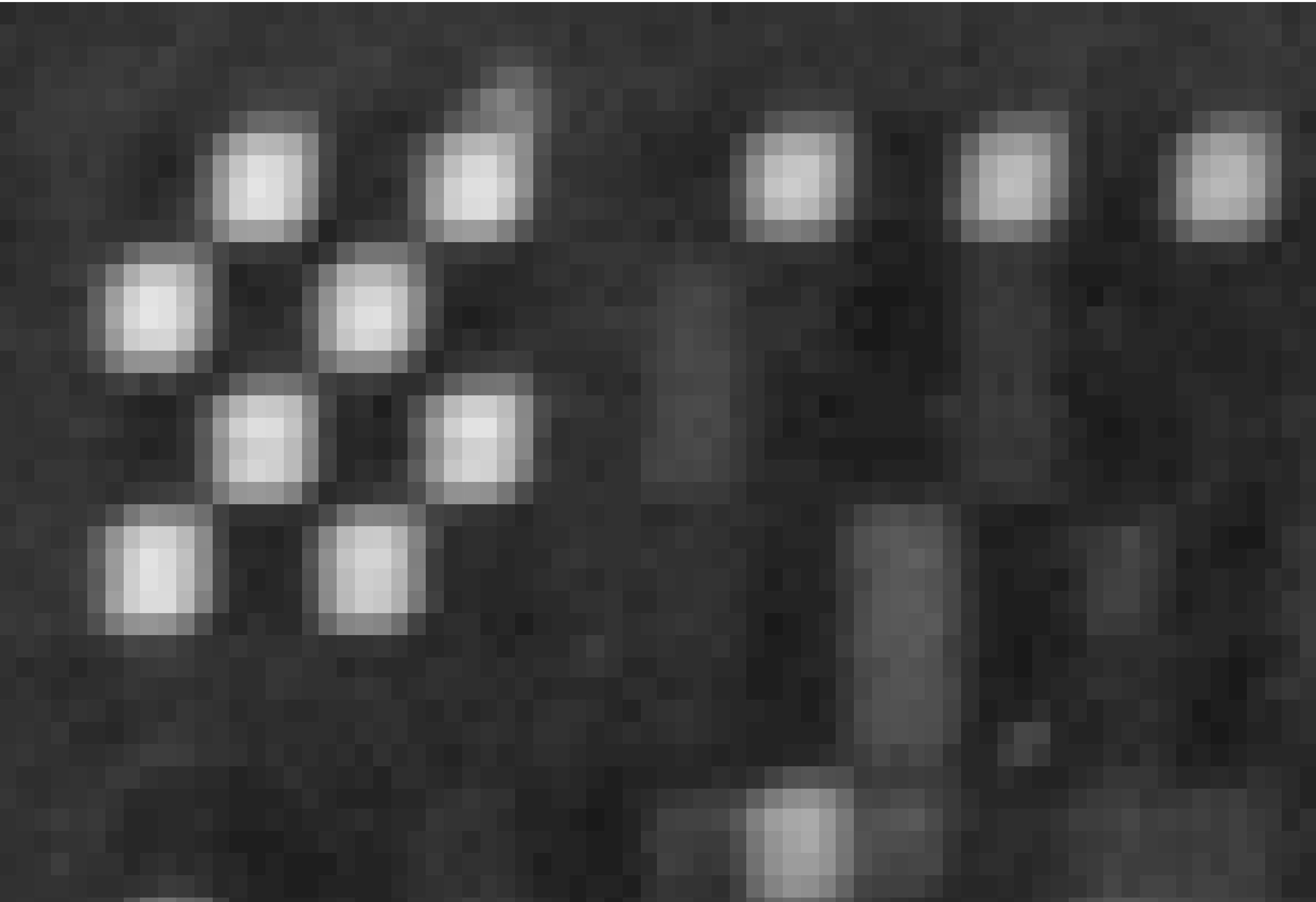
Post-hybe QA: Visual inspection of image

- **Biotinylated B2 oligonucleotide hybridization**: check that checkerboard, edge and array name cells are all o.k.
- **Quality of features**: discrete squares with pixels of slightly varying intensity
- **Grid alignment**
- **General inspection**: scratches (ignored), bright SAPE residue (masked out)

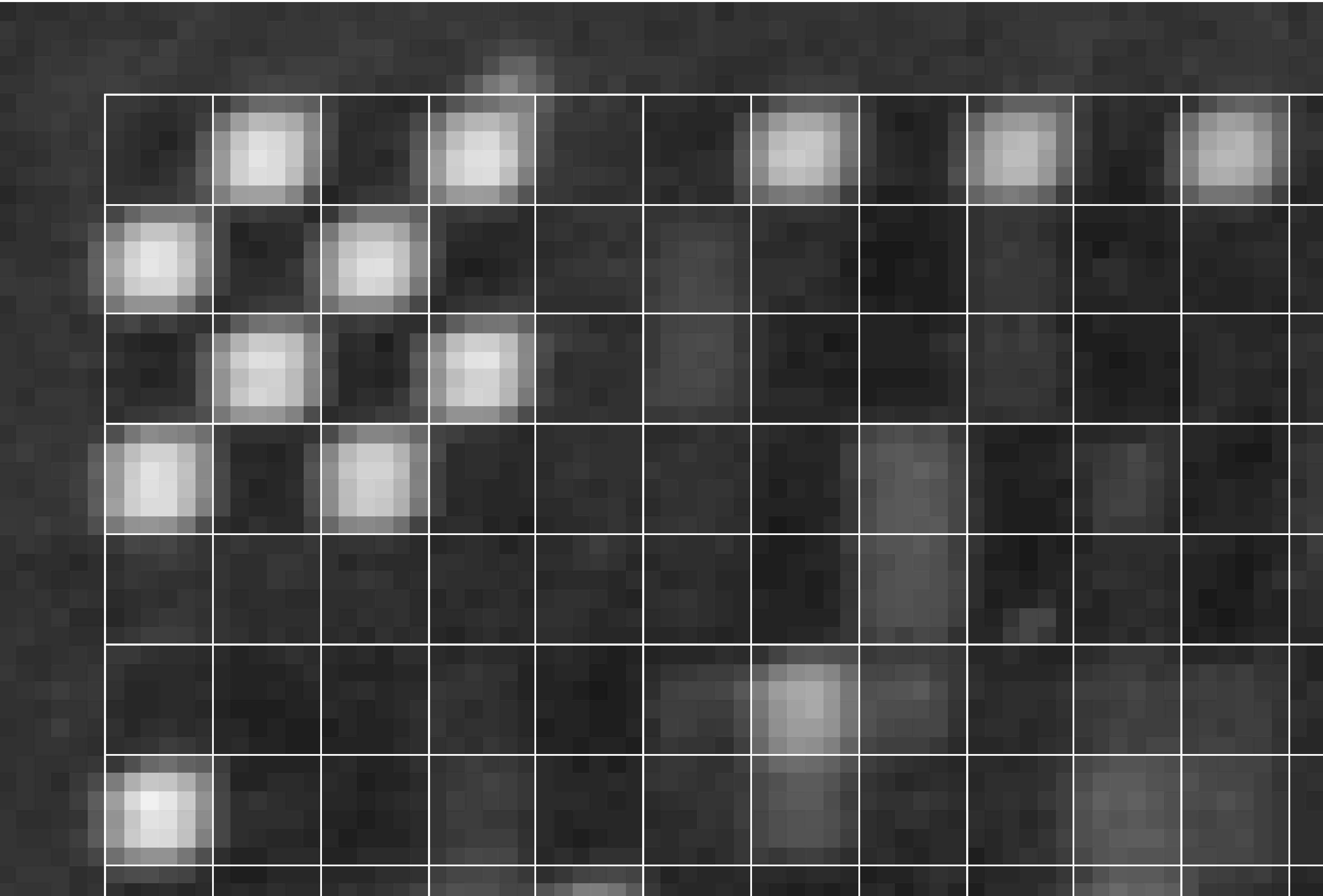
Checkerboard pattern



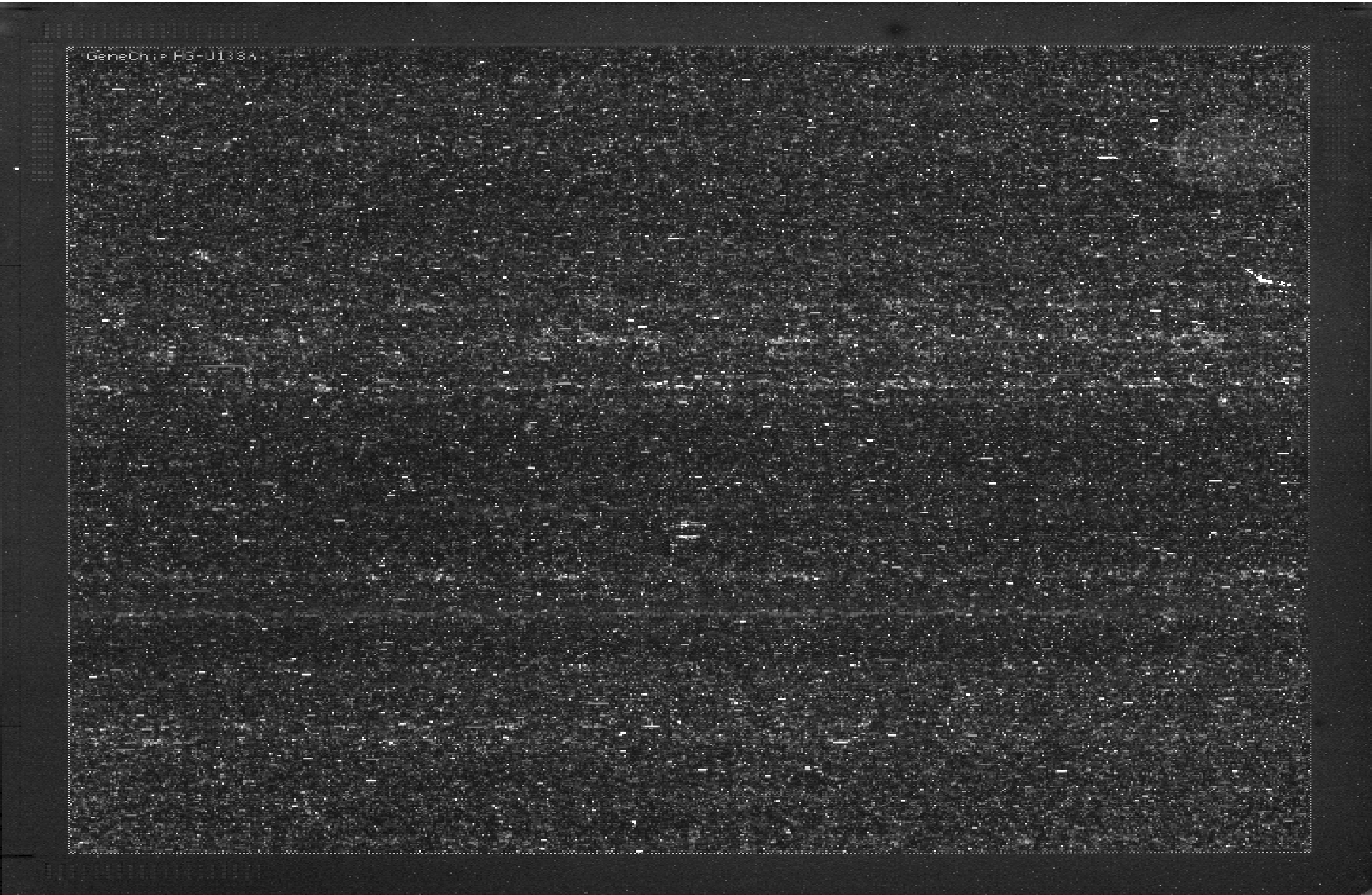
Quality of featutre



Grid alignment



General inspection



MAS 5 algorithms

- Present calls: from the results of a Wilcoxon's signed rank test based on:

$$(PM_i\text{-}MM_i)/(PM_i+MM_i)\text{-}\tau$$

for small τ (~.015). ie. PM-MM > τ*(PM+MM)?

- Signal:

$$\log(\mathit{Signal}) = \sum_i w_i \log(PM_i - MM_i^*)$$

where w_i is Tukey biweight from initial fit.

Post-hybe QA: Examination of quality report

- **Percent present calls** : Typical range is 20-50%. Key is consistency.
- **Scaling factor**: Target/(2% trimmed mean of Signal values). No range. Key is consistency.
- **Background**: average of of cell intensities in lowest 2%. No range. Key is consistency.
- **Raw Q (Noise):** Pixel-to-pixel variation among the probe cells used to calculate the background. Between 1.5 and 3.0 is ok.

Examination of spikes and controls

- **Hybridization controls**: *bioB, bioC, bioD* and *cre* from *E. coli* and *P1* phage, resp.
- **Unlabelled poly-A controls**: *dap, lys, phe, thr, tryp* from *B. subtilis*. Used to monitor wet lab work.
- **Housekeeping/control genes**: GAPDH, Beta-Actin, ISGF-3 (STAT1): 3' to 5' signal intensity ratios of control probe sets.

How do we use these indicators for identifying bad chips?

We illustrate with 17 chips from a large publicly available data set from St Jude’s Children’s Research Hospital in Memphis, TN.

Hyperdip_chip A - MAS5 QualReport

	Noise	Background	ScaleFactor	% Present	GAPDH 3'/5'	BetaActin 3'/5'
Hyperdip>50-#12	5.55	119.1	10.98	0.38	0.99	1.47
Hyperdip>50-#14	3.79	91.25	6.35	0.44	1.18	1.76
Hyperdip>50-#8	2.23	75.89	29.64	0.28	0.86	1.33
Hyperdip>50-C1	3.06	70.03	8.4	0.4	1.05	1.64
Hyperdip>50-C11	1.76	58.04	20.39	0.37	0.87	1.34
Hyperdip>50-C13	3.35	78.77	8.09	0.42	0.97	1.62
Hyperdip>50-C15	3.06	77.15	11.39	0.37	1.13	1.98
Hyperdip>50-C16	1.34	54.05	33.33	0.31	0.94	1.49
Hyperdip>50-C18	1.35	52.18	28.49	0.34	1.49	2.92
Hyperdip>50-C21	1.43	56.89	29.48	0.34	1.29	2.55
Hyperdip>50-C22	1.24	52.75	41.17	0.31	1.01	2.87
Hyperdip>50-C23	1.35	46.69	26.96	0.36	1.07	2.57
Hyperdip>50-C32	1.95	65.86	16.21	0.38	0.86	1.37
Hyperdip>50-C4	1.6	60.11	22.57	0.34	1.17	2.61
Hyperdip>50-C6	2.42	60.73	8.18	0.4	1.39	2.38
Hyperdip>50-C8	3.01	75.65	8.56	0.4	0.91	1.57
Hyperdip>50-R4	1.36	48.19	36.34	0.29	2	3.95

#12 bad in Noise, Background and ScaleFactor

#14? #8? C1? C11? C13-15? C16-C4? C8? R4?

Only C6 ***passes*** all tests. Conclusion?

Limitations of Affymetrix QA/QC procedures

- Assessments are based on features of the arrays which are only indirectly related to numbers we care about – the gene expression measures.

- The quality of data gauged from spike-ins requiring special processing may not represent the quality of the rest of the data on the chip. We risk **QCing the chip QC** process itself, but not the gene expression data.

New quality measures

Aim:

- To use QA/QC measures directly based on expression summaries and that can be used routinely.

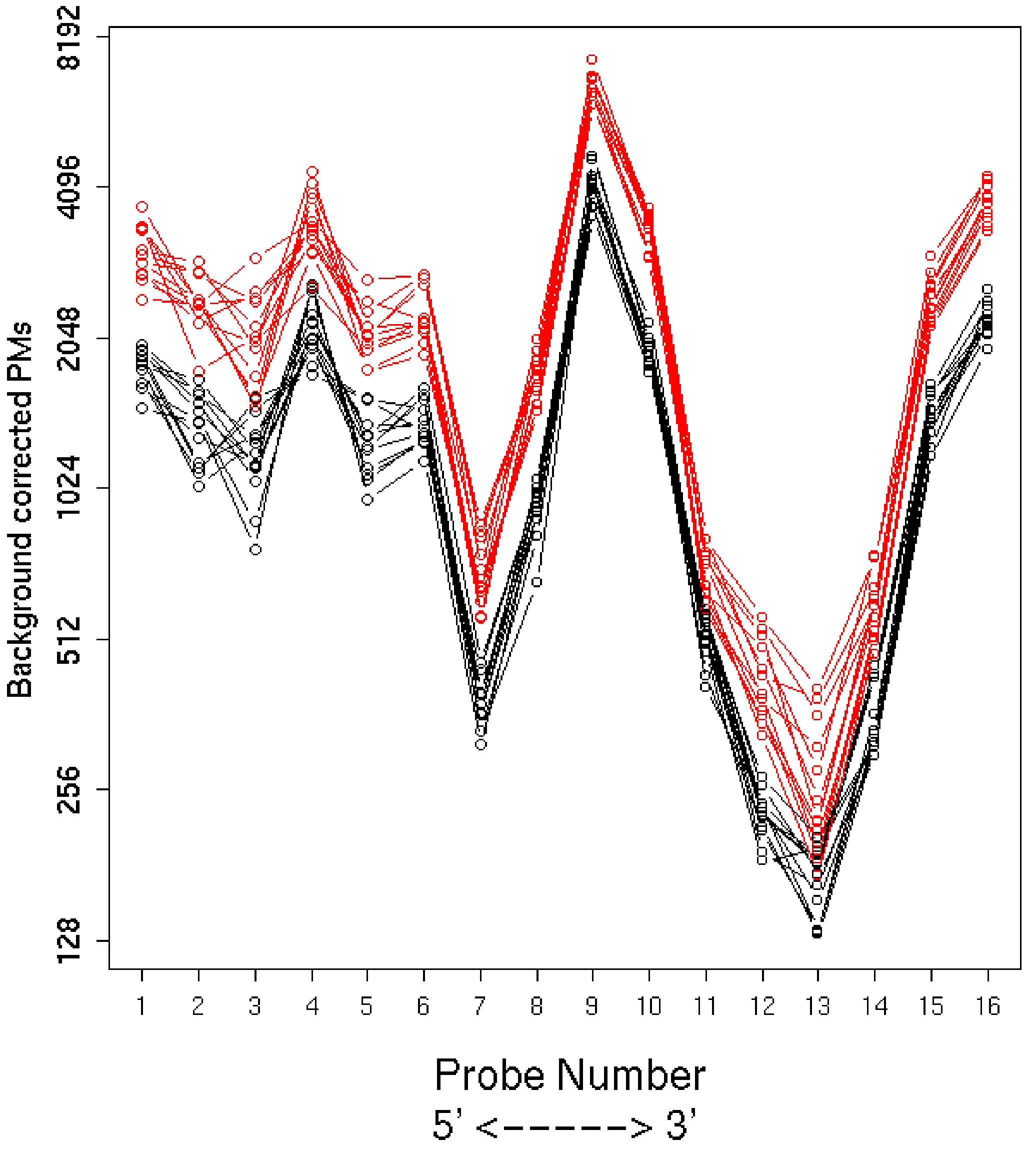
To answer the question “are chips different in a way that affects expression summaries?” we focus on residuals from fits in probe intensity models.

The RMA model for probe intensity data

Summary of Robust Multi-chip Analysis

- Uses only PM values
- Chips analysed in sets (e.g. an entire experiment)
- Background adjustment of PM made
- These values are normalized
- Normalized bg-adjusted PM values are $\log_2$-d
- A linear model including probe and chip effects is fitted robustly to probe × chip arrays of $\log_2 N(PM\text{-}bg)$ values

12 chips at 16 and 32 pM
Probe set= 36085_at



The probe intensity model

On a probe set by probe set basis (fixed k), the $\log_2$ of the normalized bg-adjusted probe intensities, denoted by $\mathbf{Y}_{\mathbf{kij},}$ are modelled as the sum of a **probe** effect $\boldsymbol{p_{ki}}$ and a **chip** effect $\boldsymbol{c_{kj}}$, and an **error** $\boldsymbol{\varepsilon_{kij}}$

$$\boldsymbol{Y_{kij} = p_{ki} + c_{kj} + \varepsilon_{kij}}$$

To make this model identifiable, we constrain the sum of the probe effects to be zero. The $\boldsymbol{p_{ki}}$ can be interpreted as probe relative non-specific binding effects.

The parameters $\boldsymbol{c_{kj}}$ provide an index of gene expression for each chip.

Least squares vs robust fit

Robust procedures perform well under a range of possible models and greatly facilitates the detection of anomalous data points.

Why robust?

- Image artifacts
- Bad probes
- Bad chips
- Quality assessment

M-estimators

(a one slide caption)

One can estimate the parameters of the model as solutions to

$$\min_{p_i, c_j} \sum_{i,j} \rho\left(\frac{Y_{ij} - p_i - c_j}{\hat{\sigma}}\right)^2 = \min_{p_i, c_j} \sum_{i,j} \rho(u_{ij})^2$$

where ρ is a symmetric, positive-definite function that increasing less rapidly than x. One can show that solutions to this minimization problem can be obtained by an IRLS procedure with weights:

$$w_{ij} = \rho'(u_{ij}) / u_{ij} = \psi(u_{ij})$$

Robust fit by IRLS

At each iteration $r_{ij} = Y_{ij}$ ***- current est(***p_i***) - current est(***c_j***),***

$S = MAD(r_{ij})$ a robust estimate of the scale parameter σ

$u_{ij} = r_{ij}/S$ standardized residuals

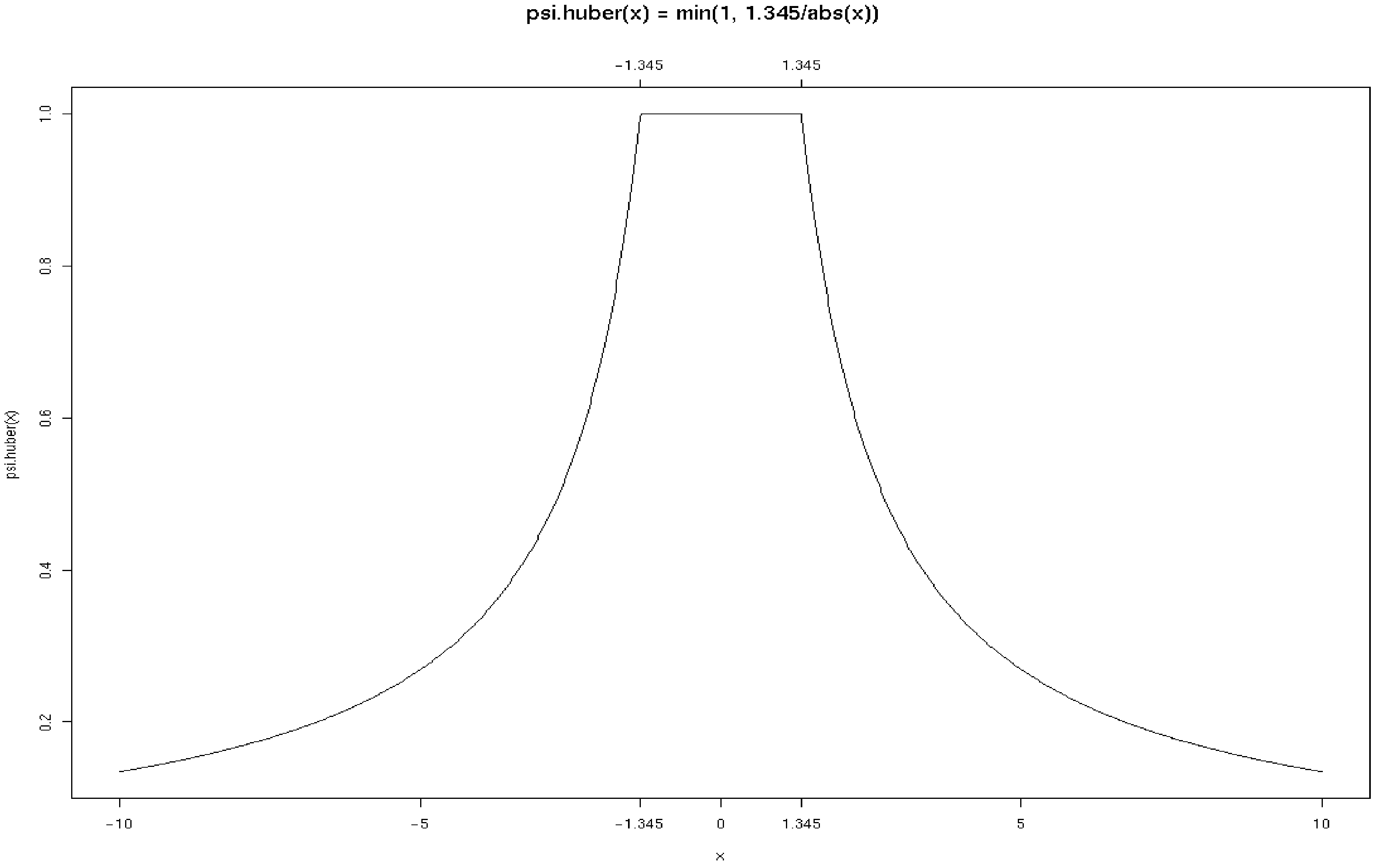
$w_{jj} = \psi(|u_{ij}|)$ weights to reduce the effect of discrepant points on the next fit

Next step estimates are:

est(p_i***) =*** *weighted row i mean – overall weighted mean*

est(c_j***)*** *= weighted column j mean*

Example – Huber ψ function



Application of the model to data quality assessment

Picture of the data – k=1,…, K

		Chip				Probe Effect
Probe Set	Probe	1	2	…	J	
k	1	Y_{k11}	Y_{k12}	…	Y_{k1J}	p_{k1}
	2	Y_{k21}	Y_{k22}	…	Y_{k2J}	p_{k2}
	…	…	…	…	…	…
	P	Y_{kP1}	Y_{kP2}	…	Y_{kPJ}	p_{kP}
	Chip Effect	c_{k1}	c_{k2}	…	c_{kJ}	S_k

- Robust vs Ls fit: whether c_{kj} is weighted average or not.
- **Single chip vs multi chip**: whether probe effects are removed from residuals or not – has huge impact on weighting and assessment of precision.

Model components – role in QA

- 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 per array
  - can examine distribution of relative expressions across arrays.
- Probe effects > 200K per model for hg_u133
  - can be compared across fitting sets.

Chip index of relative quality

We assess gene expression index variability by it's unscaled SE:

$$\text{unscaled SE}(\hat{c}_{\text{kj}}) = 1 \Big/ \sqrt{\sum_i w_{kij}}$$

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

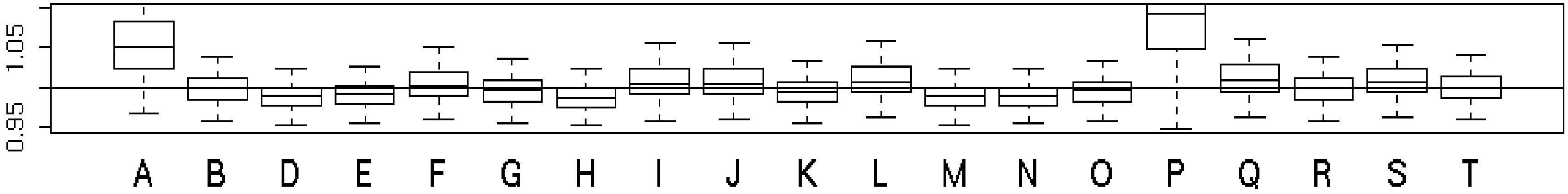
$$\text{NUSE}(\hat{c}_{\text{kj}}) = {1 \Big/ \sqrt{\sum_i w_{kij}}} \bigg/ {median_j (1 \Big/ \sqrt{\sum_i w_{kij}})}$$

Example –
NUSE + residual images

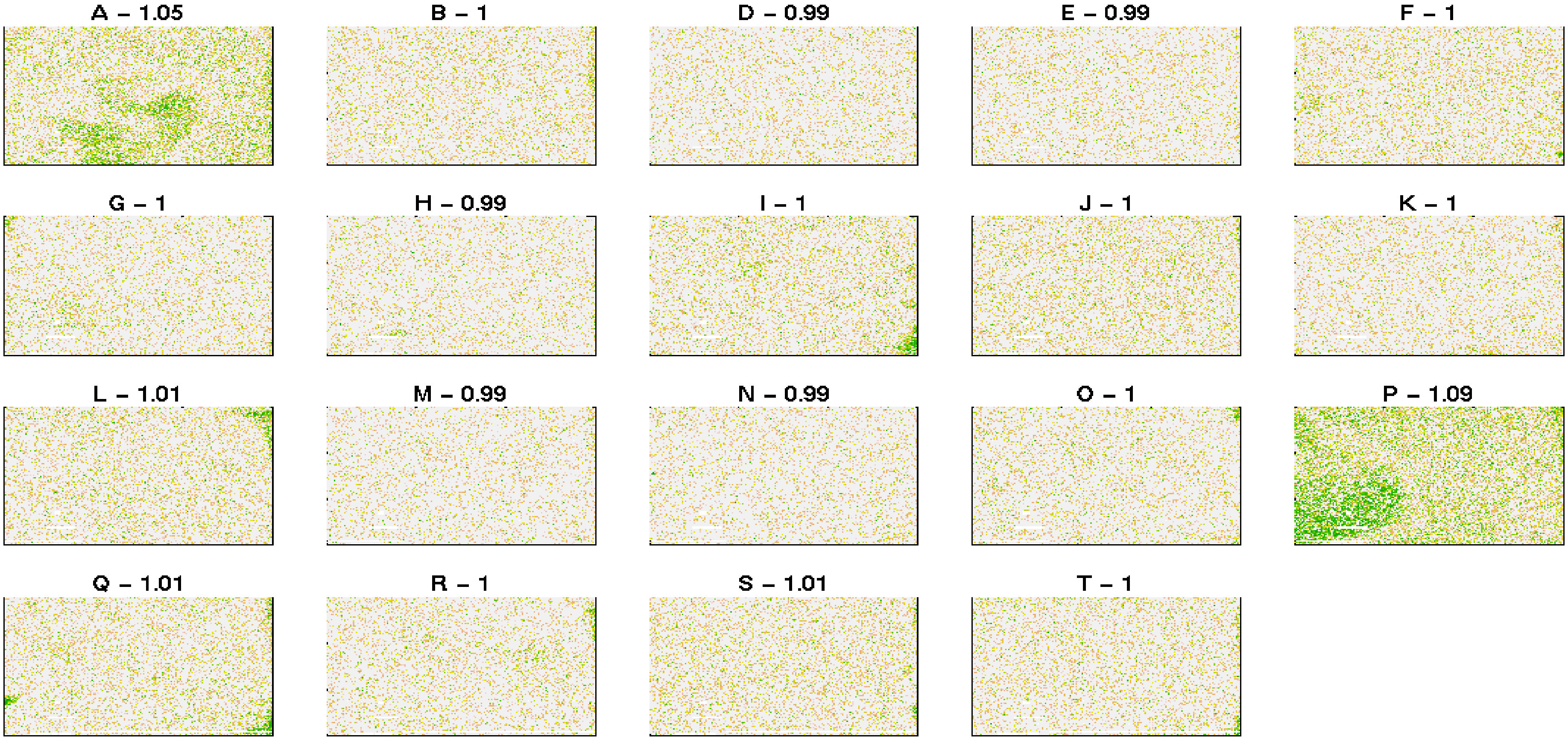
- Affymetrix hg-u95A spike-in, 1532 series – next slide.
- St-Judes Childern's Research Hospital- several groups – slides after next.

Note – special challenge here is to detect differences in perfectly good chips!!!

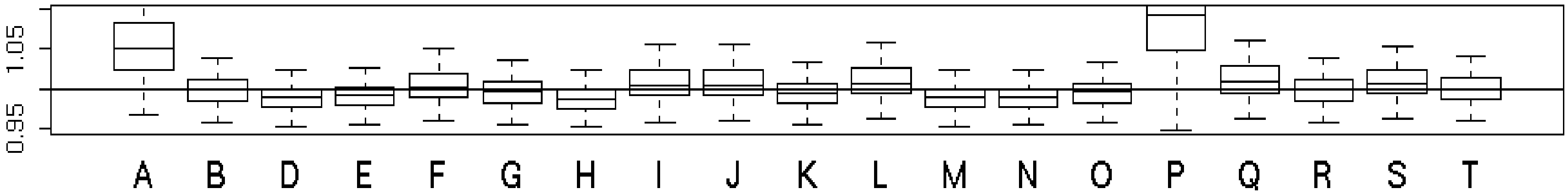
Normalized Unscaled Probeset Standard Errors – L2353



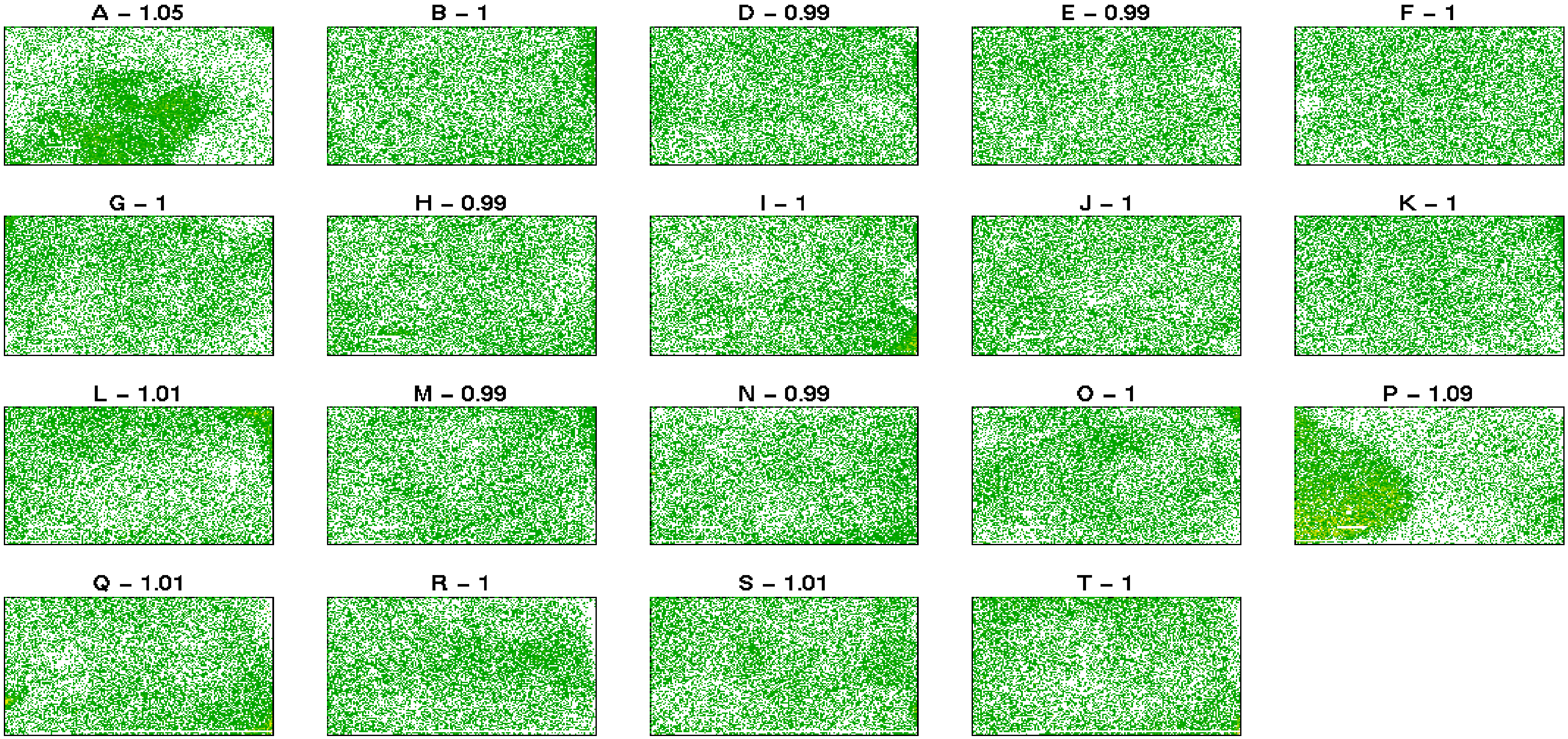
Pseudo images of weights: Chip – median(NUSE)



Normalized Unscaled Probeset Standard Errors - L2353



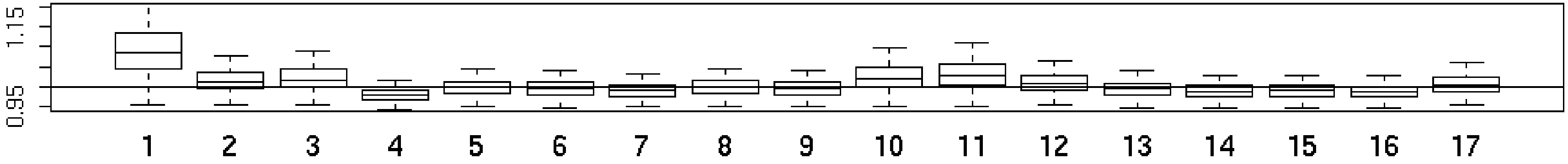
Pseudo images of positive residuals: Chip – median(NUSE)



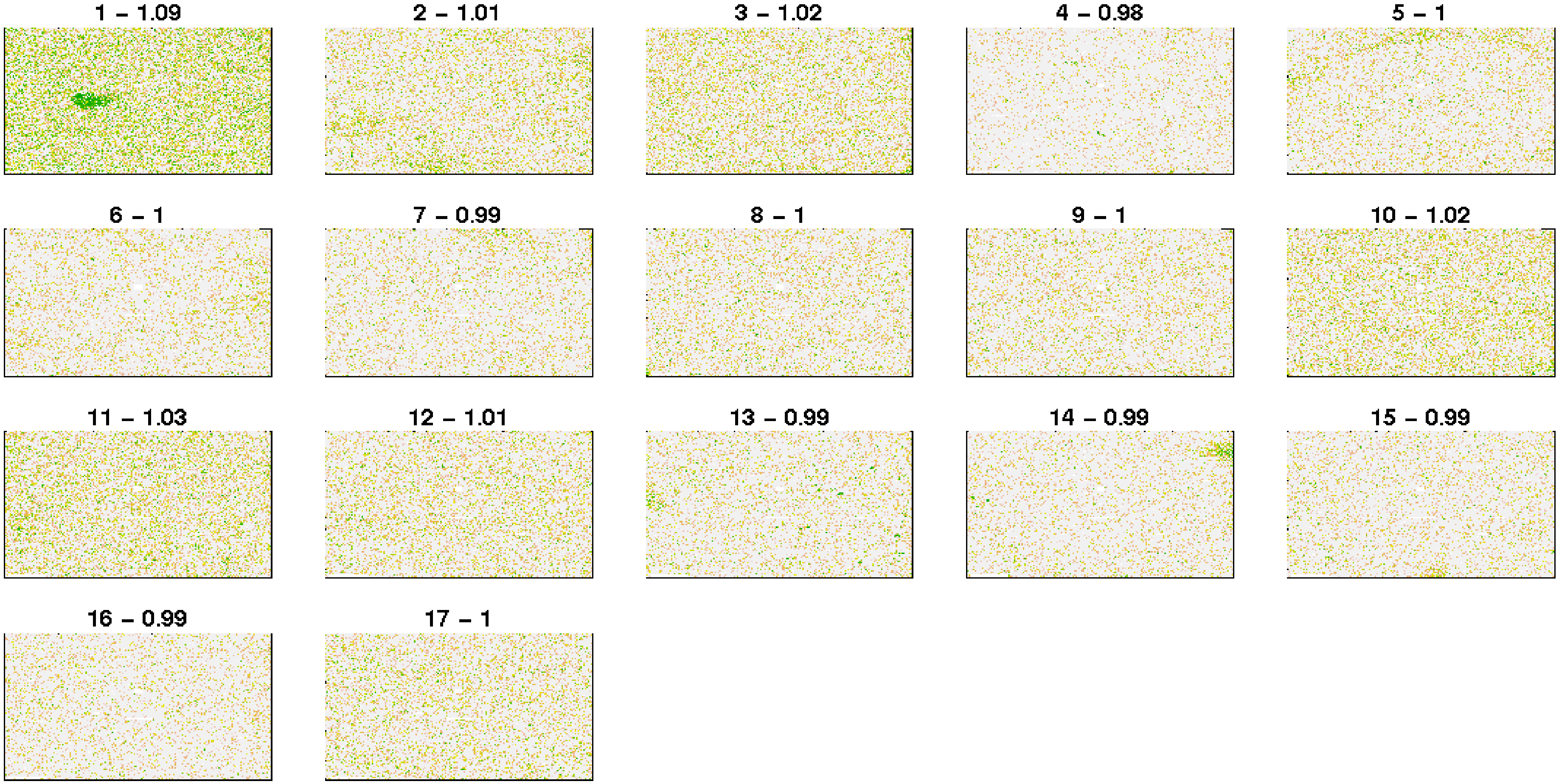
St Jude hosptial NUSE + wts images HERE

- St-Judes Childern’s Research Hospital- two groups selected from over all fit assessment which follows.

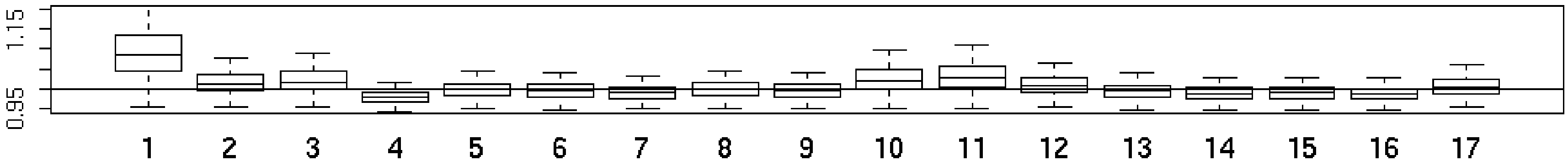
Normalized Unscaled Probeset Standard Errors – Hyperdip_chipA



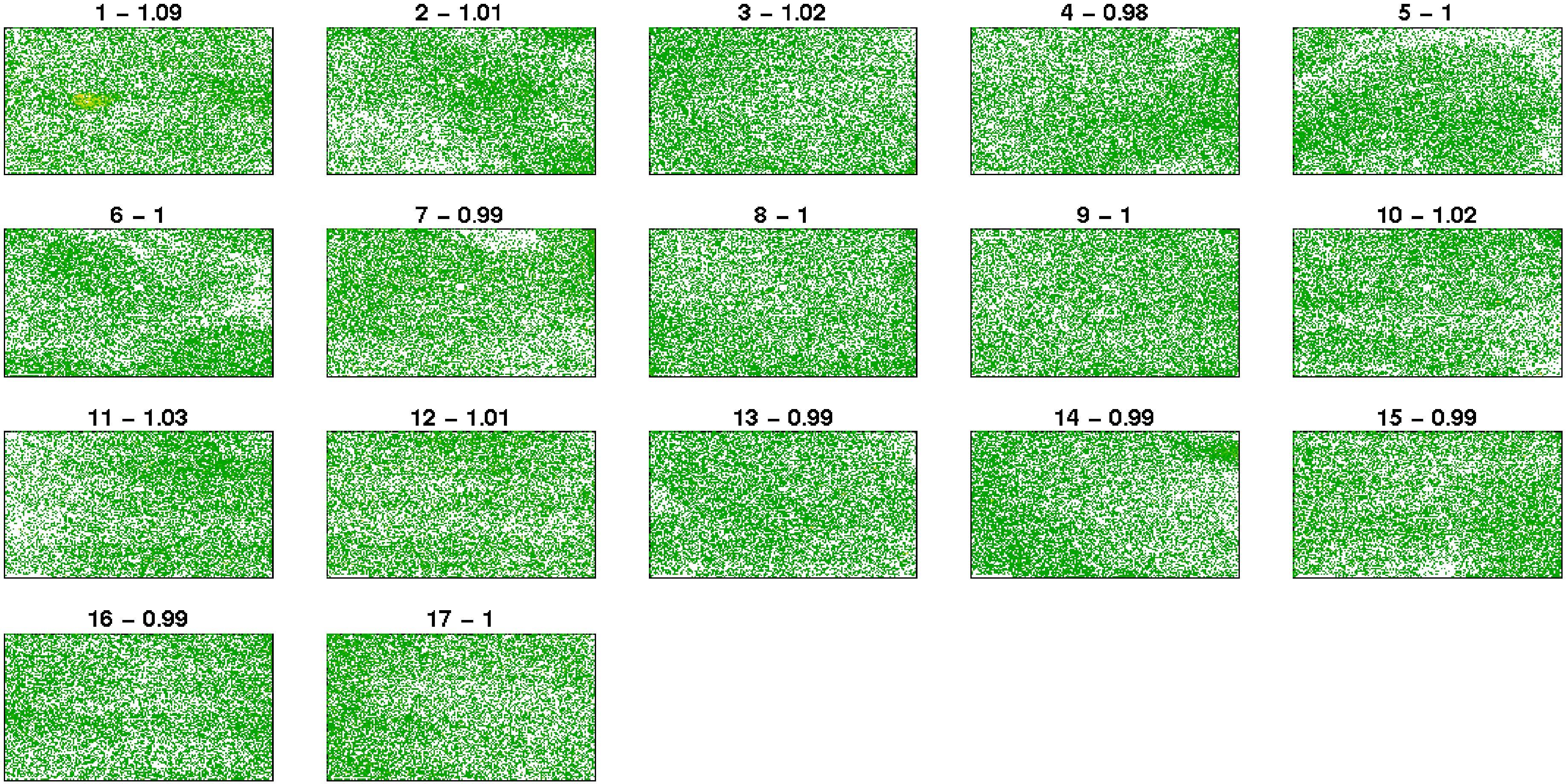
Pseudo images of weights: Chip – median(NUSE)



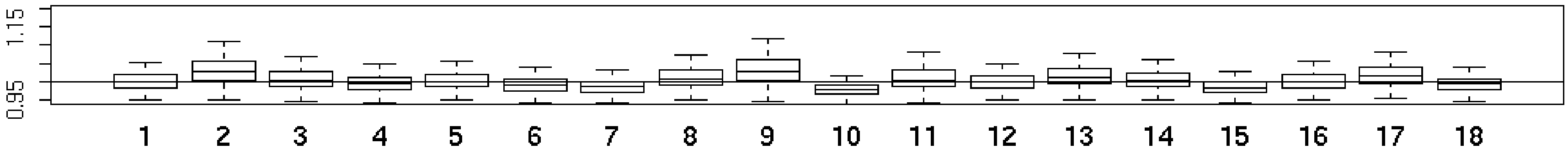
Normalized Unscaled Probeset Standard Errors - Hyperdip_chipA



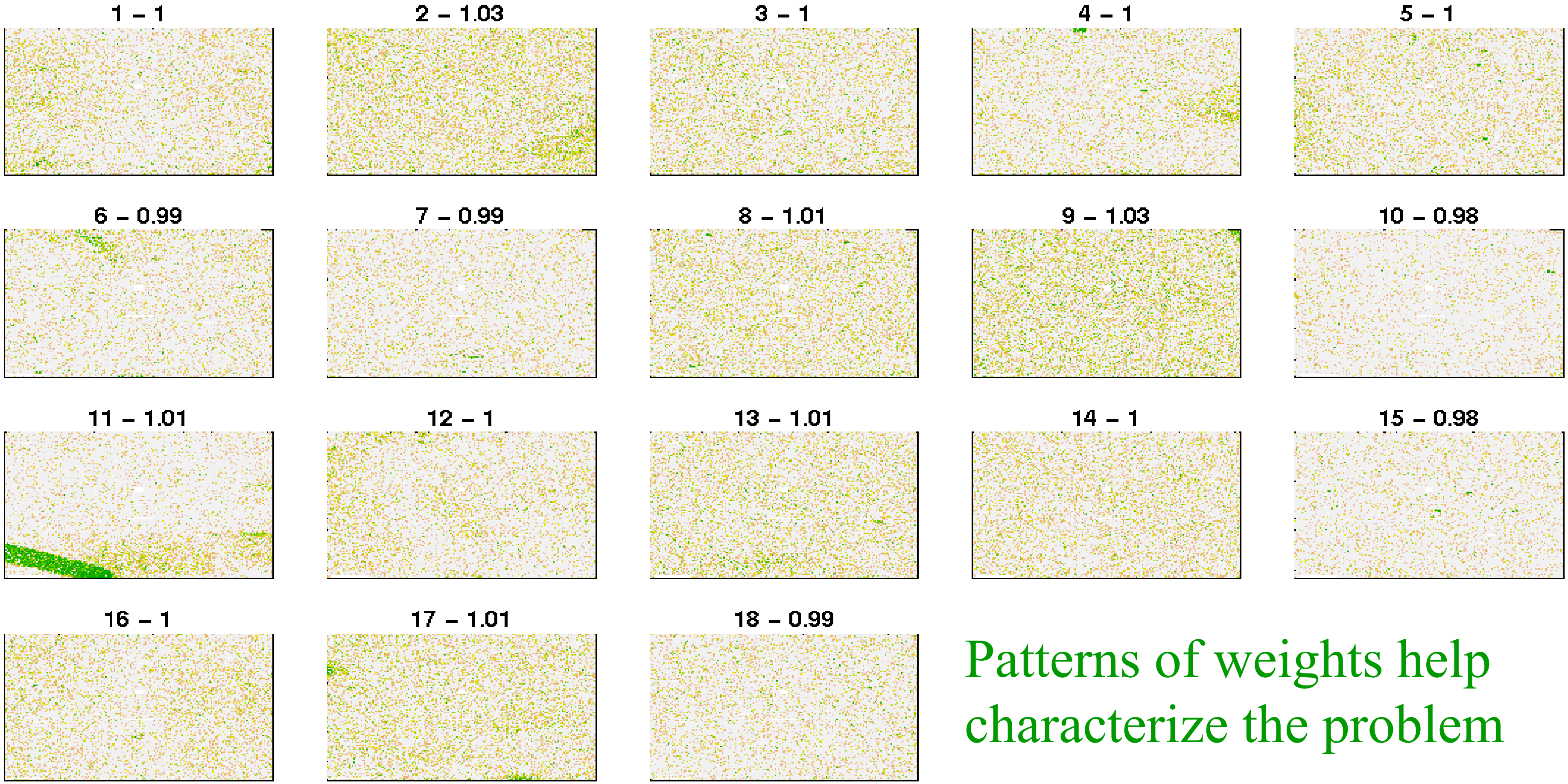
Pseudo images of positive residuals: Chip – median(NUSE)



Normalized Unscaled Probeset Standard Errors – E2A_PBX1_chipA

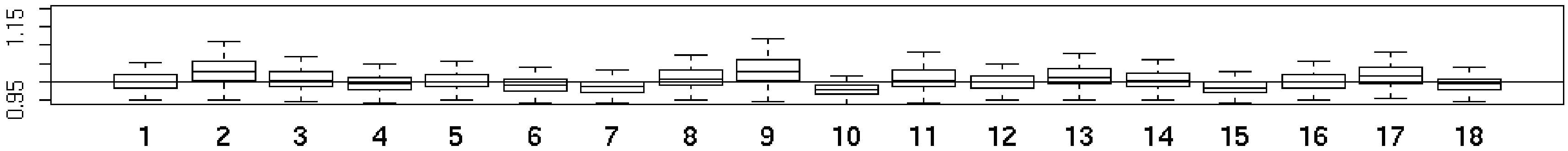


Pseudo images of weights: Chip – median(NUSE)

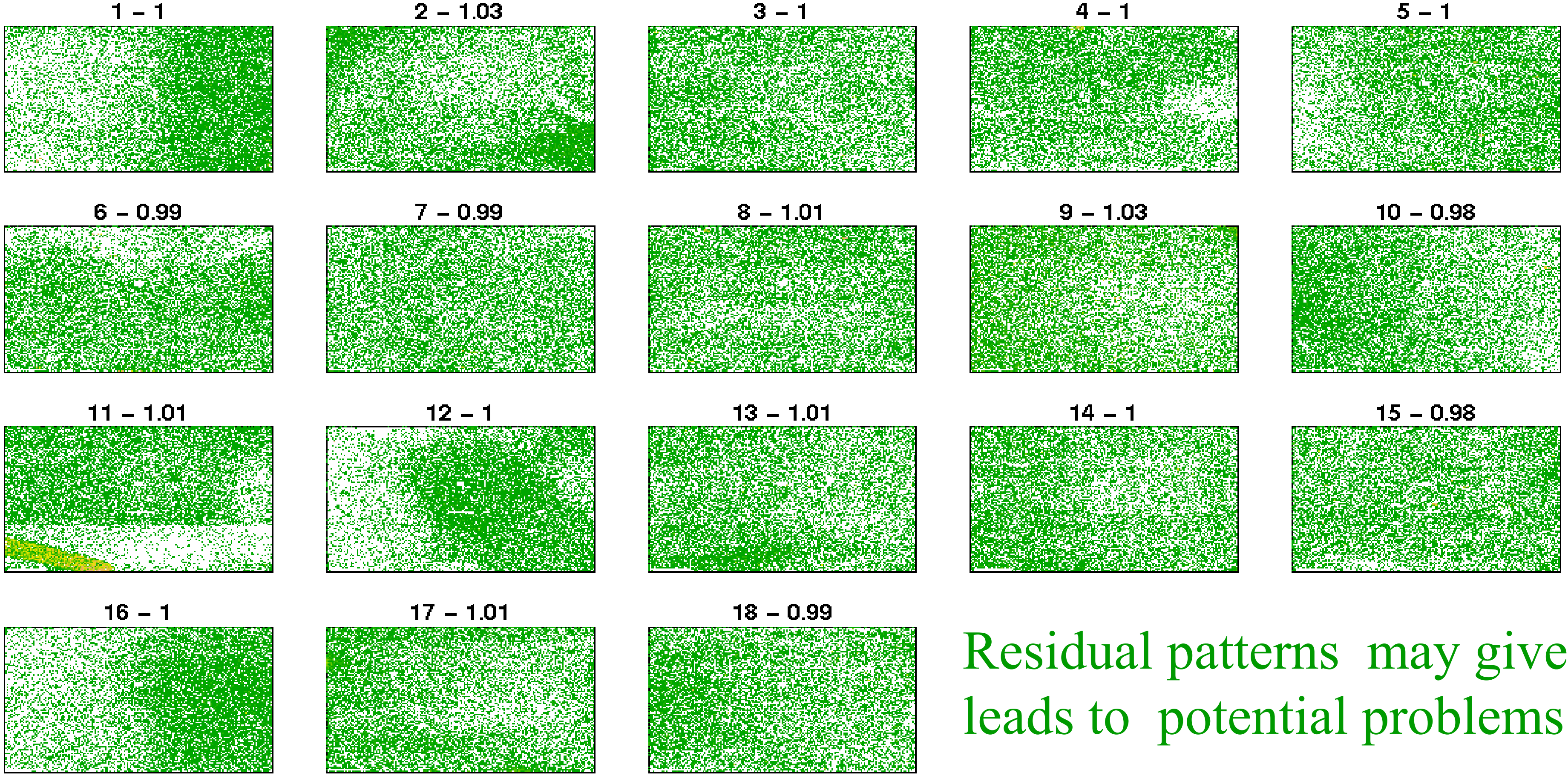


Patterns of weights help characterize the problem

Normalized Unscaled Probeset Standard Errors – E2A_PBX1_chipA

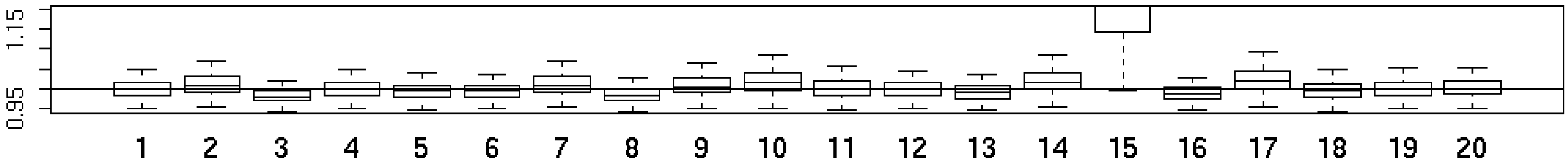


Pseudo images of positive residuals: Chip – median(NUSE)

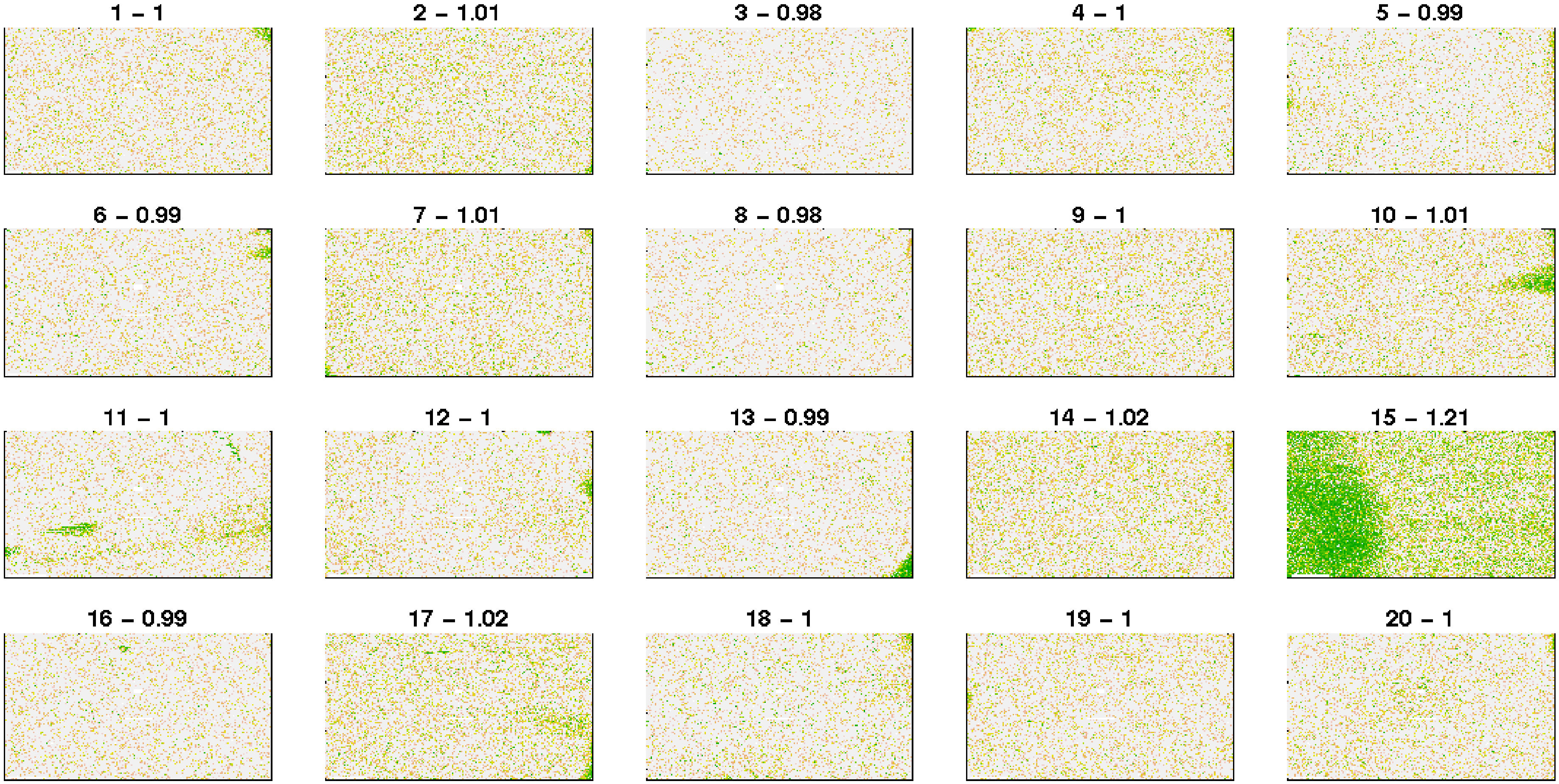


Residual patterns may give leads to potential problems.

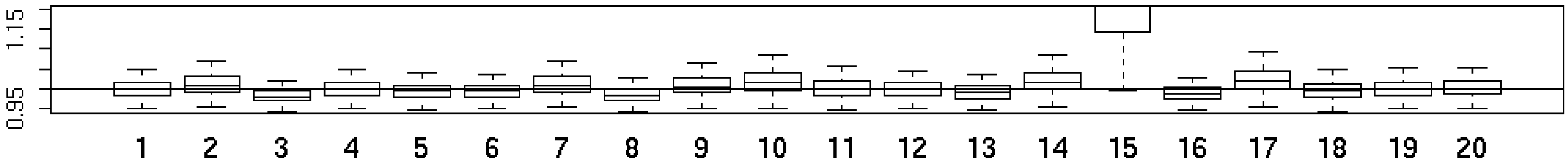
Normalized Unscaled Probeset Standard Errors – MLL_chipB



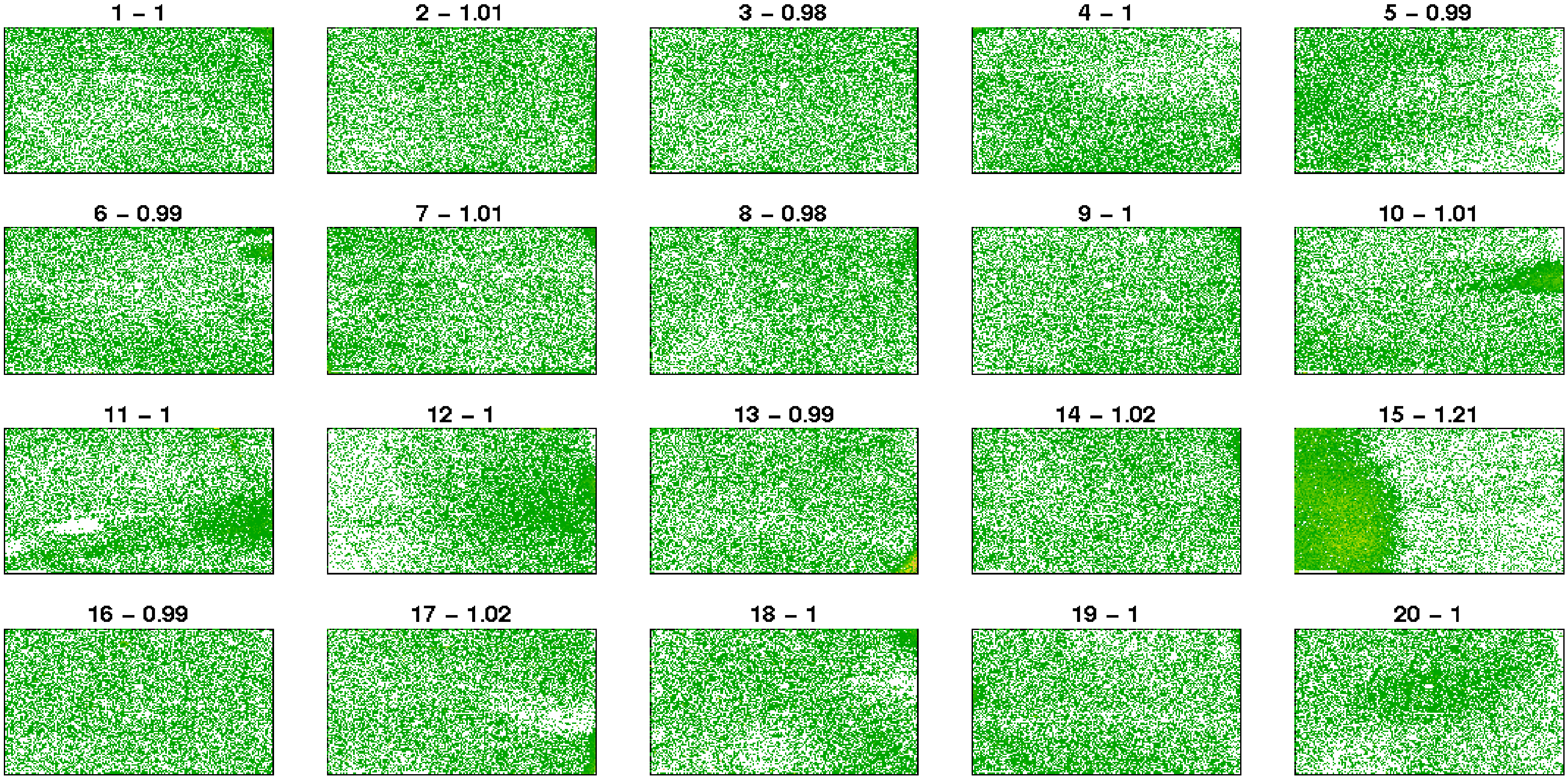
Pseudo images of weights: Chip – median(NUSE)



Normalized Unscaled Probeset Standard Errors – MLL_chipB



Pseudo images of positive residuals: Chip – median(NUSE)



Another quality measure: variability of relative log expression

How much are robust summaries affected?

We can gauge reproducibility of expression measures by summarizing the distribution of **relative log expressions:**

$$LR_{kj} = \hat{c}_{kj} - \widetilde{c}_k$$

where $\widetilde{c}_k$ is a reference expression for gene k.

For reference expression, in the absence of technical replicates, we use the median expression value for that gene in a set of chips.

Relative expression summaries

- IQR(LR_{kj}) measures variability which includes **Noise + Differential expression** in biological replicates.
- When biological replicates are similar (eg. RNA from same tissue type), we can typically detect processing effects with IQR(LR)
- Median(LR_{kj}) should be close to zero if No. up and regulated genes are roughly equal.

IQR(LR_{kj})+|Median(LR_{kj})| can be combined to give a measure of chip expression measurement error.

Other Chip features: Signal + Noise

We consider the Noise + Signal model:

PM = N + S

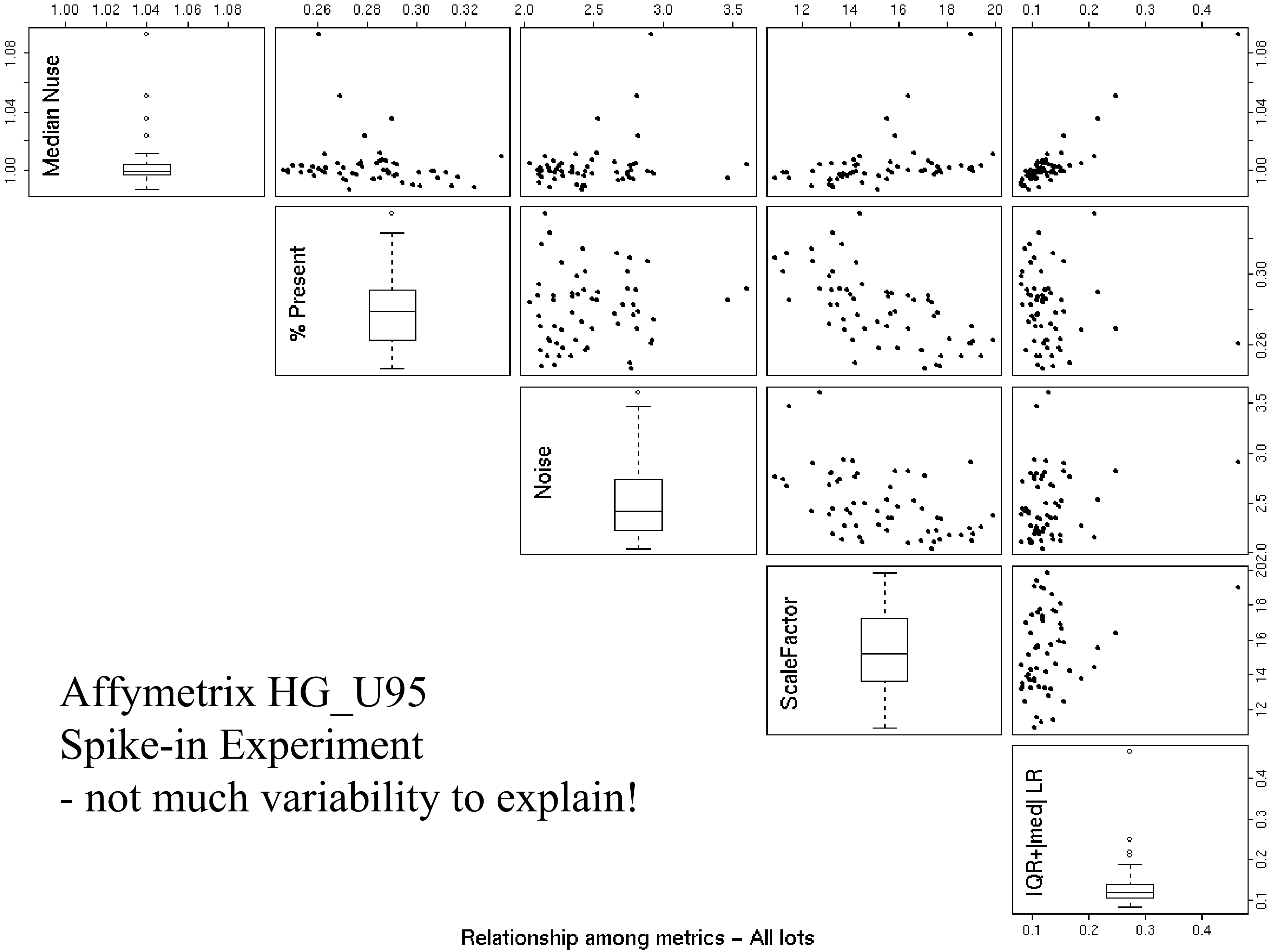
Where N ~ $N(\mu, \sigma^2)$ and S ~ $\mathrm{Exp}(1/\alpha)$

We can use this model to obtain “background corrected” PM values – won’t discuss here.

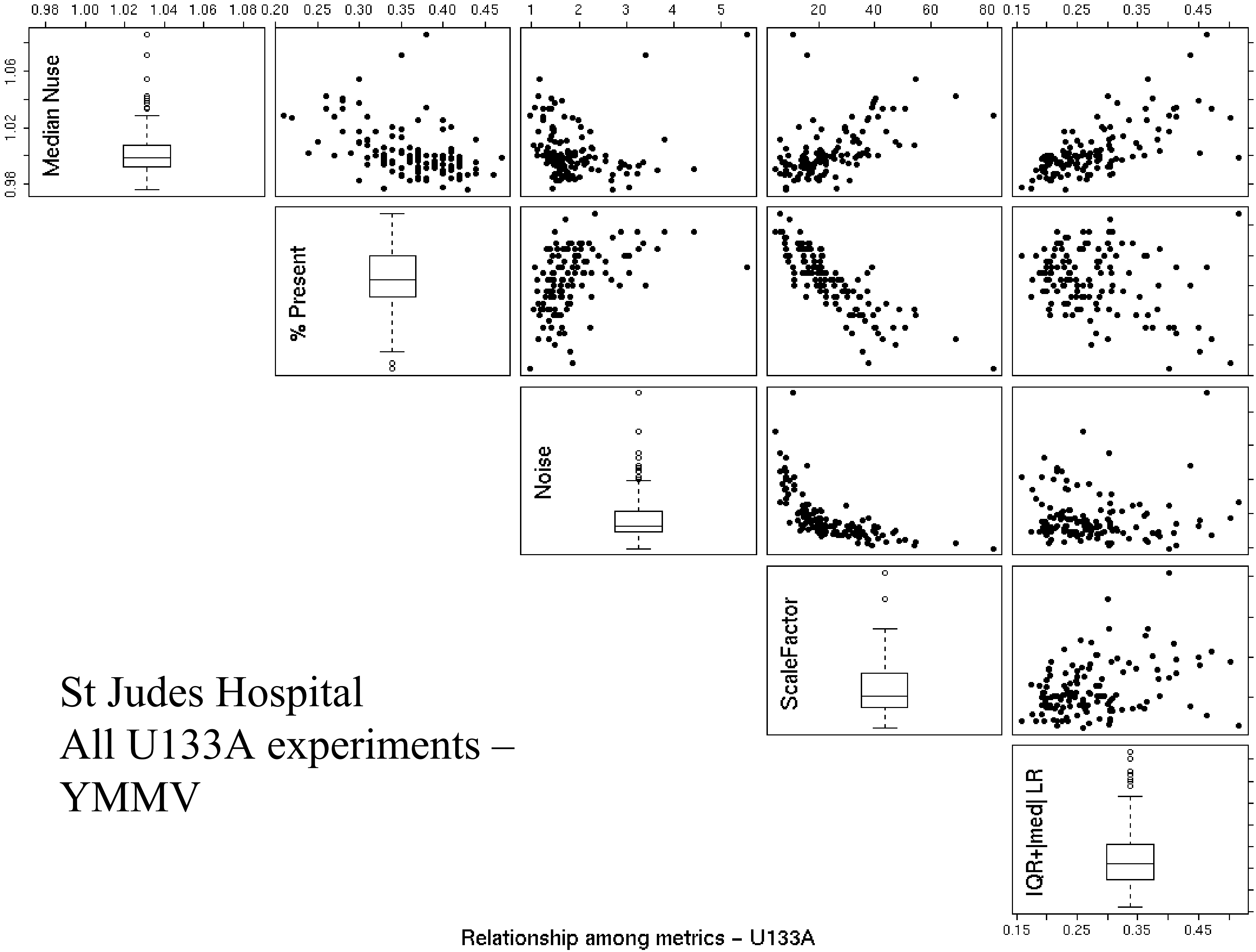
Our interest here is to see how measures of level of signal ($1/\alpha$) and noise (μ) relate to other indicators.

* In the example data sets used here, %P, SF and RMA S/N measures correlate similarly with median NUSE *

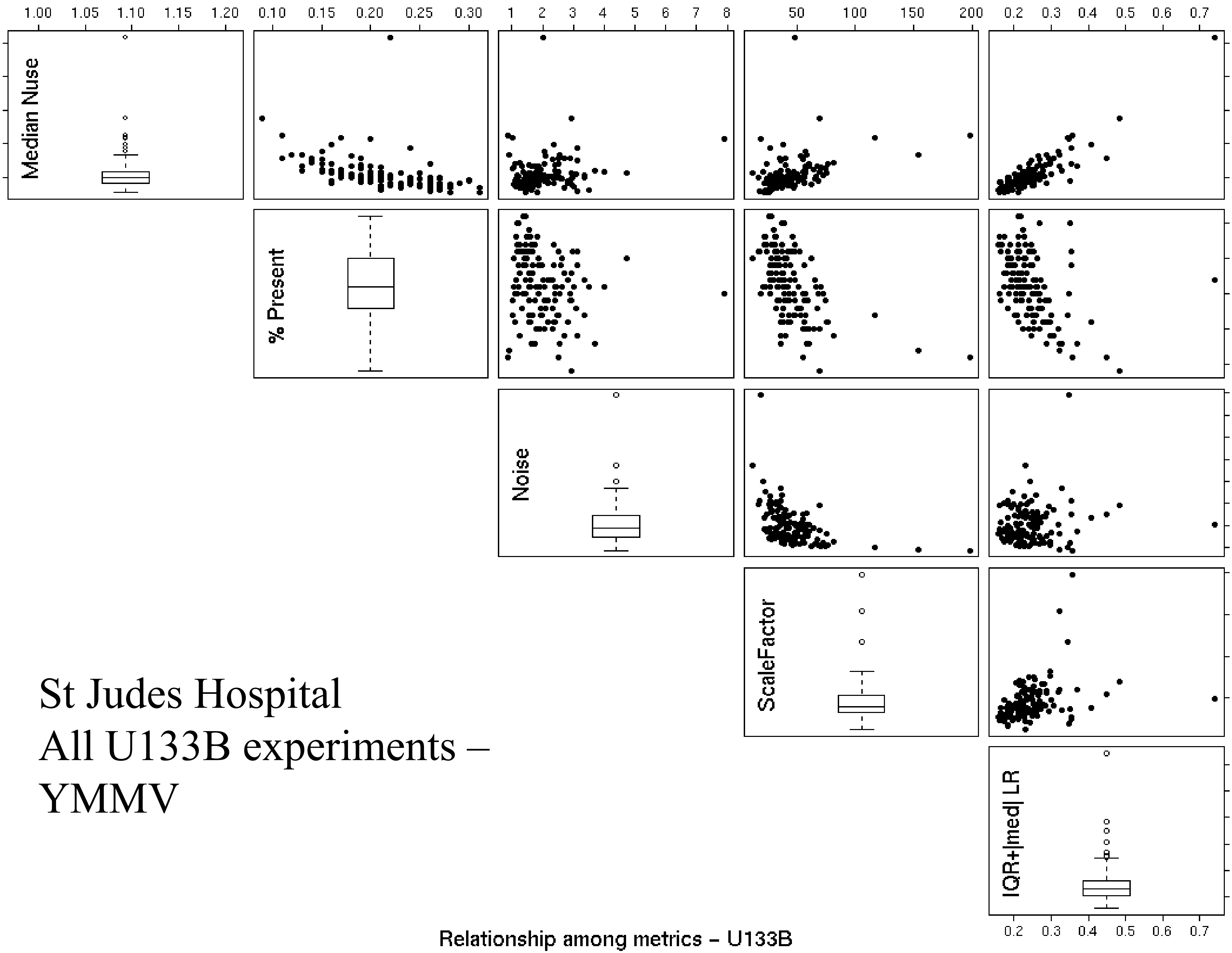
Comparison of quality indicators



Affymetrix HG_U95
Spike-in Experiment
- not much variability to explain!



St Judes Hospital
All U133A experiments –
YMMV



St Judes Hospital
All U133B experiments –
YMMV

Correlation among measures for U133A chips

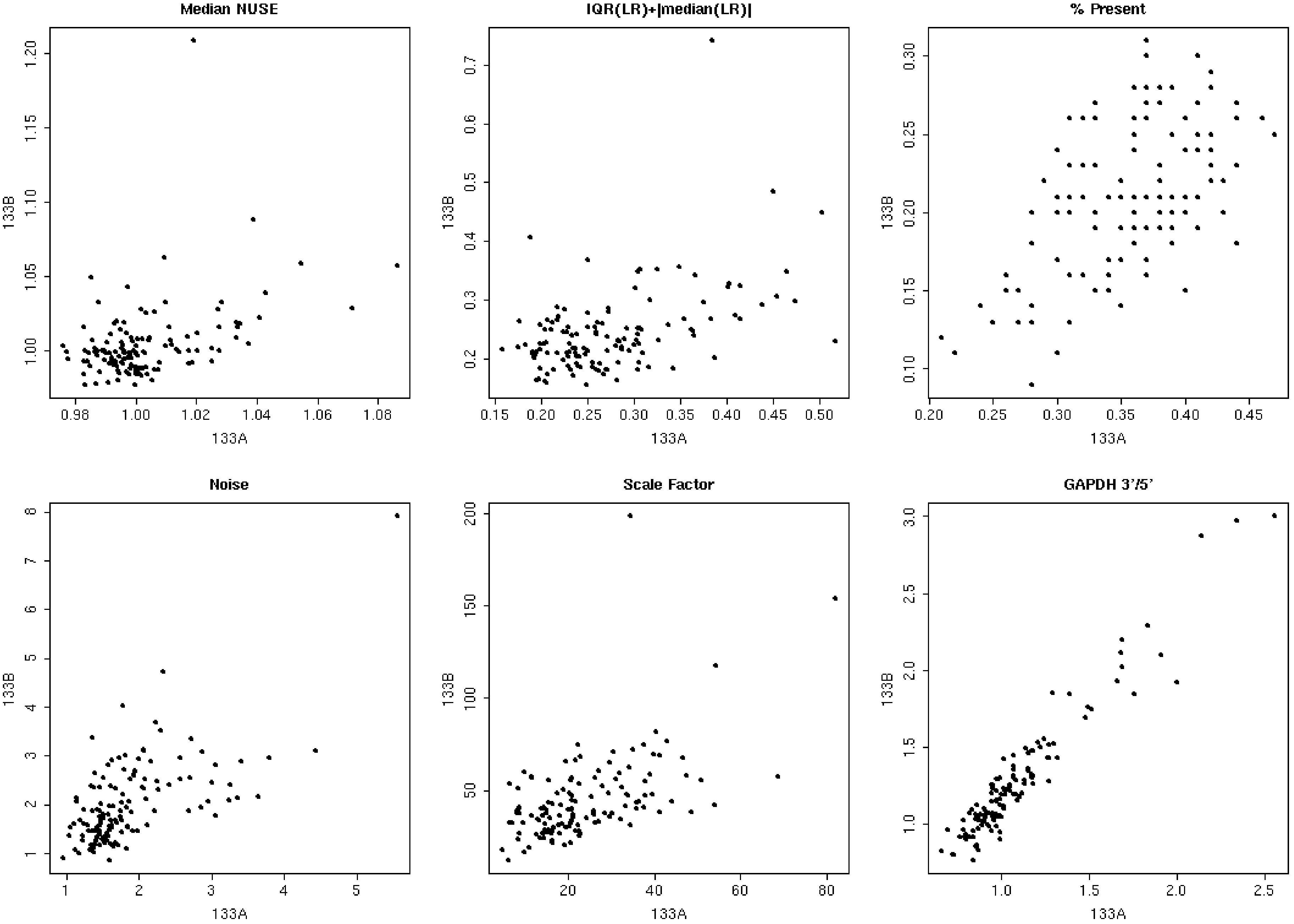
	Median.Nuse	IQRplusB	PercPresent	Noise	Background	ScaleFactor	Gapdh.3P5P	RMA S/N
Median.Nuse	1.00	0.69	-0.46	0.00	0.03	0.52	0.09	-0.54
IQRplusB	0.69	1.00	-0.29	-0.01	0.02	0.32	0.02	-0.31
PercPresent	-0.46	-0.29	1.00	0.44	0.36	-0.83	-0.09	0.75
Noise	0.00	-0.01	0.44	1.00	0.90	-0.64	-0.01	0.60
Background	0.03	0.02	0.36	0.90	1.00	-0.57	-0.09	0.41
ScaleFactor	0.52	0.32	-0.83	-0.64	-0.57	1.00	0.09	-0.87
Gapdh.3P5P	0.09	0.02	-0.09	-0.01	-0.09	0.09	1.00	-0.03
RMA S/N	-0.54	-0.31	0.75	0.60	0.41	-0.87	-0.03	1.00

Your Mileage May Vary – ie. depending on chip selection, relationships may differ in your chip set

Correlation among measures for U133B chips

	Median.Nuse	IQRplusB	PercPresent	Noise	Background	ScaleFactor	Gapdh.3P5P	RMA S/N
Median.Nuse	1.00	0.88	-0.47	0.18	0.24	0.38	0.08	-0.31
IQRplusB	0.88	1.00	-0.42	0.12	0.17	0.33	0.06	-0.26
PercPresent	-0.47	-0.42	1.00	-0.18	-0.35	-0.54	-0.20	0.74
Noise	0.18	0.12	-0.18	1.00	0.92	-0.45	0.02	-0.01
Background	0.24	0.17	-0.35	0.92	1.00	-0.34	0.06	-0.22
ScaleFactor	0.38	0.33	-0.54	-0.45	-0.34	1.00	0.13	-0.62
Gapdh.3P5P	0.08	0.06	-0.20	0.02	0.06	0.13	1.00	-0.23
RMA S/N	-0.31	-0.26	0.74	-0.01	-0.22	-0.62	-0.23	1.00

Scatter plots of measures for U133A chip vs U133B chip



Median NUSE

IQR(LR)+|median(LR)|

% Present

Noise

Scale Factor

GAPDH 3'/5'

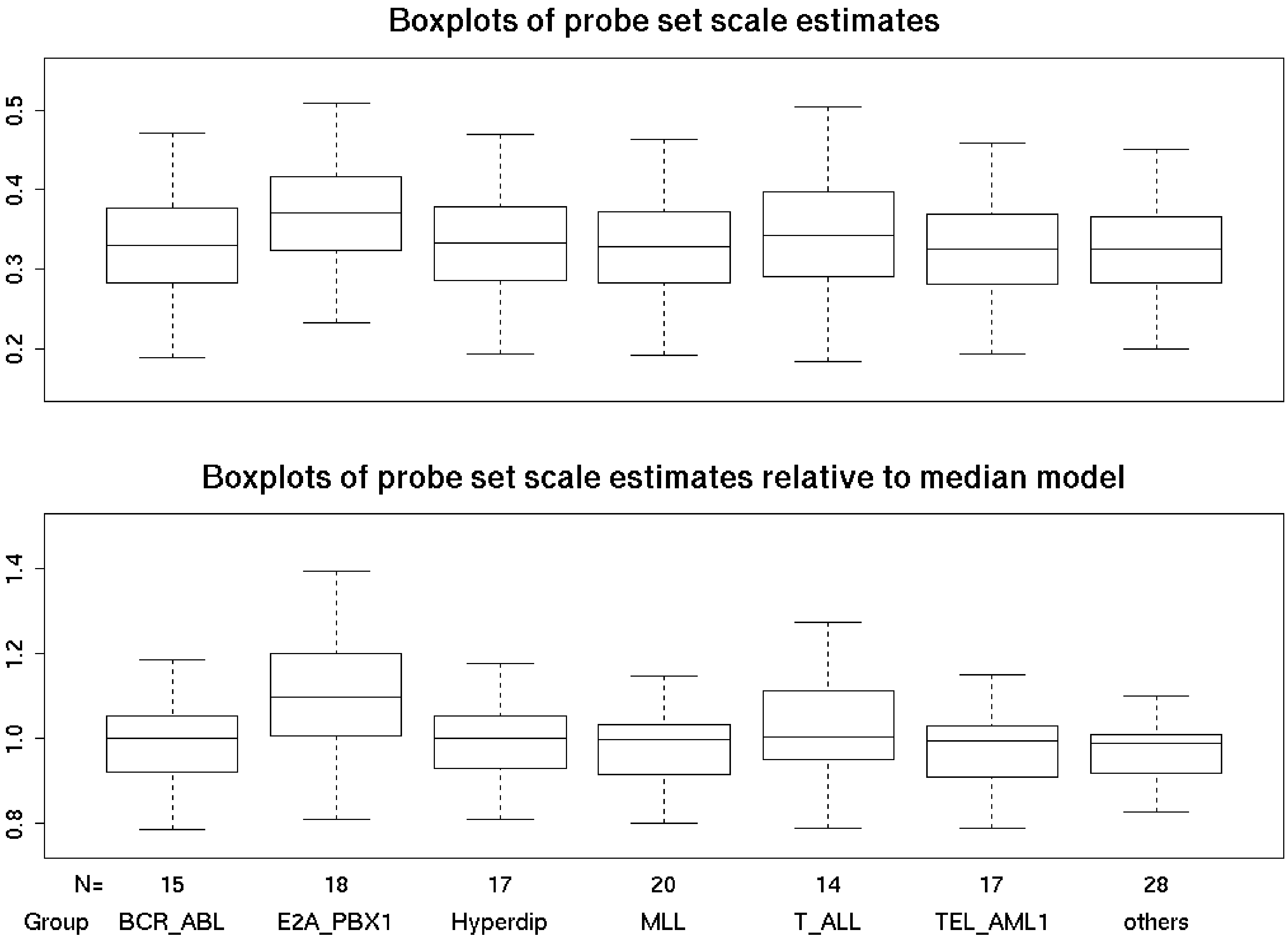
Comparing experiments

- NUSE: have no units – only get relative quality within chip set (could use a ref. QC set)
- IQR(LR): include some biological variability which might vary between experiments

Can use model residual scales (S_k) to compare experiments (assuming the intensity scale was standardized)

Next: Analyzed St-Judes chips by treatment group (14-28 chips per group). Compare scale estimates.

Comparison of experiments through residual scales
U133A chips

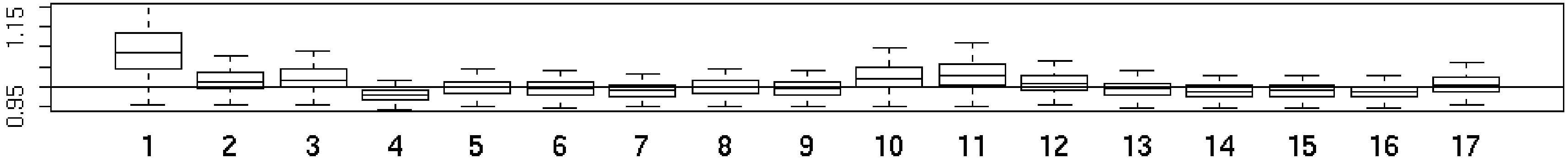


Boxplots of probe set scale estimates

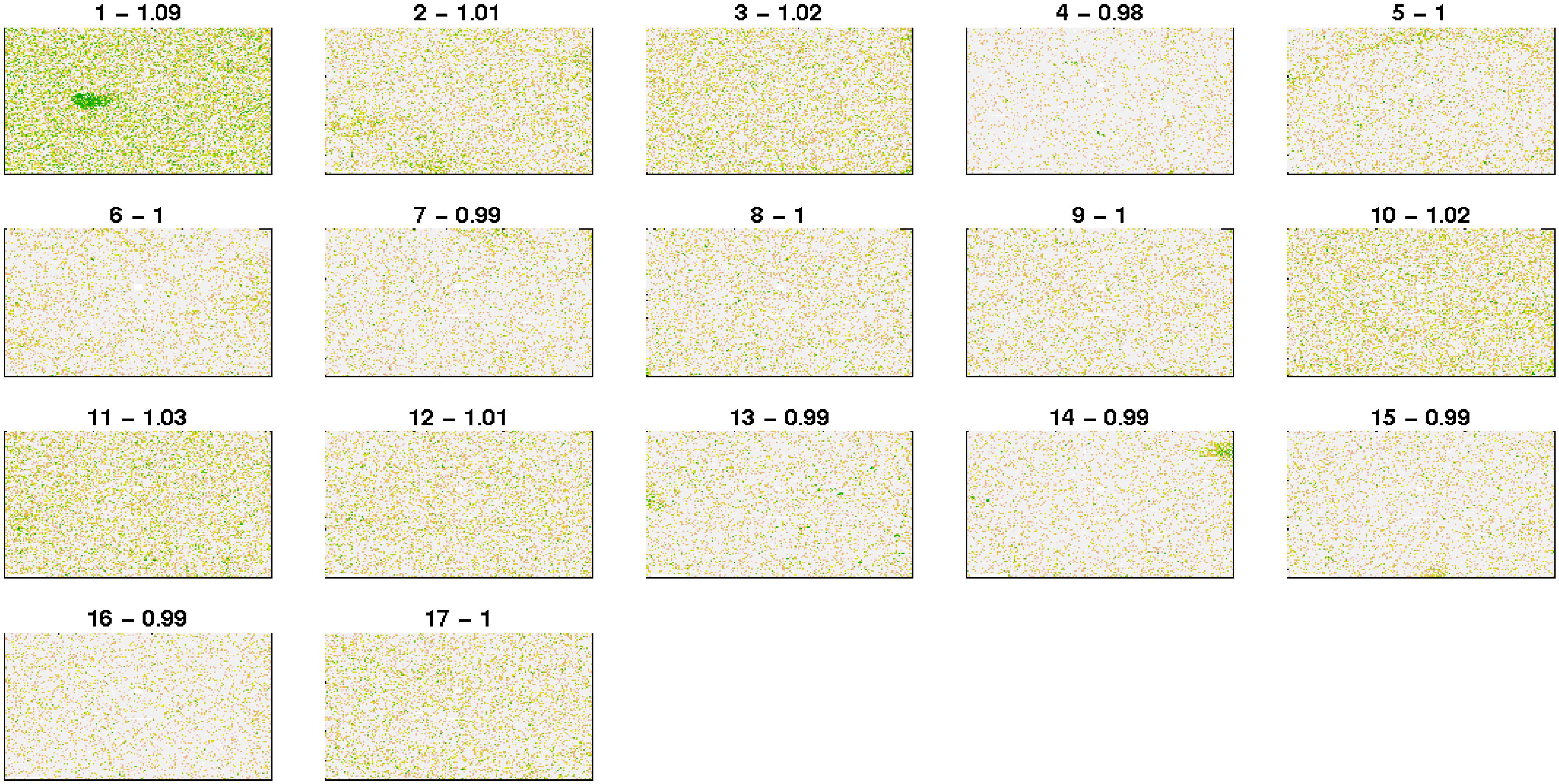
Boxplots of probe set scale estimates relative to median model

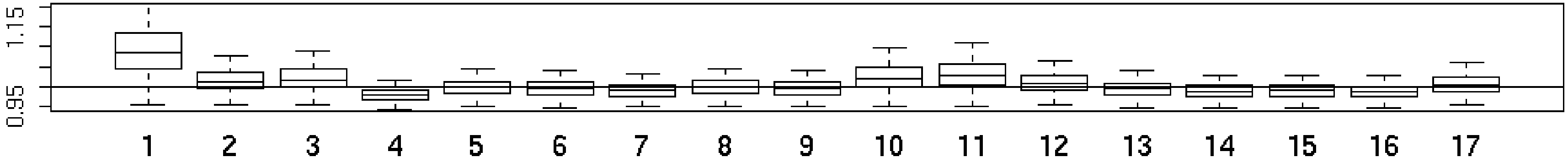
Next contrast the good and the less good

Normalized Unscaled Probeset Standard Errors – Hyperdip_chipA

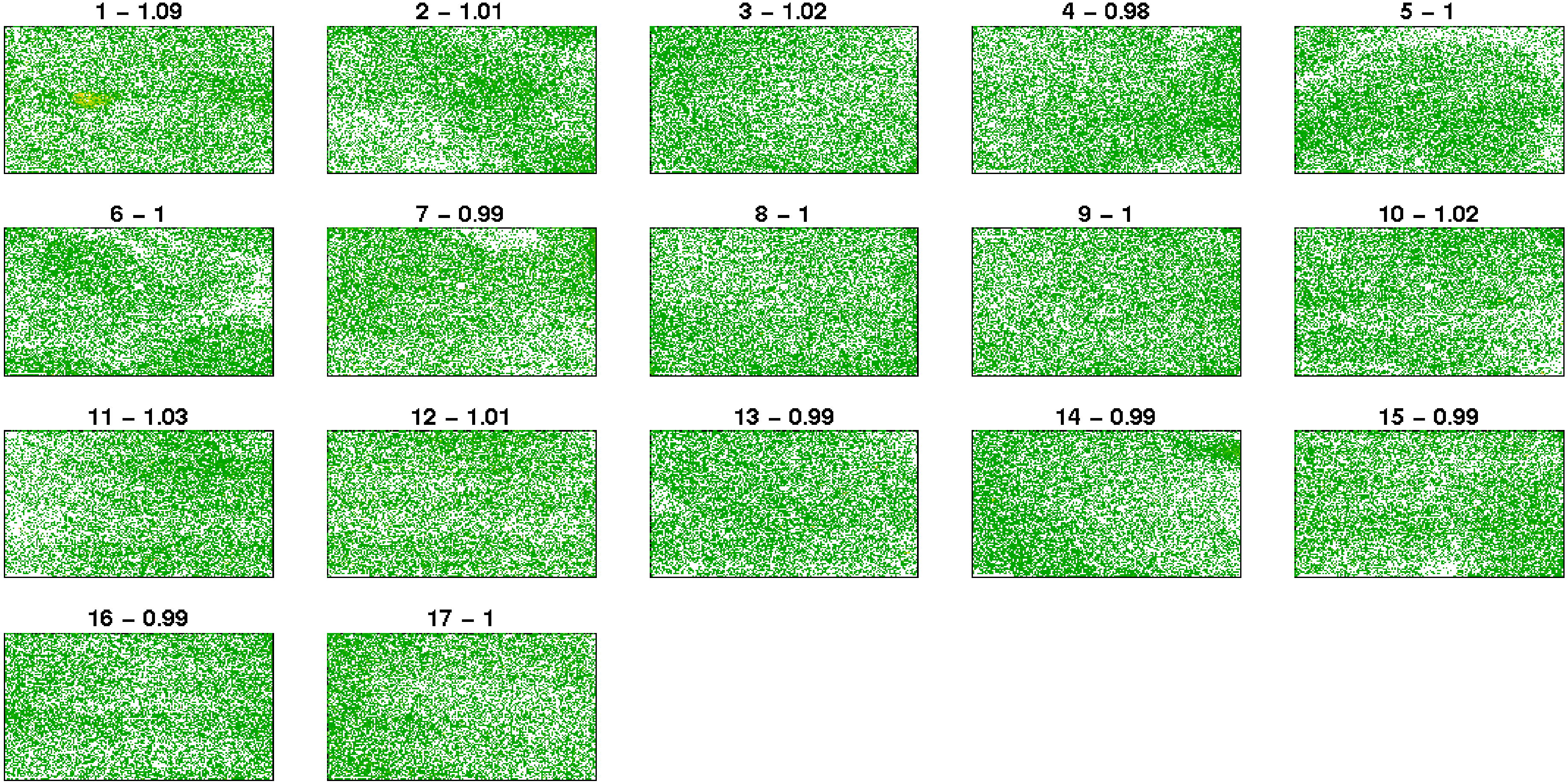


Pseudo images of weights: Chip – median(NUSE)



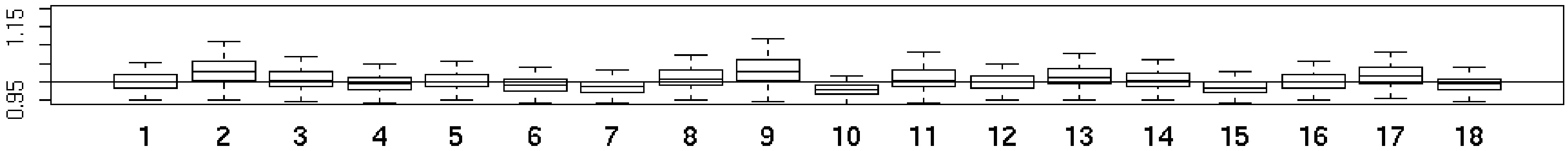


Normalized Unscaled Probeset Standard Errors – Hyperdip_chipA

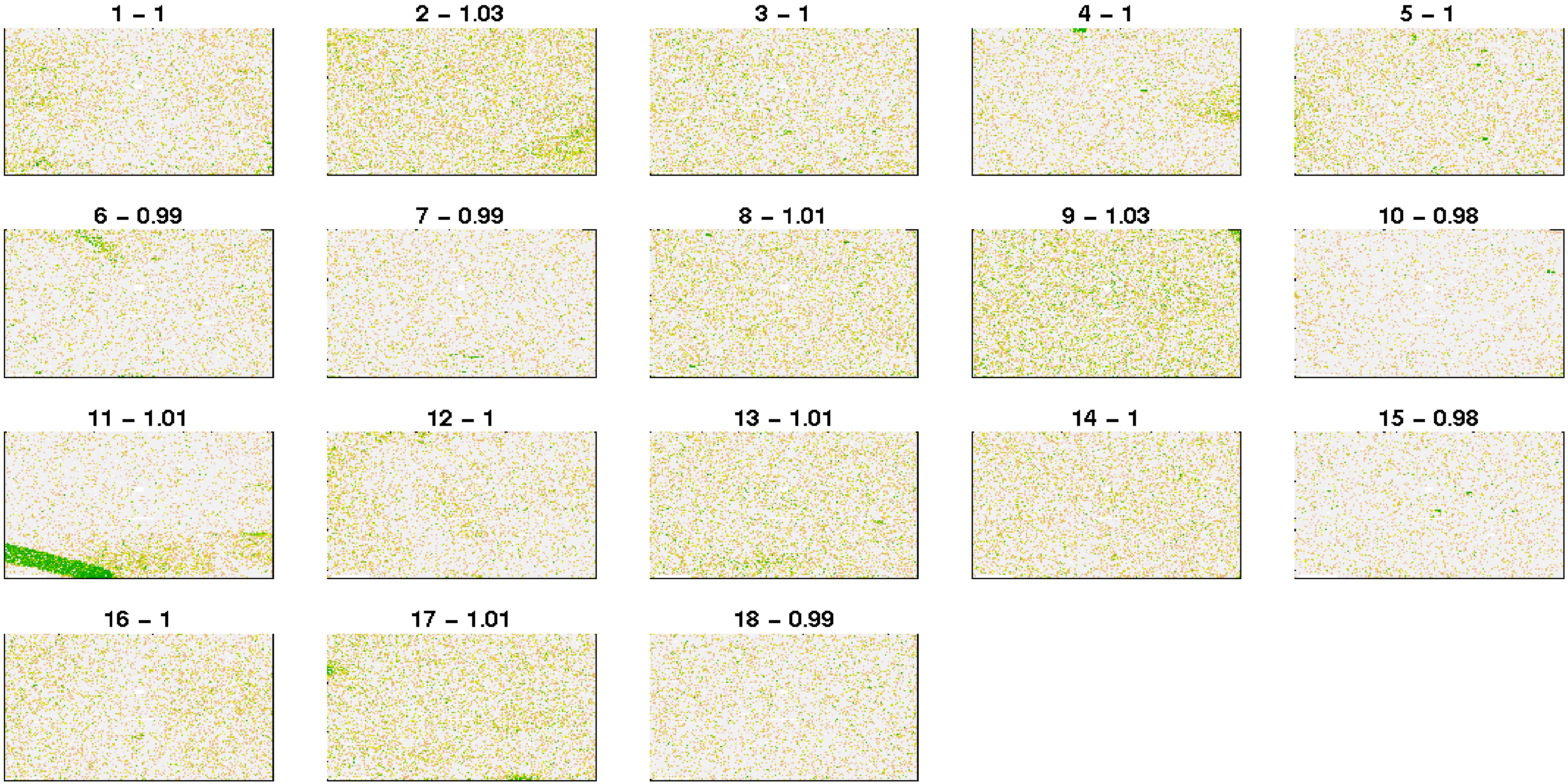


Pseudo images of positive residuals: Chip – median(NUSE)

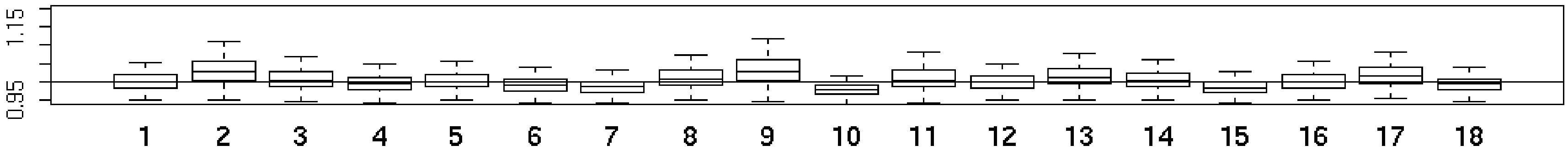
Normalized Unscaled Probeset Standard Errors – E2A_PBX1_chipA



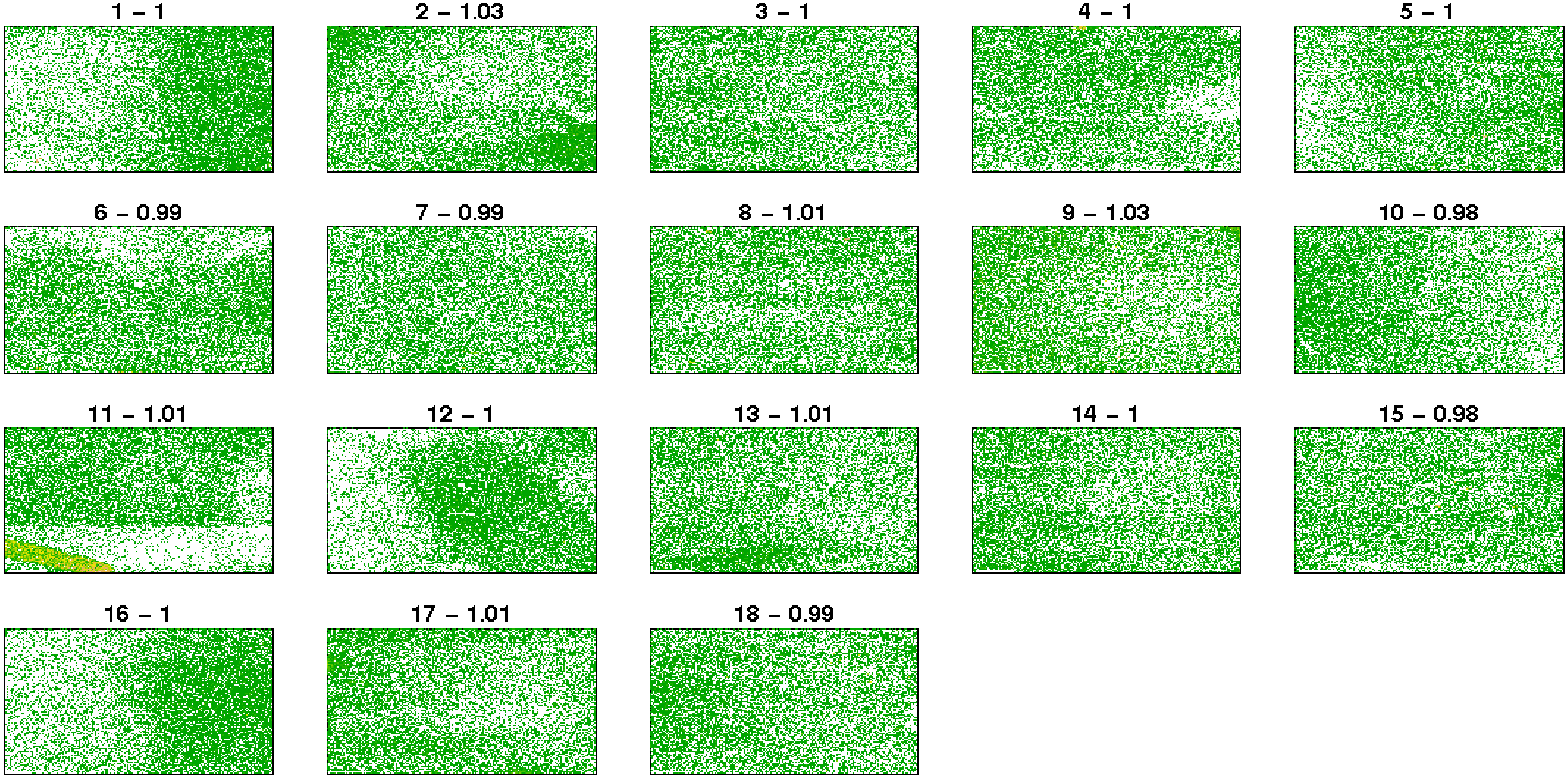
Pseudo images of weights: Chip – median(NUSE)



Normalized Unscaled Probeset Standard Errors – E2A_PBX1_chipA



Pseudo images of positive residuals: Chip – median(NUSE)



1 – 1

2 – 1.03

3 – 1

4 – 1

5 – 1

6 – 0.99

7 – 0.99

8 – 1.01

9 – 1.03

10 – 0.98

11 – 1.01

12 – 1

13 – 1.01

14 – 1

15 – 0.98

16 – 1

17 – 1.01

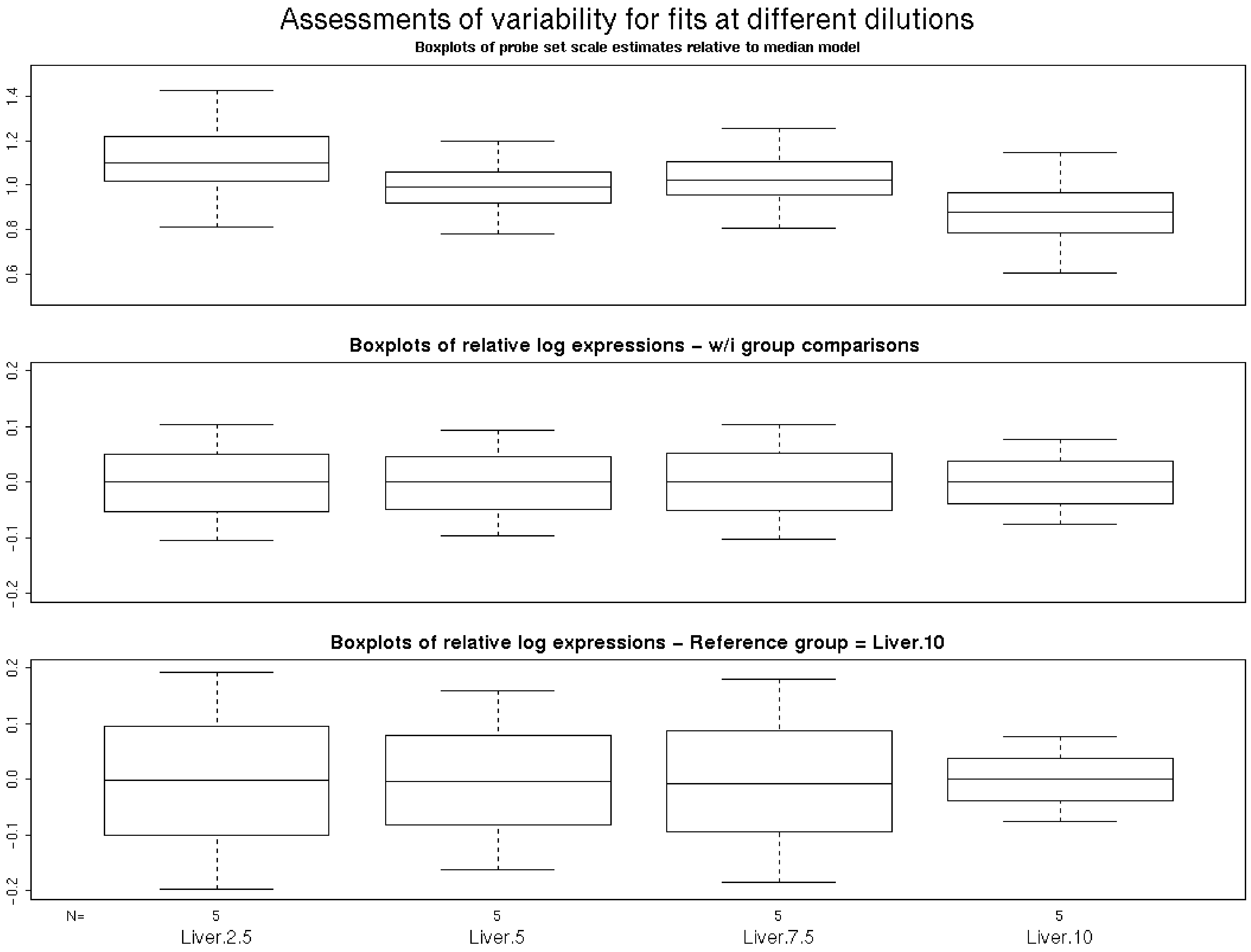
18 – 0.99

More model comparisons

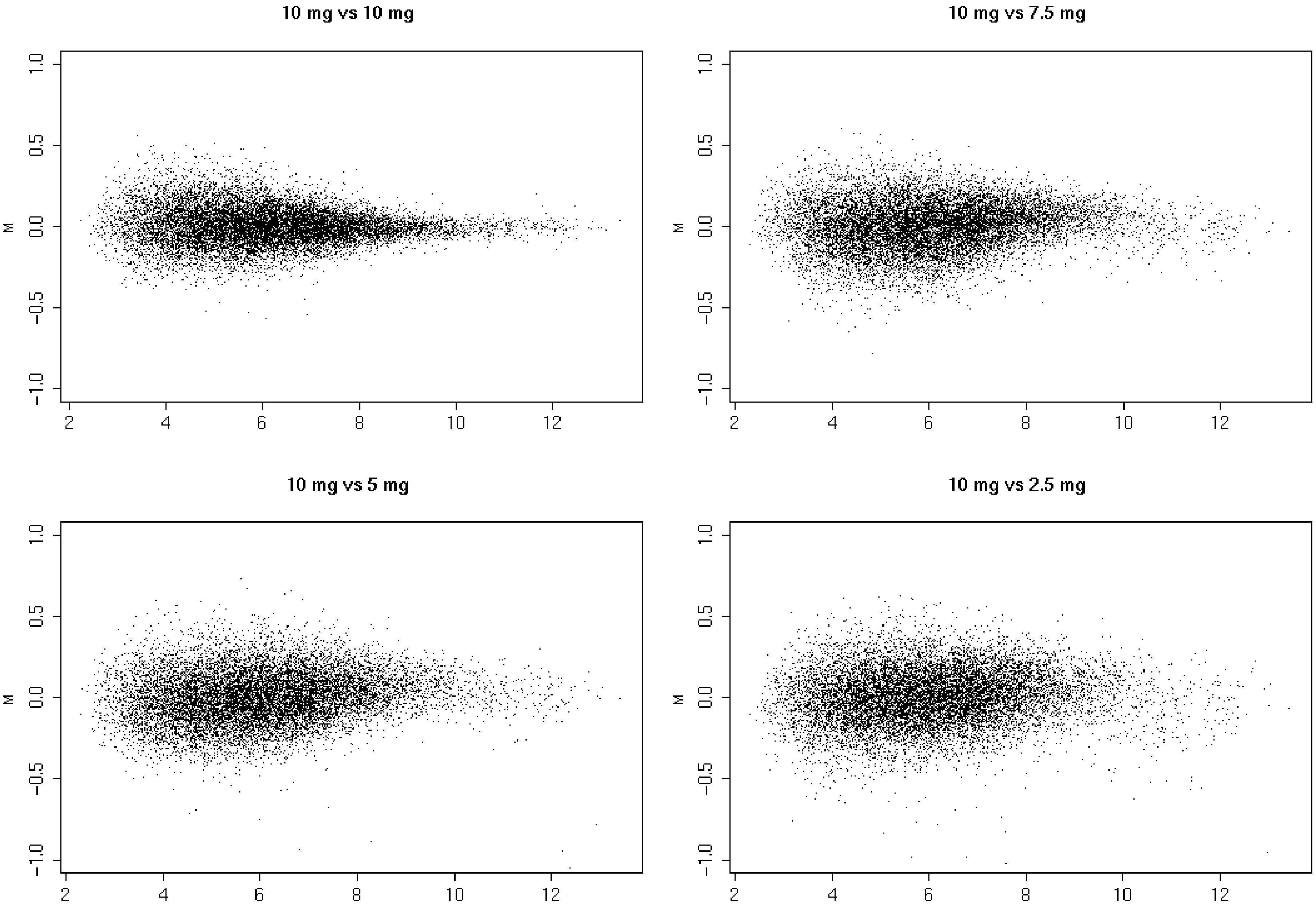
- Recommended amount of cRNA to hybe to chip is 10μg.
- In GLGC dilution have chips with 1.25, 2.5, 5, 7.5, 10 and 20 μg of the same cRNA in replicates of 5

Questions:

- can we use less cRNA?
- can we combine chips with different amounts of cRNA in an experiment?


Assessments of variability for fits at different dilutions
Boxplots of probe set scale estimates relative to median model
0.6
0.8
1.0
1.2
1.4
Boxplots of relative log expressions – w/i group comparisons
-0.2
-0.1
0.0
0.1
0.2
Boxplots of relative log expressions – Reference group = Liver.10
-0.2
-0.1
0.0
0.1
0.2
N=
5
5
5
5
Liver.2.5
Liver.5
Liver.7.5
Liver.10

MVA plots comparing expressions for a 10mg reference chip with 4 other chips



10 mg vs 10 mg

10 mg vs 7.5 mg

10 mg vs 5 mg

10 mg vs 2.5 mg

Where we are?

- We have measures that are good at detecting differences
- Need more actionable information:
- ➢ What is the impact on analysis?
- ➢ What are the causes?
- ➢ Gather more data to move away from relative quality and toward absolute quality.
- ➢ Other levels of quality to investigate – individual probes and probe sets, individual summaries.

Acknowledgements

- Terry Speed and Julia Brettschneider
- Gene Logic, Inc.
- Affymetrix, Inc.
- St-Jude's Children’s Research Hospital

- The BioConductor Project
- The R Project

References

1. Mei, R., et. al. (2003), Probe selection for high-density oligonucleotide arrays, PNAS, 100(20):11237-11242
2. Dai, Hongyue et. al. (2003), Use of hybridization kinetics for differentiating specific from non-specific binding to oligonucleotide microarrays, NAR, Vol. 30, No. 16 e86
3. Irizarry, R. et.al (2003) Summaries of Affymetrix GeneChip probe level data, Nucleic Acids Research, 2003, Vol. 31, No. 4 **e15**
4. Irizarry, R. et. al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, in press.
5. http://www.stjuderesearch.org

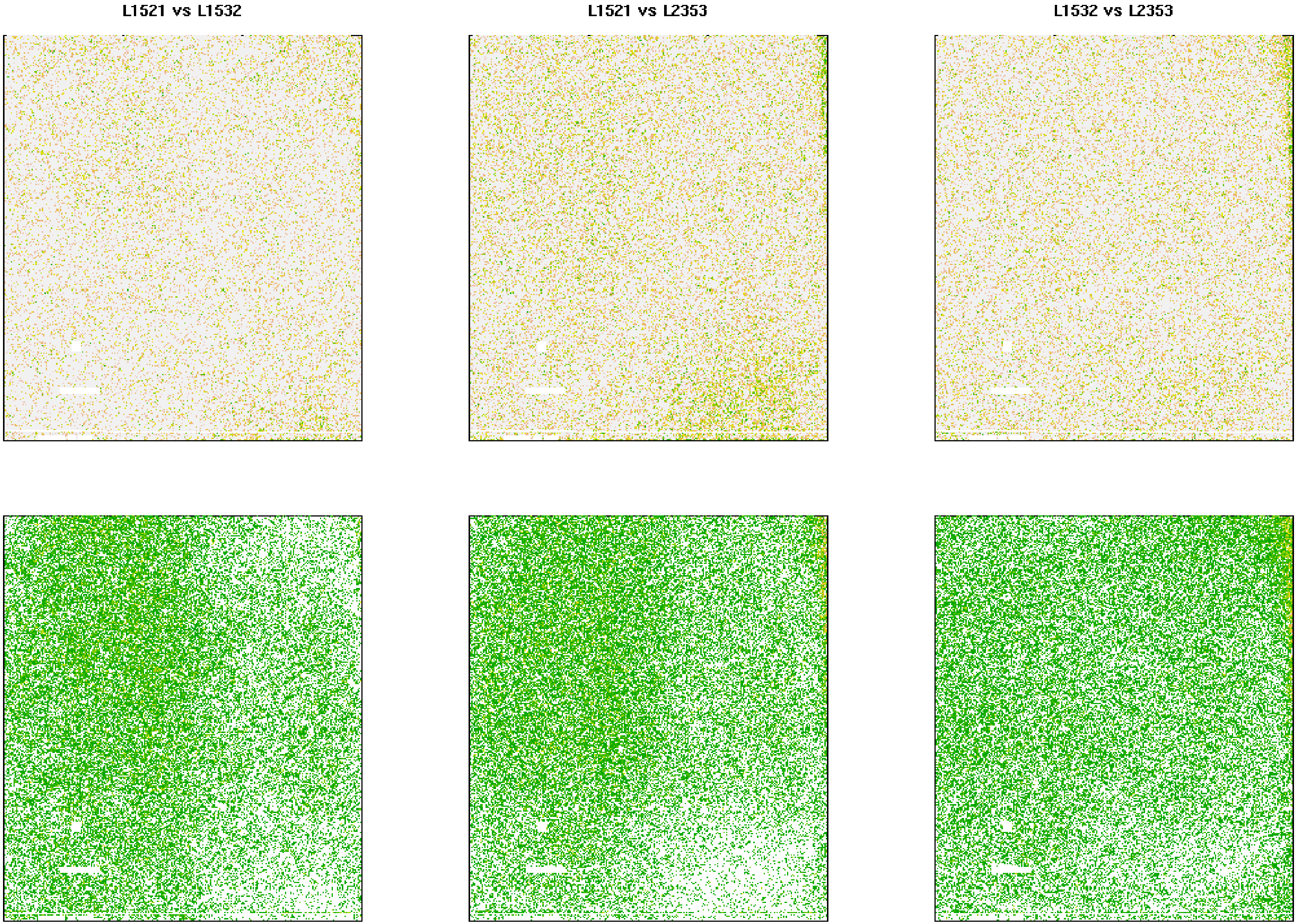
Additional slides

Example – comparing experiments: probe effects

- Affy hg-u95A
- We compare probe effects from models fitted to data from chips from different lots (3 lots)
- For pairs of lots, image est(p_1)-est(p_2) properly scaled and transformed into a weight.
- Also look at sign of difference

Comparing probe effects by spatial location

Models fitted to chips from different lots


L1521 vs L1532
L1521 vs L2353
L1532 vs L2353