Some Principles for the Design and Analysis of Experiments using Gene Expression Arrays and Other High-Throughput Assay Methods

David M. Rocke

Division of Biostatistics, School of Medicine,

Department of Applied Science, College of Engineering, &

Institute for Data Analysis and Visualization

University of California, Davis

ARTP
MIND Institite
May 2005

The -Omics Revolution

The advent of gene expression microarrays, proteomics by mass spectrometry, and metabolomics by mass spectrometry and NMR spectroscopy presents enormous opporunities for fundamental biological research and for applications in medicine, agriculture, and environmental science

They also present many challenges in design and analysis of laboratory experiments, population studies, and clinical trials. We present some lessons learned from our experience with these studies.

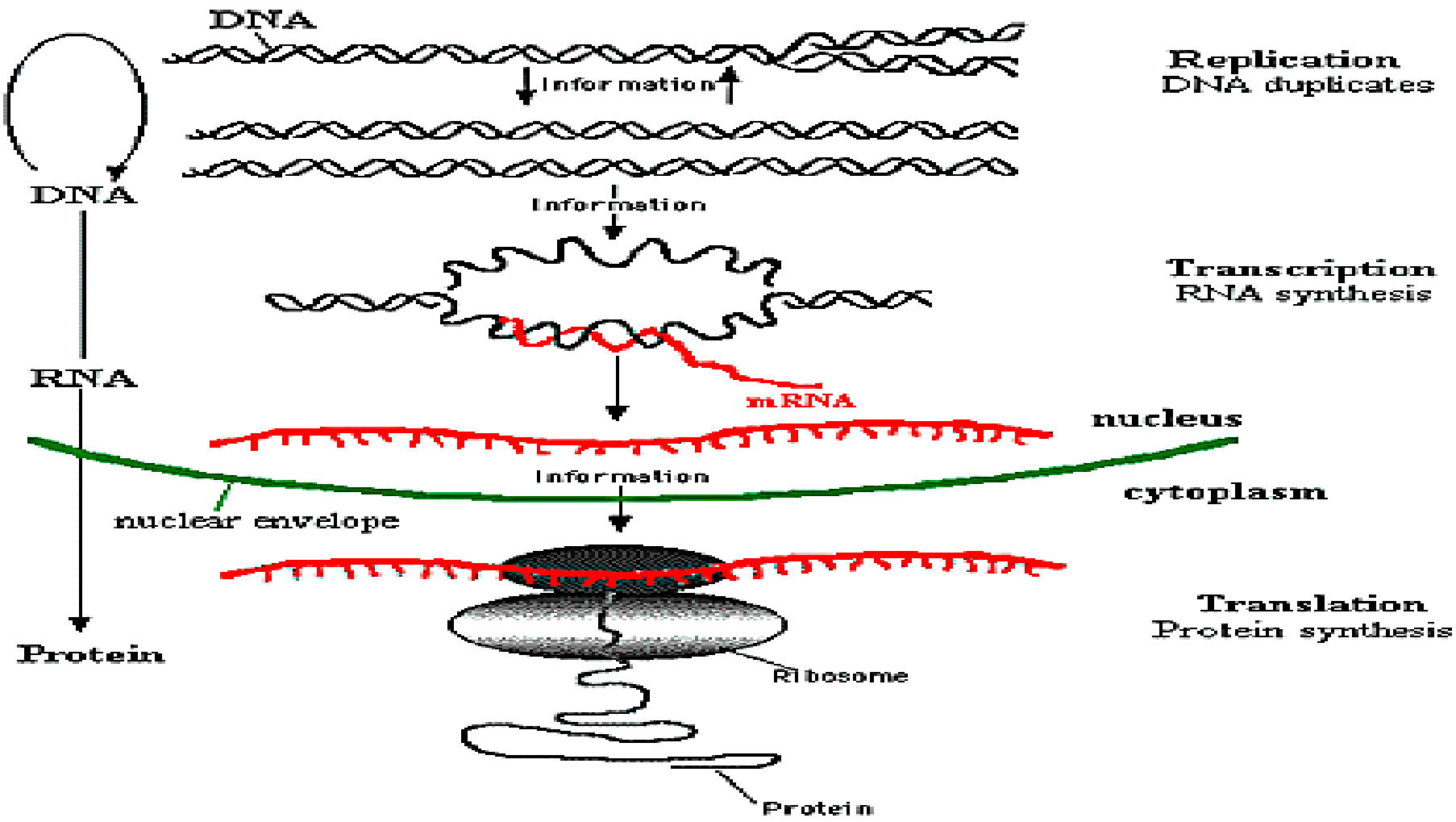
Omics Data

Genome Complement of all genes, or of all components of genetic material in the cell (mostly static).

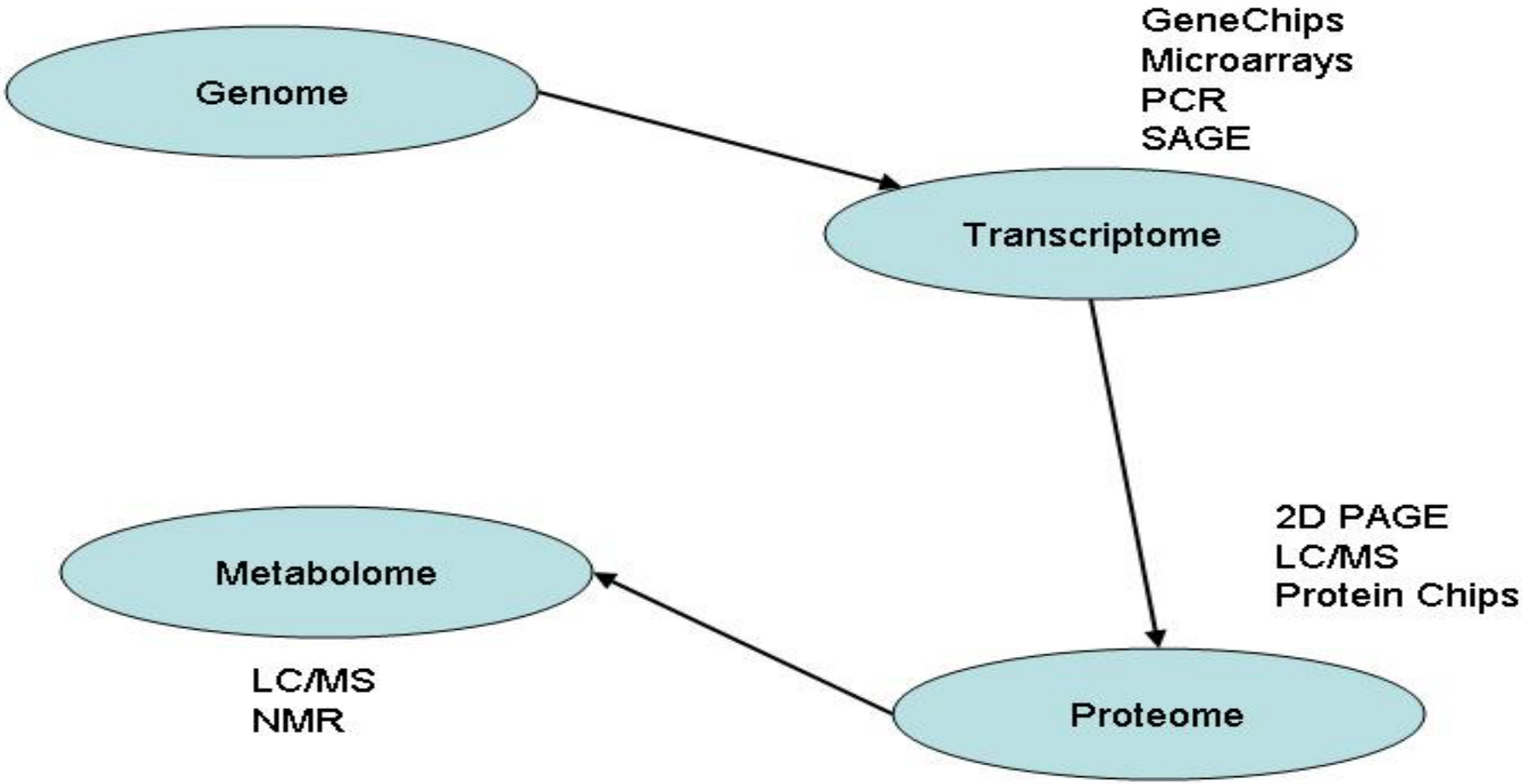
Transcriptome Complement of all mRNA transcripts produced by a cell (dynamic).

Proteome Complement of all proteins in a cell, whether directly translated or produced by post-translational modification (dynamic).

Metabolome Complement of all metabolites other than proteins and mRNA; e.g., lipids, saccharides, etc (dynamic).



The Central Dogma of Molecular Biology


Genome
GeneChips
Microarrays
PCR
SAGE
Transcriptome
2D PAGE
LC/MS
Protein Chips
Metabolome
LC/MS
NMR
Proteome

The Principles of Experimental Design Have not Changed

- A design that is not adequate to measure a change in one indicator across populations is probably not adequate to measure the change in 20,000 indicators.

- Usually, biological variability (within or between organisms) is much larger than the technical variability of measurements.

- Thus, most replications should be across organisms, not repeats of the same sample.

- The measurement of difference between types of cancer, between varieties of wheat, or between animal populations will often require many samples

We Need Internal Controls

- We learned long ago that clinical studies need internal controls to be believable. Comparisons with past history are too frequently deceptive to be useful.

- Genomics data are an obvious exception because the genetic structure of (for example) humans varies only a little between individuals, and mostly varies not at all over time in a given individual.

- Gene expression data, proteomics data, and metabolomics data are more like clinical data than genomics data: they vary over time and over conditions, some of which are hard to measure.

- Databases of expression, proteomics, etc. will mostly be useful as archives of studies; direct comparisons across studies will need to be interpreted cautiously.

- What we hope will be reproducible is differences between groups, not absolute measurements.

Detecting Statistically Significant Effects

- Mostly, we do not yet have quantitative knowledge of what changes in gene expression, protein content, etc. are biologically significant. Until we do have such knowledge, we should detect all changes that we are sure have occurred without regard to size. Twofold may be a large or small change. A 10% change may be important.

- If we measure 10,000 things at once, and test each one for significance, we may have too many false positives to be useful.

- A 5% statistical test will generate an average of 500 false positives in 10,000. If we have 1,000 "significant" genes in tests for differential expression, then about half will likely be "false discoveries."

- One way to control this is to use the Bonferroni method for family-wise error rates, in which each gene is tested at a significance level of 5%/10,000 = 0.000005, or one in 200,000. This guarantees that there will be no genes identified in 19 of 20 studies where there are no real diffences. It may lack sensitivity.

- With a sample of 5 in each of two groups, the smallest difference that is significant at the 5% level is about 1.7 standard deviations. With the Bonferroni adjustment on 10,000 variables, the detectable change is over four times as large (7.5 standard deviations).

False Discovery Rate

- There are a series of False Discovery Rate (FDR) methods that provide good protection but are more sensitive than the Bonferroni Method.

- If there are 10,000 genes and 500 are identified by a 5% FDR method, then approximately 95% of these 500 will be really different and no more than about 5% of them will be false discoveries. This means that only about 25 of the 500 will be false leads.

Experimental Design

- Often investigating multiple factors in the same experiment is better. We can use a full factorial design (all possible combinations) or a fractional factorial. Fractional factorial designs can investigate as many as 7 factors in 8 experiments, each one with the full precision of a comparison of 4 vs. 4.

- Consider a study of the response of mice to a toxic insult. We can examine 2 ages of mice, 2 sexes, treatment and control, for a total of eight conditions. With 2 mice per condition, we are well placed to investigate even complex relationships among the three factors.

- Two color arrays generate more complexity in the design, with possible dye bias, and with the most accurate comparisions being between the two samples on the same slide.

The Analysis of Variance

- The standard method of analyzing designs with categorical variables is the analysis of variance (ANOVA).

- The basic principle is to compare the variability of group means with an estimate of how big the variability could be at random, and conclude the difference is real if the ratio is large enough.

- Consider an example with four groups and two measurements per group.

Example Data

Group	Sample 1	Sample 2	Mean
A	2	4	3
B	8	10	9
C	14	16	15
D	20	22	21

- The variabiliy among the four group means is 120 (Mean Square for groups). This has three degrees of freedom.

- The variability within groups is 2 (Mean Square Error or MSE). This has four degrees of freedom.

- The significance of the ratio uses the F distribution. The more df in the MSE, the more sensitive the test is.

- The observed F ratio of $120/2 = 60$ is highly significant. If there were no real difference, the F ratio would be near 1.

Measurement Scales

- Standard statistical methods are additive: we compare differences of means.

- Often with gene expression data and other kinds of assay data we prefer ratios to means.

- This is equivalent to taking logarithms and using differences.

$$\log(x/y) = \log(x) - \log(y)$$

- In general, we often take logs of data and then use regression, ANOVA and other standard (additive) statistical methods. High-throughput assay data require some alteration in this method.

Variation in Microarry and other Omics Data

Some well known properties of measurement error in gene expression microarrays: include the following:

- For high gene expression, the standard deviation of the response is approximately proportional to the mean response, so that the CV is approximately constant.

- For low levels of expression, the CV is much higher.

- Expression is commonly analyzed on the log scale, so that for high levels the SD is approximately constant, but for low levels of expression it rises.

- Comparisons of expression are usually expressed as n-fold, corresponding to the ratio of responses, of which the logarithm would be well behaved, but only if both genes are highly expressed.

- These phenomena occur in many measurement technologies, but are more important in high-throughput assays like microarrays.

- What is the fold increase when a gene goes from zero expression in the control case to positive expression in the treatment case?

- Which is biologically more important: an increase in expression from 0 to 100 or an increase from 100 to 200?

Variance Model for Gene Expression and other Omics Data

At high levels, the standard deviation of replicates is proportional to the mean. If the mean is μ, then this would be

$$\begin{aligned} \mathrm{SD}(y) &= b\mu \\ \mathrm{Var}(y) &= b^2\mu^2 \end{aligned}$$

- But this cannot hold for unexpressed genes, or in general for assays where the true concentration is 0.

- So a reasonable model for the variance of microarray data is

$$\text{Var}(y) = a^2 + b^2\mu^2$$

(Rocke and Durbin 2001).

Often, the observed intensity (peak area, etc.) needs to be corrected for background or baseline by subtraction of the average signal α corresponding to genes unexpressed (compounds not present) in the sample. This may be a single number, a single number per slide, or a more complex expression. This can be estimated from negative controls or by more complex methods.

So if y is the signal, and $z = y - \alpha$ is the background corrected signal, our mean/variance model is

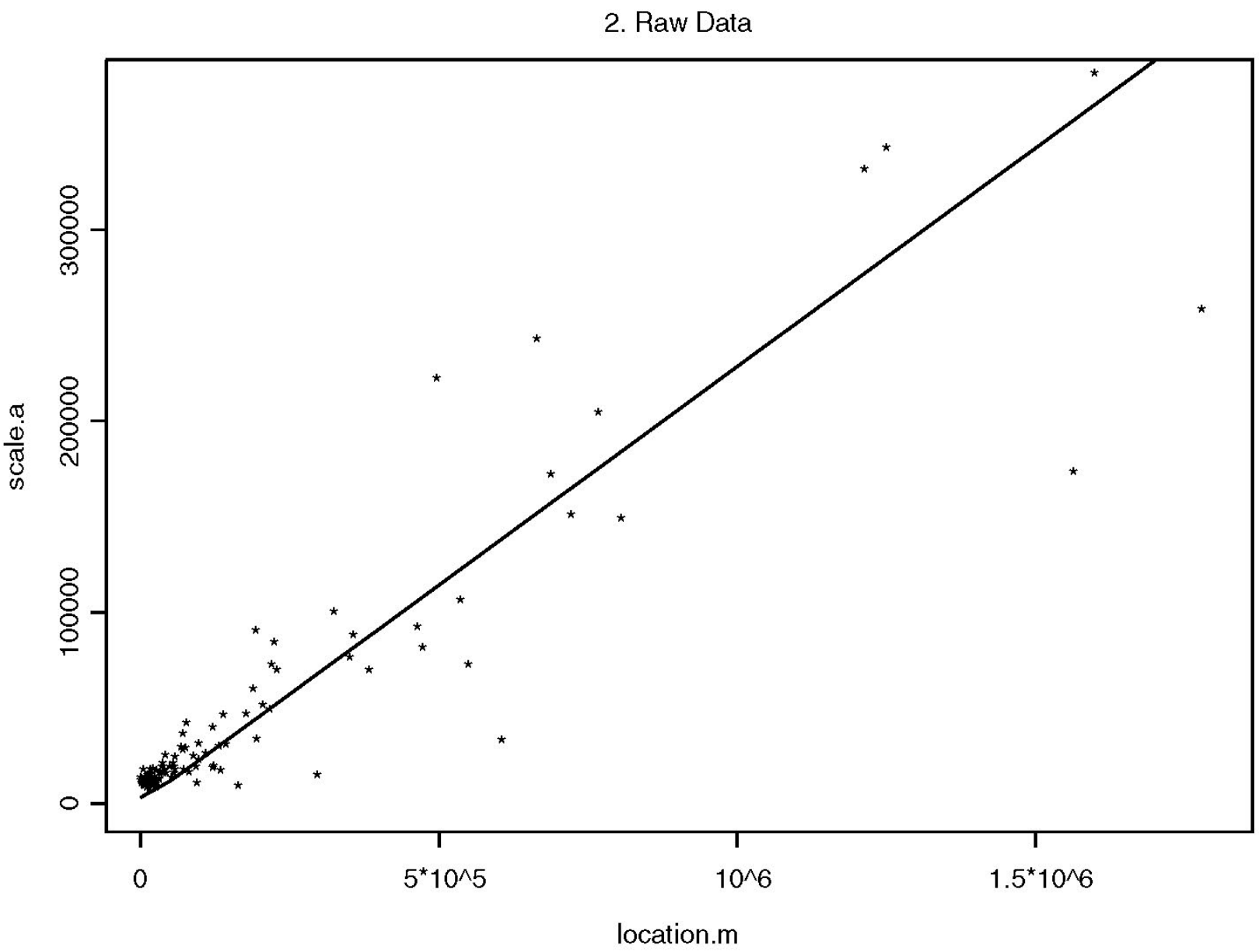
$$\begin{aligned} E(z) &= \mu \\ V(z) &= a^2 + b^2\mu^2 \end{aligned}$$

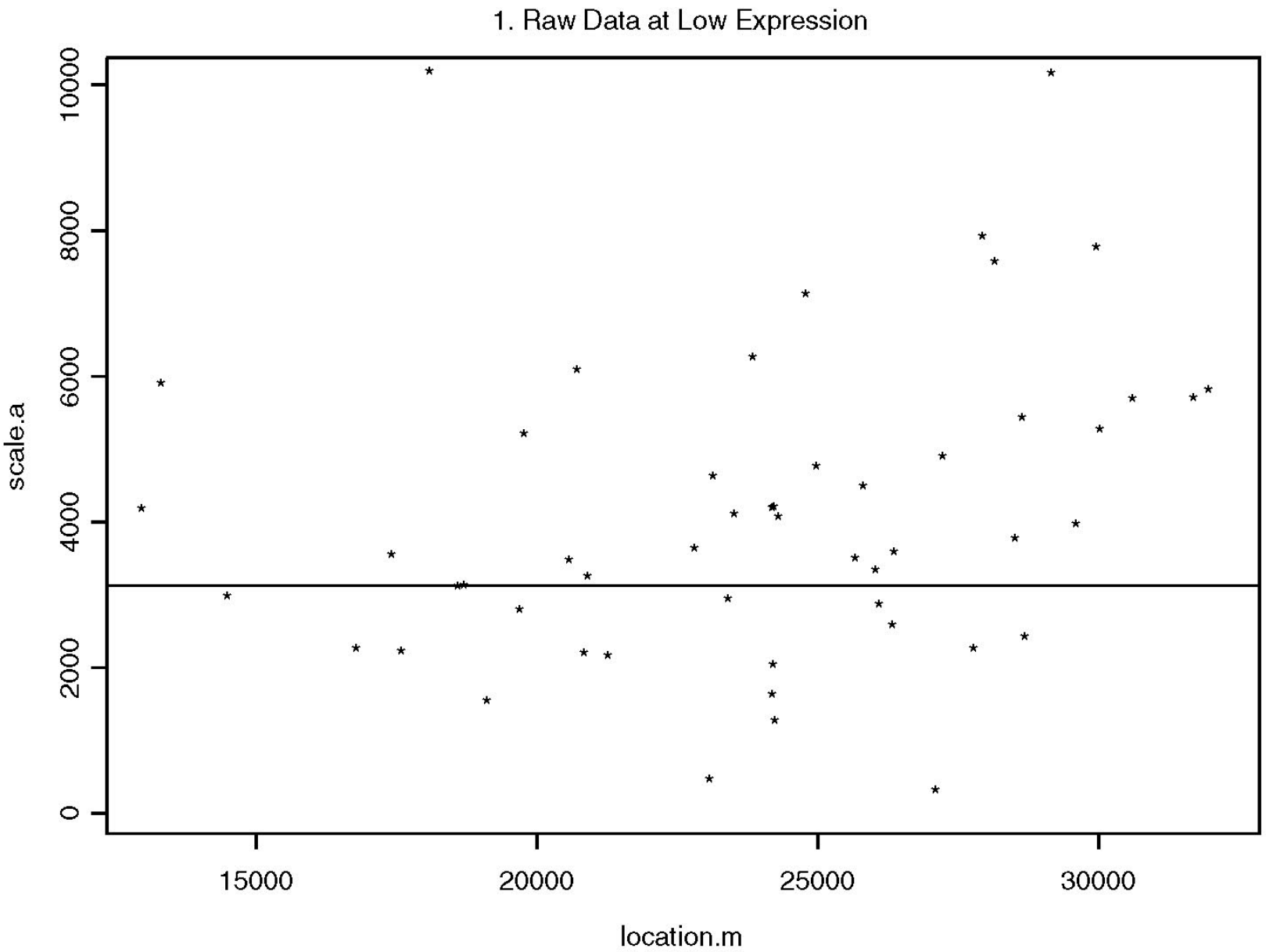
It can be shown that

$$\text{Var}\{\ln(y - \alpha)\} \approx \sigma_\eta^2 + \sigma_\epsilon^2/\mu^2.$$

An Example

We illustrate this with one slide from an experiment on the response of male Swiss Webster mice to a toxic substance. The treated animal received 0.15mg/kg ip of Naphthoflavone, while the control mouse had an injection of the carrier (corn oil). Genes were replicated usually eight times per slide.


2. Raw Data
scale.a
0
100000
200000
300000
0
5*10^5
10^6
1.5*10^6
location.m


1. Raw Data at Low Expression
scale.a
0
2000
4000
6000
8000
10000
15000
20000
25000
30000
location.m

Data Transformation

- Logarithms stabilize the variance for high levels, but increase the variance for low levels.

- Log expression ratios have constant variance only if both genes are expressed well above background.

- Heterogeneity of variance is an important barrier to reliable statistical inference

- Such heterogeneity is common in biological data, including gene expression data

- Data transformations are a well-known way of dealing with this problem

- We present a new transformation family that is expressly designed for biological data, and which appears to work very well on gene expression data

- The logarithm is designed to stabilize data when the standard deviation increases proportional to the mean.

- When the data cover a wide range down to zero or near zero, this transformation performs poorly on low level data. This does not mean that these data are "bad" or "highly variable" or "unreliable". It only means that we are using the wrong transformation or measurement scale.

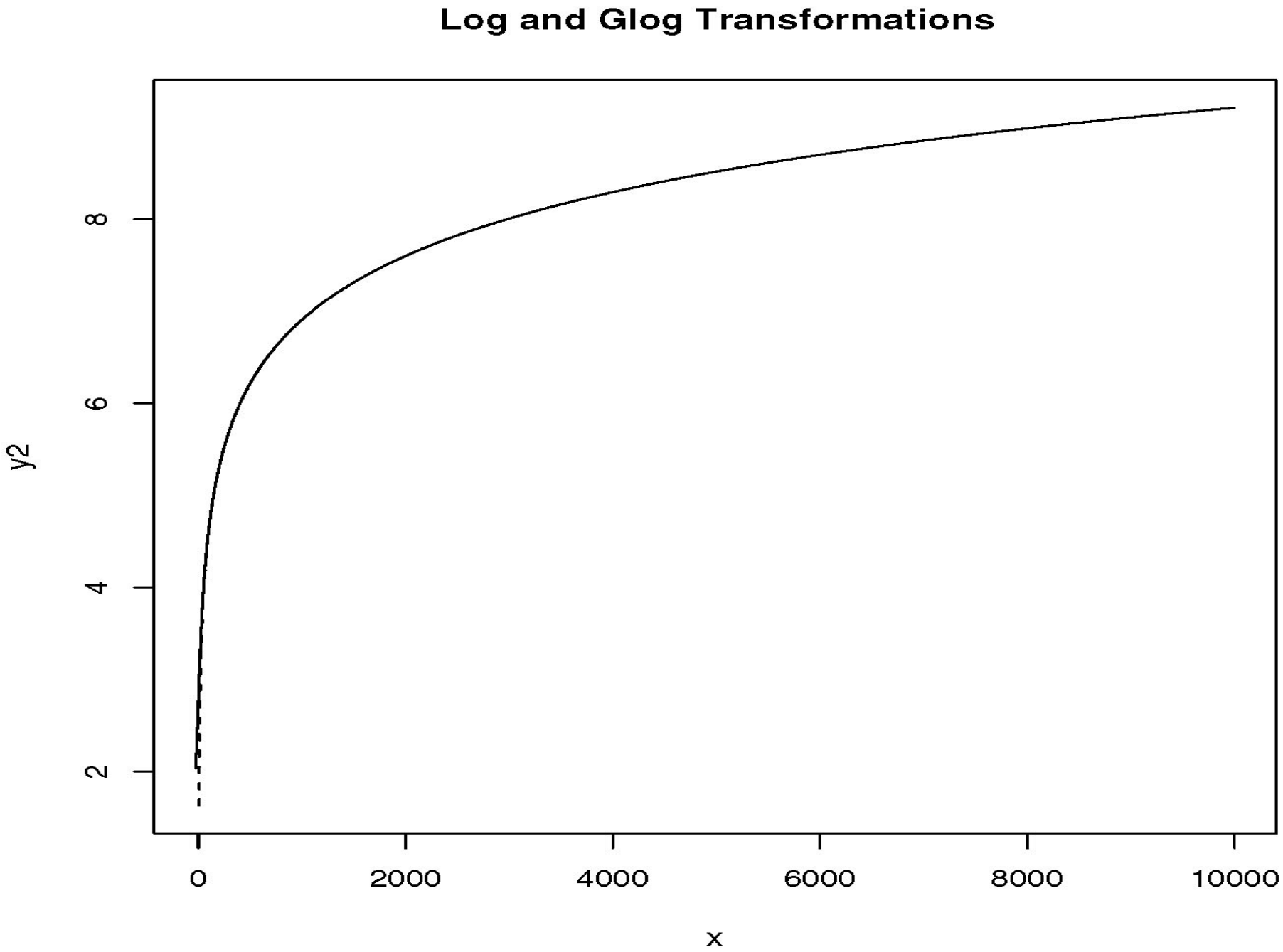
The *generalized logarithm* reproduces the logarithm at high levels, but behaves better at low levels. One way to express it is

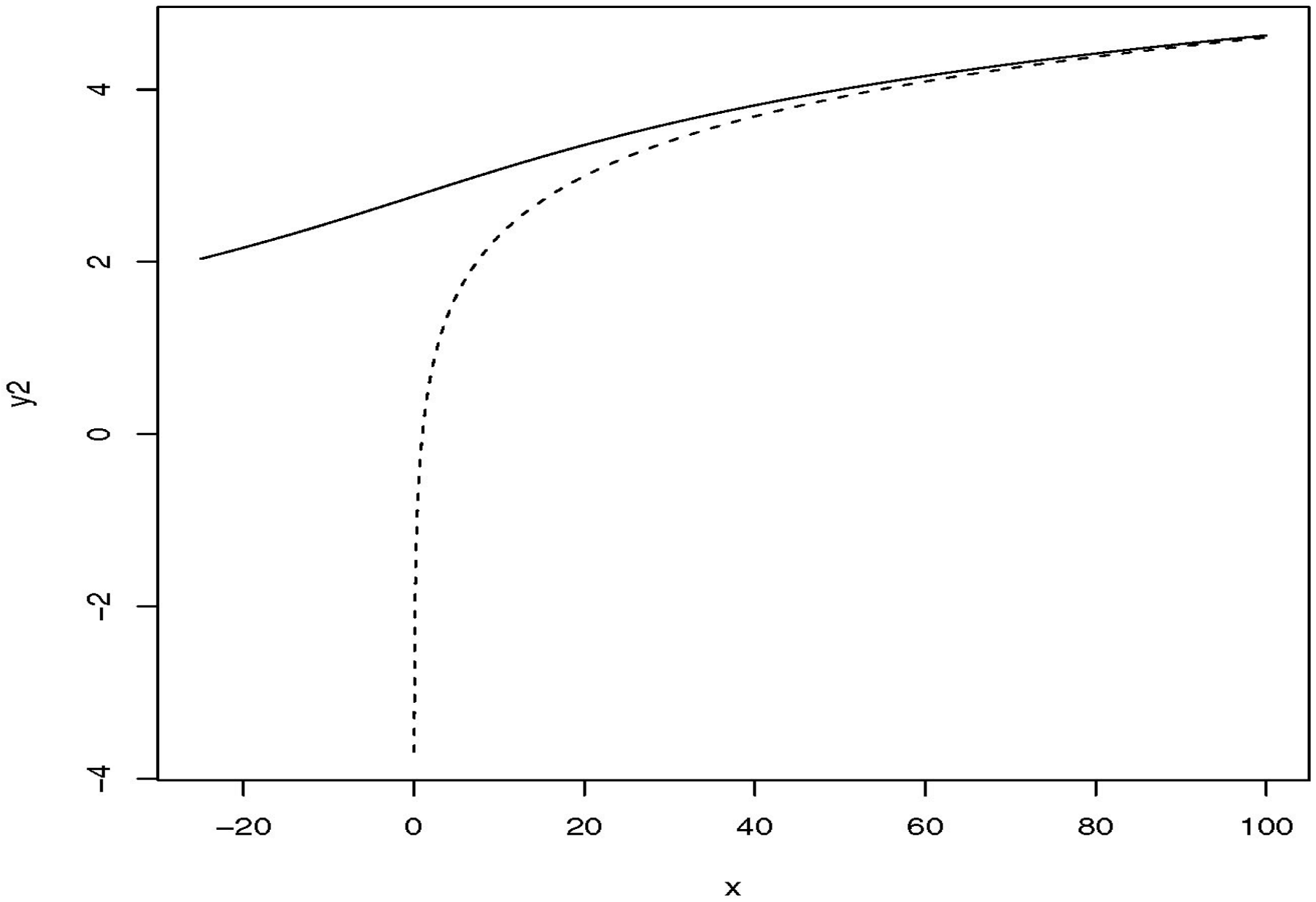
$$f(z) = \ln(z + \sqrt{z^2 + a^2/b^2})$$

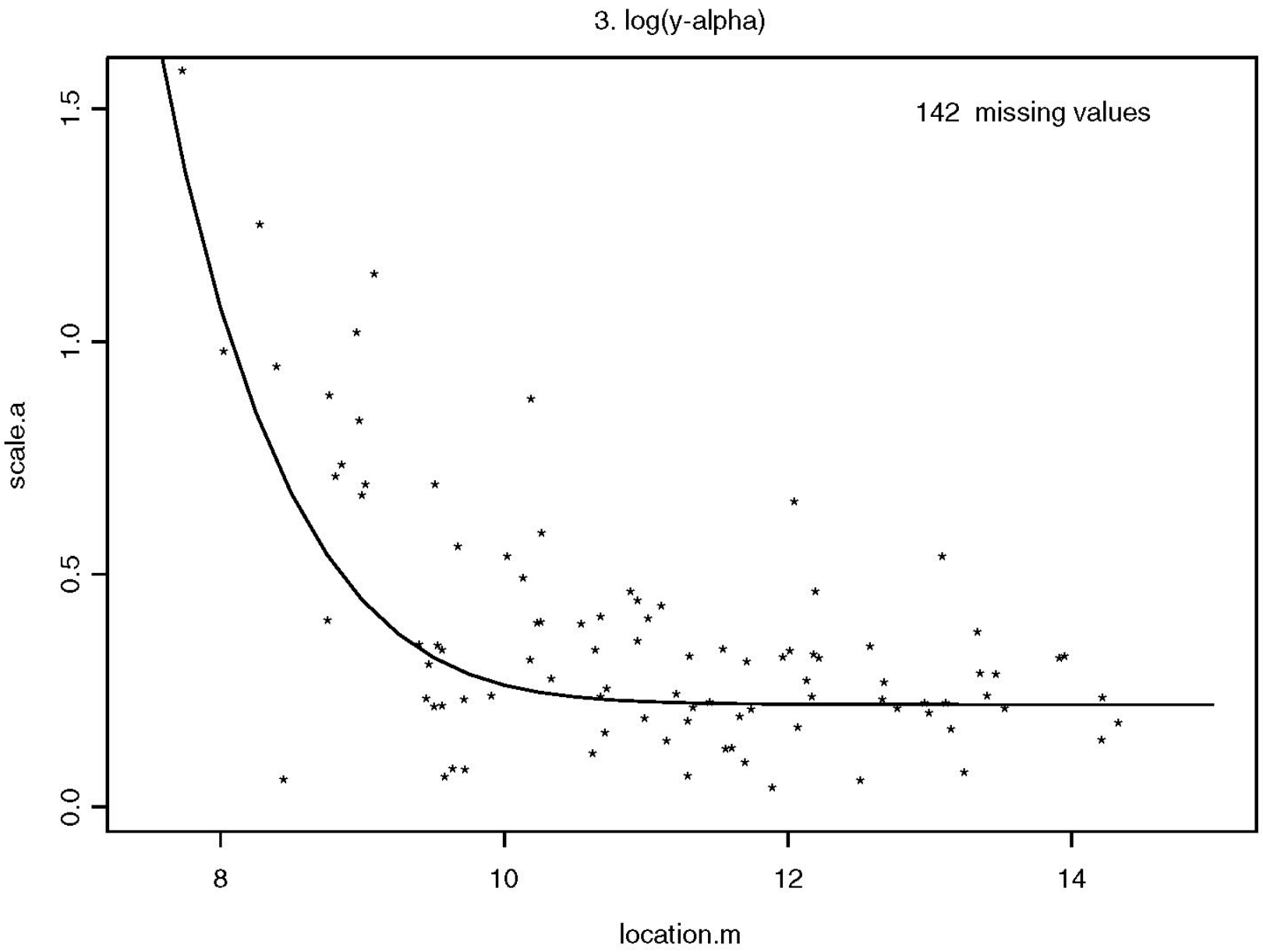
where z is the background-corrected intensity. (Durbin, Hardin, Hawkins, and Rocke 2002; Hawkins 2002; Huber, von Heydebreck, Sültmann, Poustka, and Vingron 2002; Munson 2001)

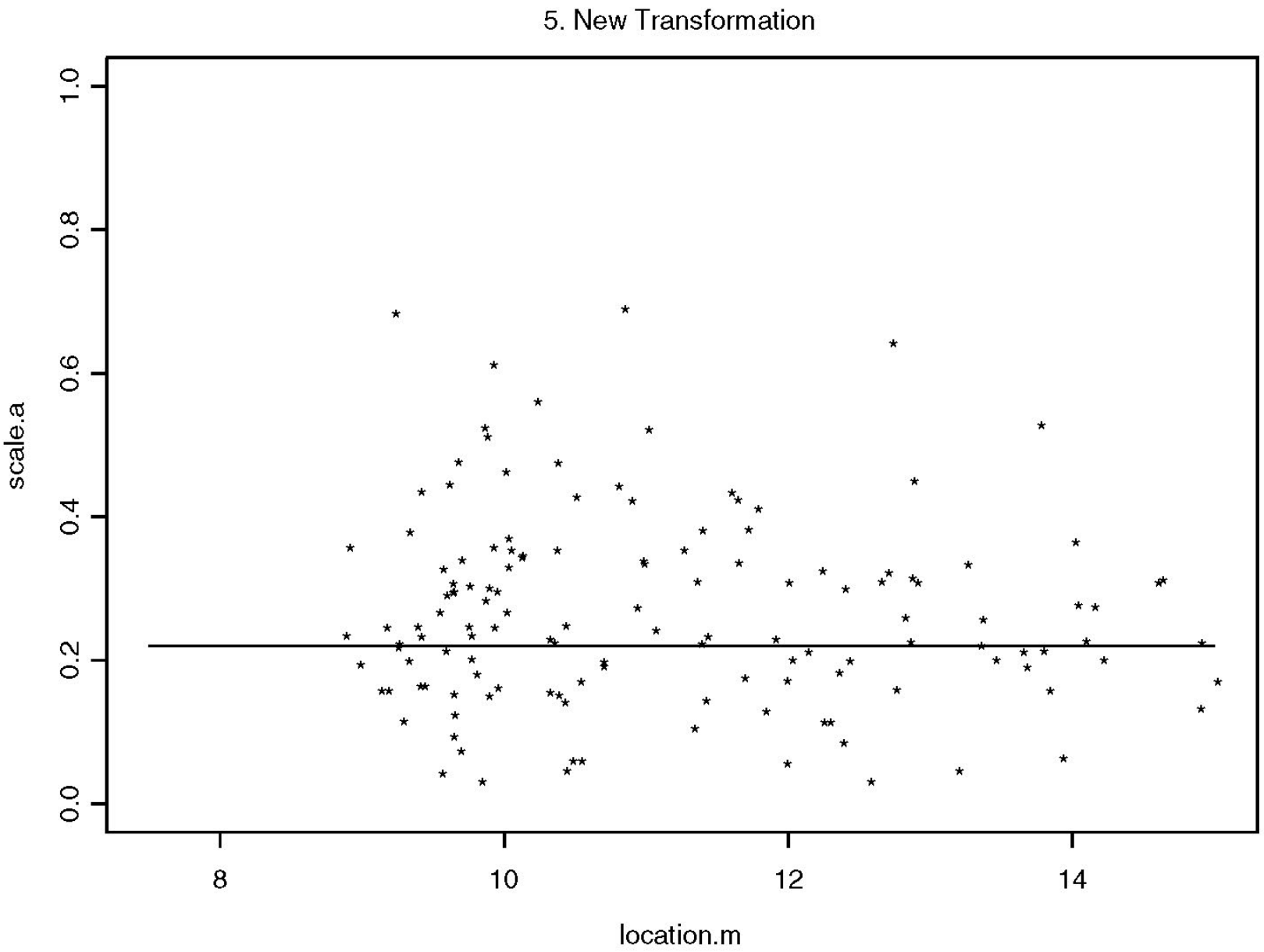
$$f(z) = \mathrm{ln}(z + \sqrt{z^2 + a^2/b^2})$$

- $f(z) \sim \mathrm{ln}(z)$ for large z.
- $f(z)$ is approximately linear for $z = 0$.
- $f(z)$ is monotonic (does not change the order of size of data).


Log and Glog Transformations
y2
2
4
6
8
0
2000
4000
6000
8000
10000
x


Log and Glog Transformations at Low Levels
y2
4
2
0
−2
−4
−20
0
20
40
60
80
100
x


3. log(y-alpha)
142 missing values
scale.a
0.0
0.5
1.0
1.5
8
10
12
14
location.m


5. New Transformation
scale.a
0.0
0.2
0.4
0.6
0.8
1.0
8
10
12
14
location.m

Estimation

This transformation has one parameter that must be estimated, as well as the background. We can do this in various ways.

$$h_{\lambda,\alpha}(y) = \ln\left(y - \alpha + \sqrt{(y-\alpha)^2 + \lambda}\right).$$

- We can background correct beforehand, or estimate the background and transformation parameter in the same step.

- We can estimate $\lambda = a^2/b^2$ by estimating the low-level variance a^2 and the high-level square CV b^2, and take the ratio.

- We can estimate the parameters in the context of a model using standard statistical estimation procedures like maximum likelihood.

- We can estimate the transformation each time, or use values estimated with a given technology in a given lab for further experiments.

This helps solve the puzzle of comparing a change from 0 to 40 to a change from 1000 to 1600. Suppose that the standard deviation at 0 is 10, and the high-level CV is 15%. Then

- A change from 0 to 40 is four standard deviations ($4 \times 10 = 40 = 40 - 0$).

- A change from 1000 to 1600 is also four standard deviations ($1600/1000 = 160\% =$ increase of $4 \times 15\%$).

- So is a change from 10,000 to 16,000 ($16,000/10,000 = 160\% =$ increase of $4 \times 15\%$).

- The biological significance of any of these is unknown. Different transcripts can be active at vastly different levels.

- But the glog transformation makes an equal change equally statistically significant.

Normalization and Transformation of Arrays

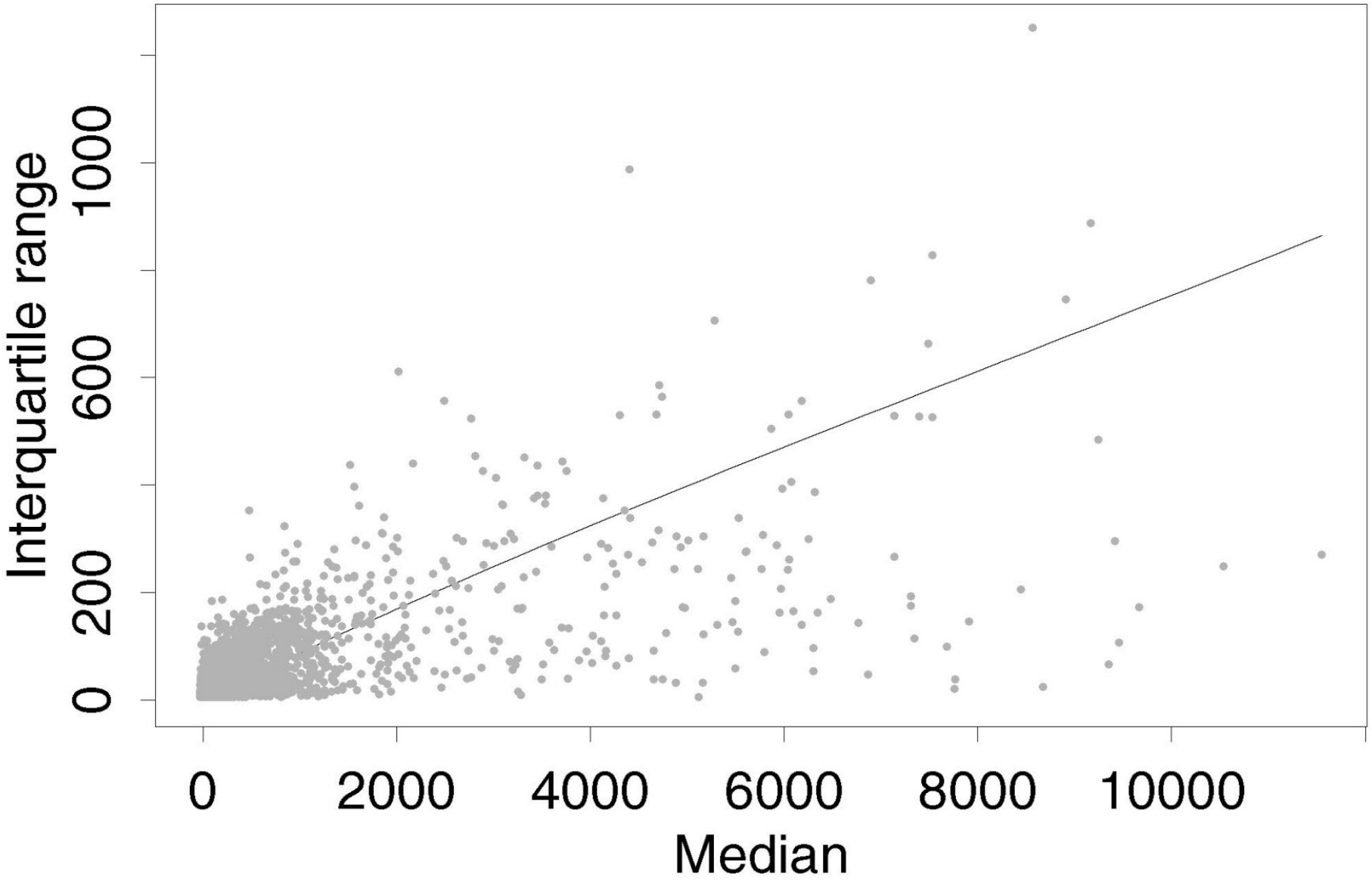
Given a set of replicate chips from the same biological sample, we can simultaneously determine the transformation parameter and the normalization.

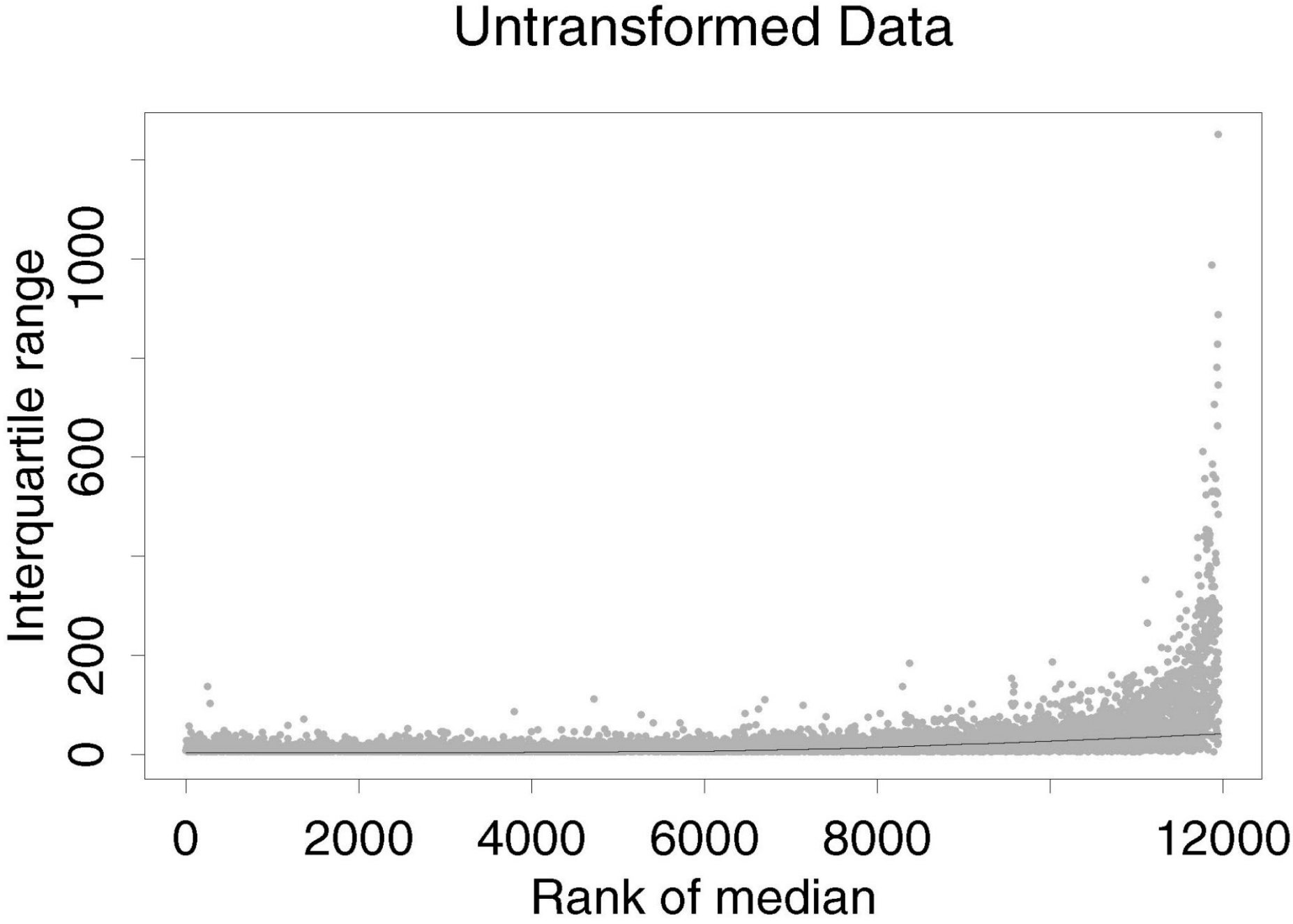
The statistical model used is

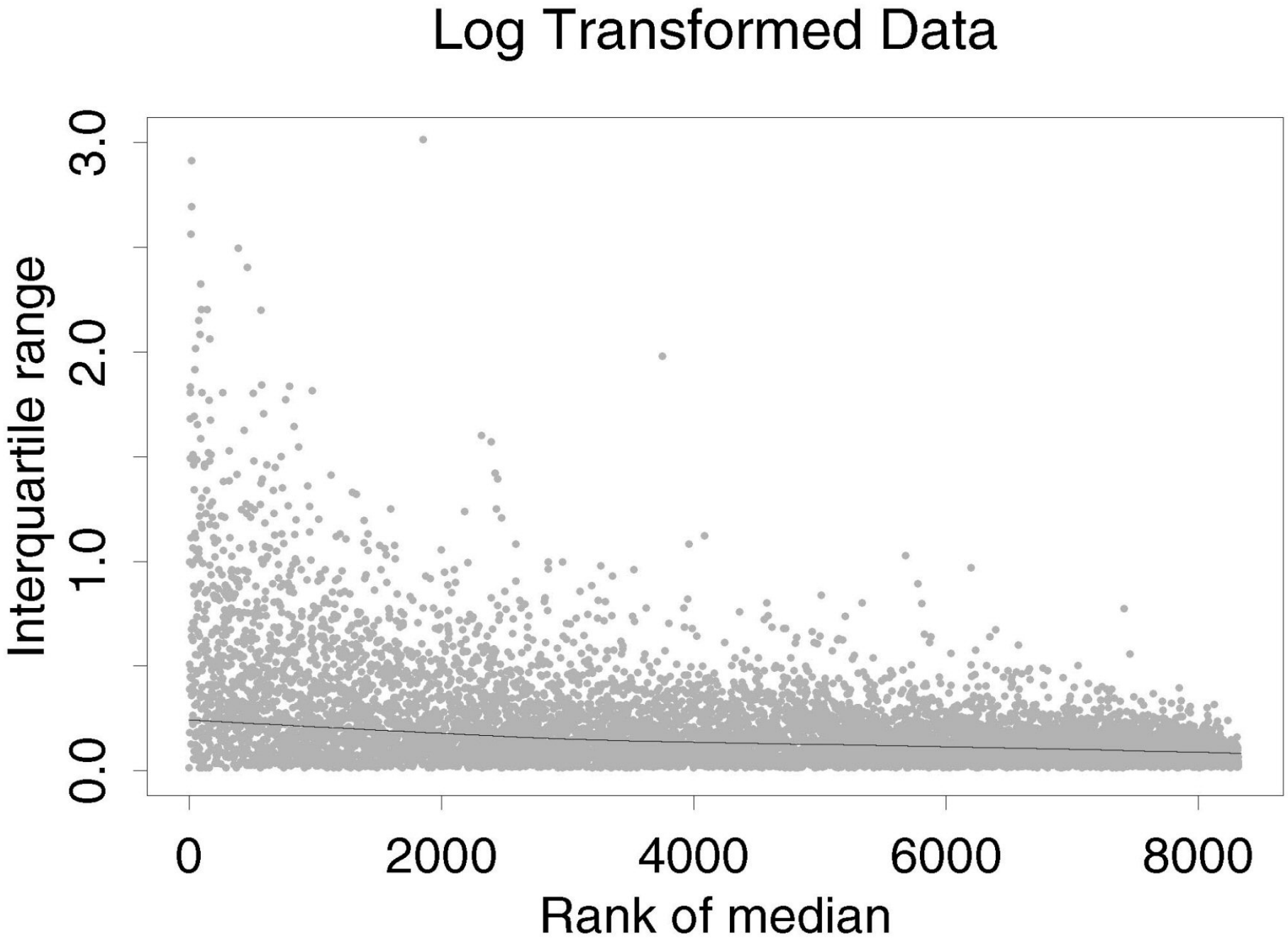
$$h_{\lambda,\alpha}(\text{intensity}) = \text{gene} + \text{chip} + \text{error}$$

and we can estimate the transformation, the gene effects, and the normalization together.

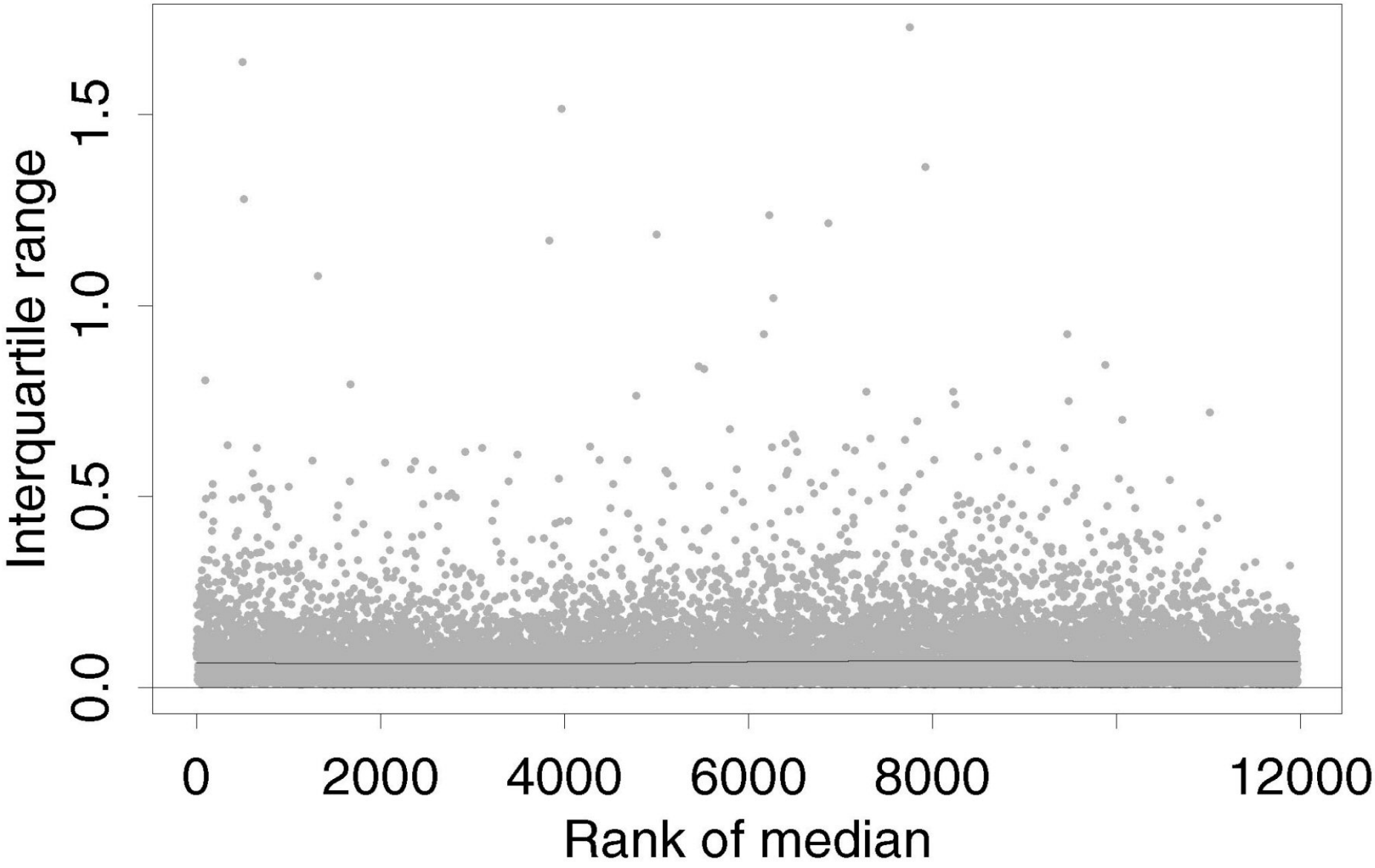
Untransformed Data




Untransformed Data
Interquartile range
0
200
600
1000
0
2000
4000
6000
8000
12000
Rank of median


Log Transformed Data
Interquartile range
0.0
1.0
2.0
3.0
0
2000
4000
6000
8000
Rank of median

Transformed Data with no Flask Effects



Two color Arrays

We wish to model two-color arrays so that after transformation the model is linear. We have a red and a green reading from each spot. Our transformation model is as follows:

$$
\begin{aligned}
h_{\lambda,\alpha}(\text{red intensity}) &= \text{red gene} + \text{chip} \\
&= + \text{spot} + \text{red error} \\
h_{\lambda,\alpha}(\text{green intensity}) &= \text{green gene} + \text{chip} \\
&= + \text{spot} + \text{green error}
\end{aligned}
$$

$$h_{\lambda,\alpha}(\text{red intensity}) - h_{\lambda,\alpha}(\text{green intensity}) =$$

$$\text{red gene} - \text{green gene} + \text{errors}$$

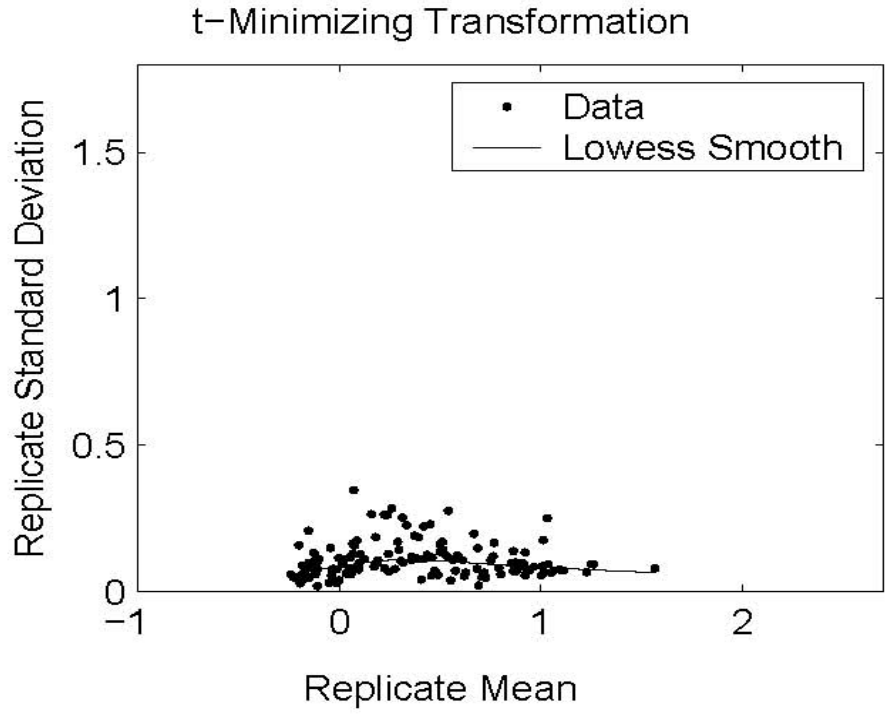
- $h_{\lambda,\alpha}(\text{red}) - h_{\lambda,\alpha}(\text{green})$ looks like the log ratio at high levels, but makes sense also at low levels. We call this a generalized log ratio.

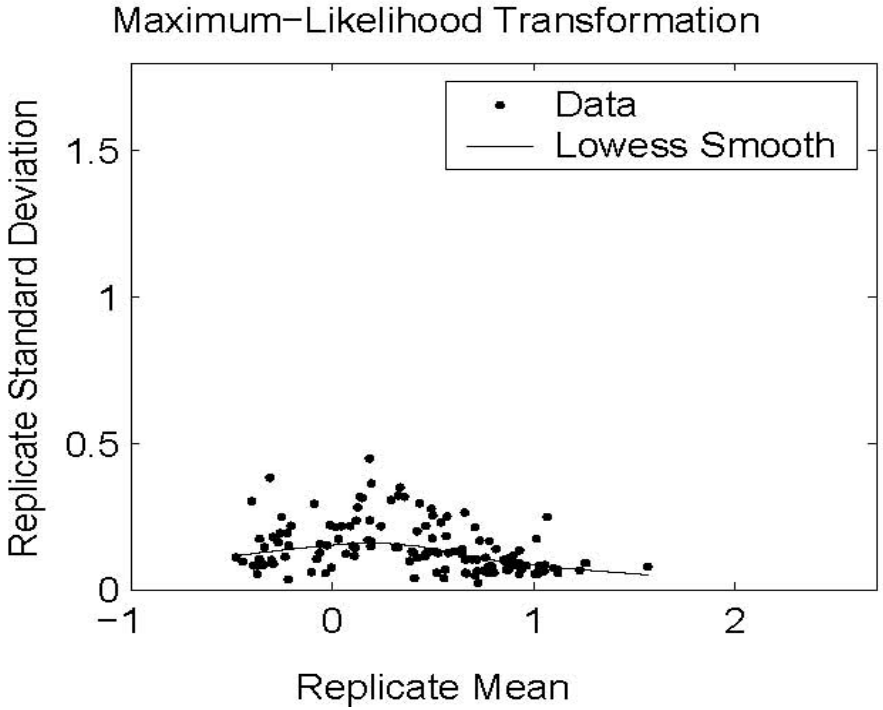
- Can solve iteratively for the transformation parameters and the gene and slide effects, for example by maximum likelihood.

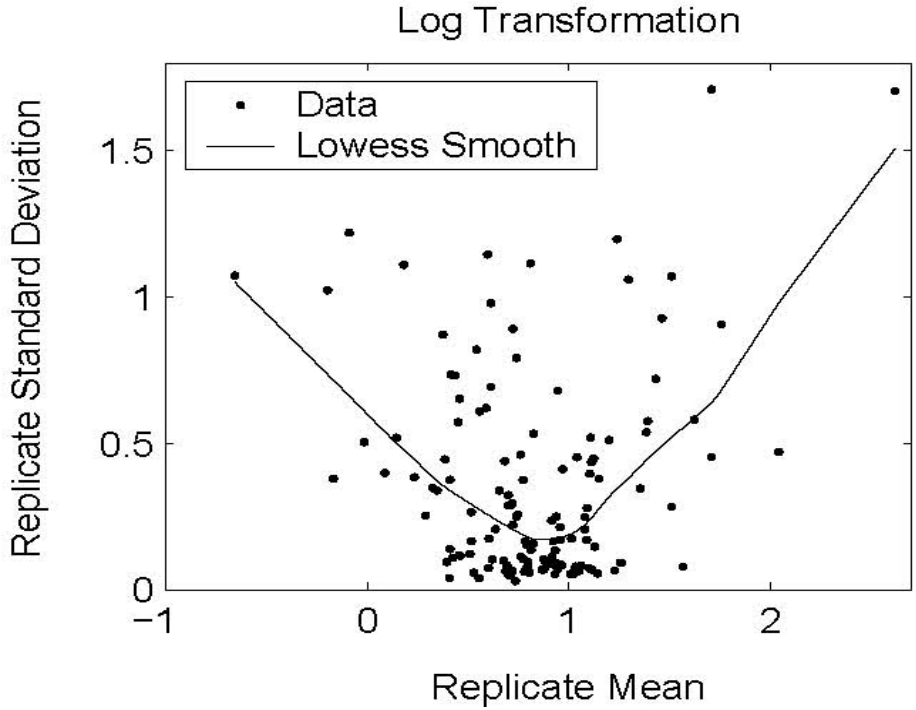
- Dye swap designs remove one source of bias.

- Exercise flexibility in choice of two types of samples (e.g., loop designs).

Figure 4: Replicate Mean and Standard Devation of Differences of Transformed Observations, Three Different Transformations







Determining Differentially Expressed Genes

Consider an experiment on four types of cell lines A, B, C, and D, with two samples per type, each of the eight measured with an Affymetrix U95A human gene array. We have a measured intensity for each gene for each sample (array) in each group. The measured expression is derived from the mean glog-transformed PM probes.

Steps in the Analysis

- Background correct each array so that 0 expression corresponds to 0 signal.

- Transform the data to constant variance using a suitably chosen glog or alternative transformation (started log, hybrid log).

- Normalize the chips additively.

- The transformation should remove systematic dependence of the gene-specific variance on the mean expresssion, but the gene-specific variance may still differ from a global average. Estimate the gene-specific variance using all the information available.

- Test each gene for differential expression against the estimate of the gene-specific variance. Obtain a p-value for each gene.

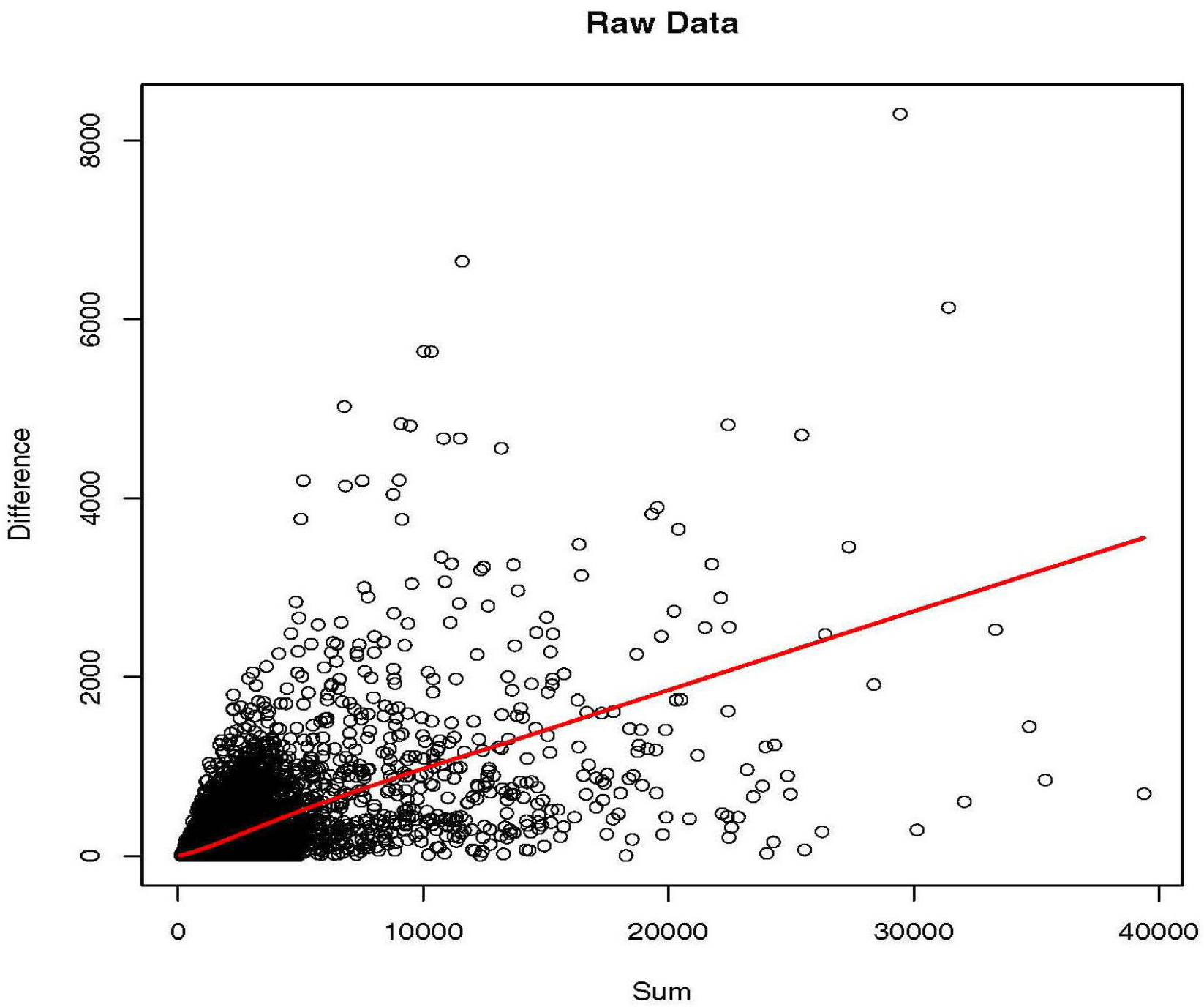
- Adjust p-values for multiplicity using, for example, the False Discovery Rate method.

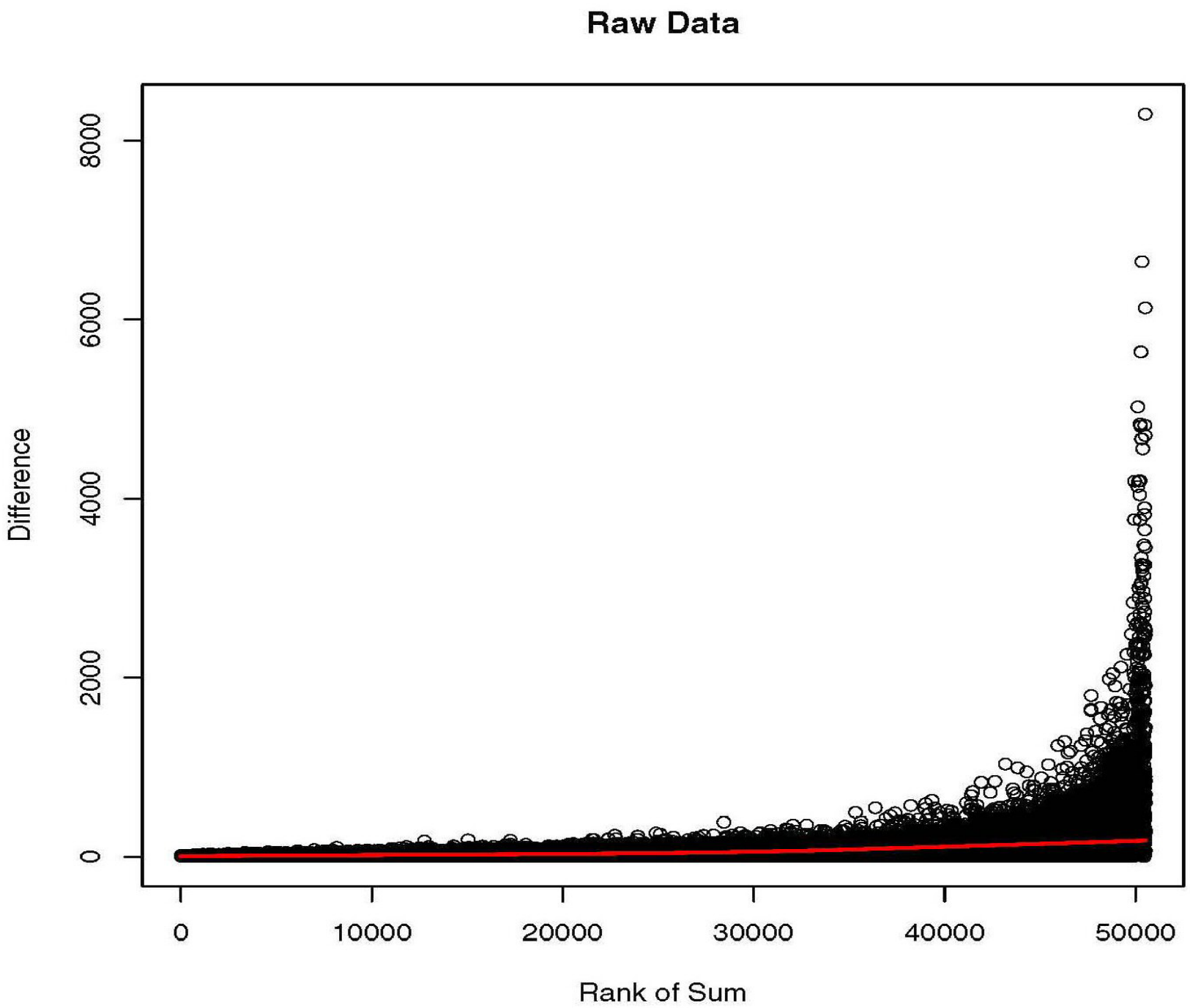
- Provide list of differentially expressed genes

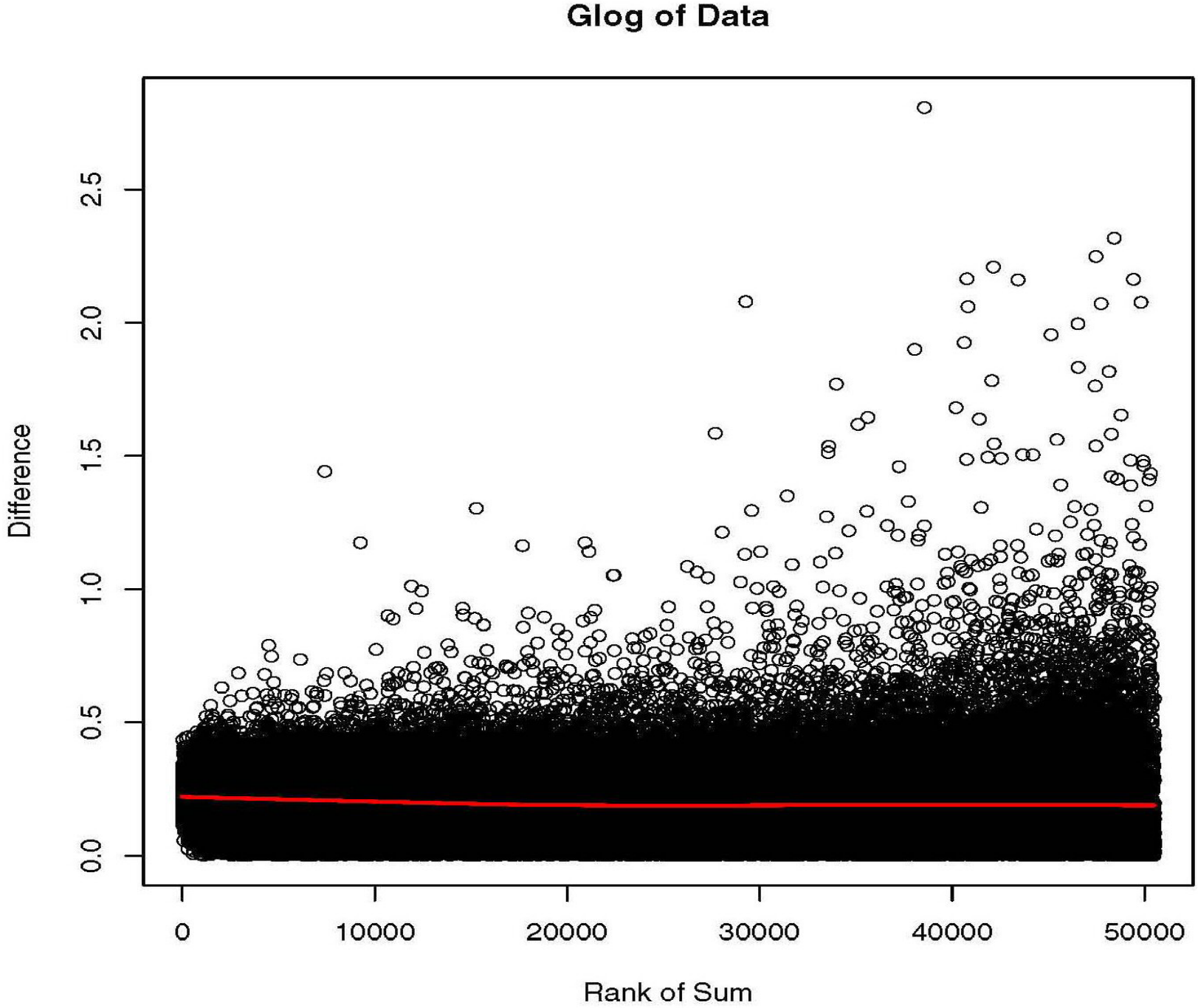
- Investigate identified genes statistically and by biological follow-up experiments.

Structure of Example Data

Gene	Group 1		Group 2		Group 3		Group 4	
ID	1	2	3	4	5	6	7	8
1	y_{111}	y_{112}	y_{123}	y_{124}	y_{135}	y_{136}	y_{147}	y_{148}
2	y_{211}	y_{212}	y_{223}	y_{224}	y_{235}	y_{236}	y_{247}	y_{248}
3	y_{311}	y_{312}	y_{323}	y_{324}	y_{335}	y_{336}	y_{347}	y_{348}
4	y_{411}	y_{412}	y_{423}	y_{424}	y_{435}	y_{436}	y_{447}	y_{448}
5	y_{511}	y_{512}	y_{523}	y_{524}	y_{535}	y_{536}	y_{547}	y_{548}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮


Raw Data
Difference
Sum
0
2000
4000
6000
8000
0
10000
20000
30000
40000


Raw Data
Difference
0
2000
4000
6000
8000
0
10000
20000
30000
40000
50000
Rank of Sum


Glog of Data
Difference
0.0
0.5
1.0
1.5
2.0
2.5
0
10000
20000
30000
40000
50000
Rank of Sum

The model we use is

$h_{\lambda,\alpha}(\text{intensity}) = \text{gene+chip+gene-by-group+error}$

For a given gene, this model is

$$z = \text{group} + \text{error}$$

where z is the transformed, chip-normalized data for the given gene. (Kerr, Martin, and Churchill 2001; Kerr 2003)

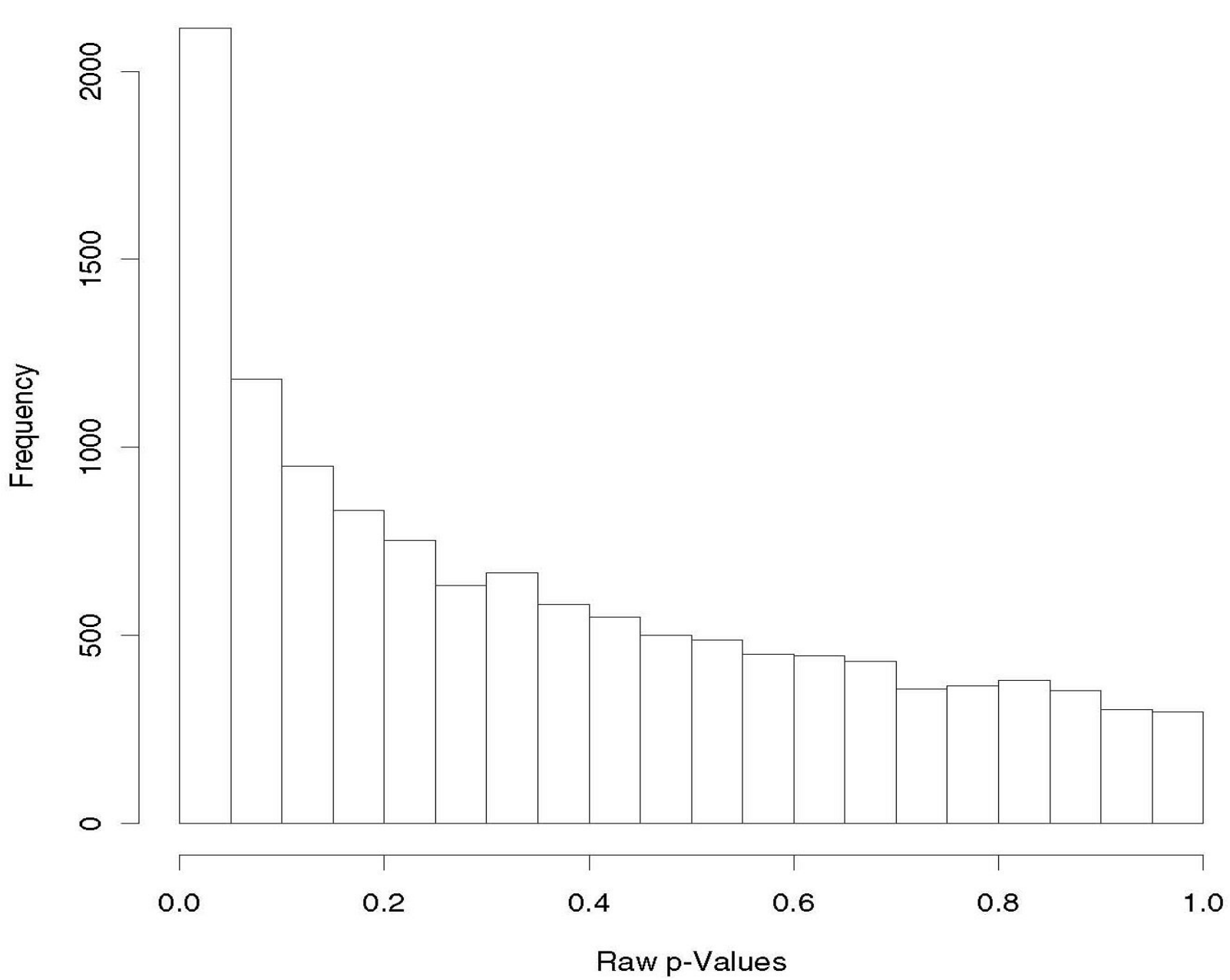
- We estimate all the parameters by normal maximum likelihood, including the transformation, and possibly the background correction.

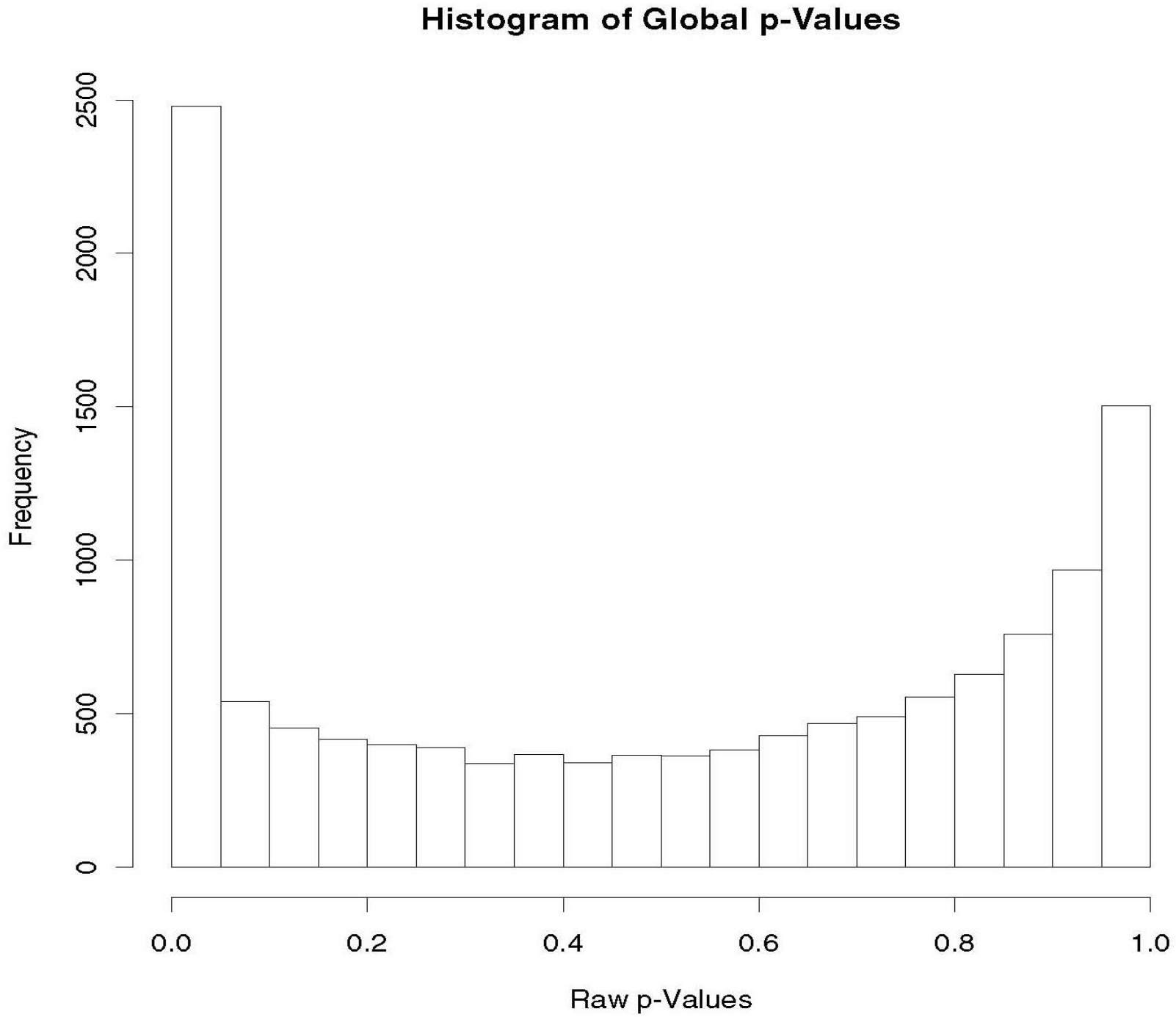
- Some care must be taken in the computations to avoid computer memory problems.

- We can test for differential expression for a given gene by analyzing the transformed, normalized data in a standard one-way ANOVA.

- We can use as a denominator the gene-specific 4df MSE from that ANOVA. This is valid but not powerful.

- We can use the overall 50,493df MSE as a denominator. This is powerful, but risky.


Histogram of Gene-Specific p-Values
Frequency
0
500
1000
1500
2000
0.0
0.2
0.4
0.6
0.8
1.0
Raw p-Values


Histogram of Global p-Values
Frequency
0
500
1000
1500
2000
2500
0.0
0.2
0.4
0.6
0.8
1.0
Raw p-Values

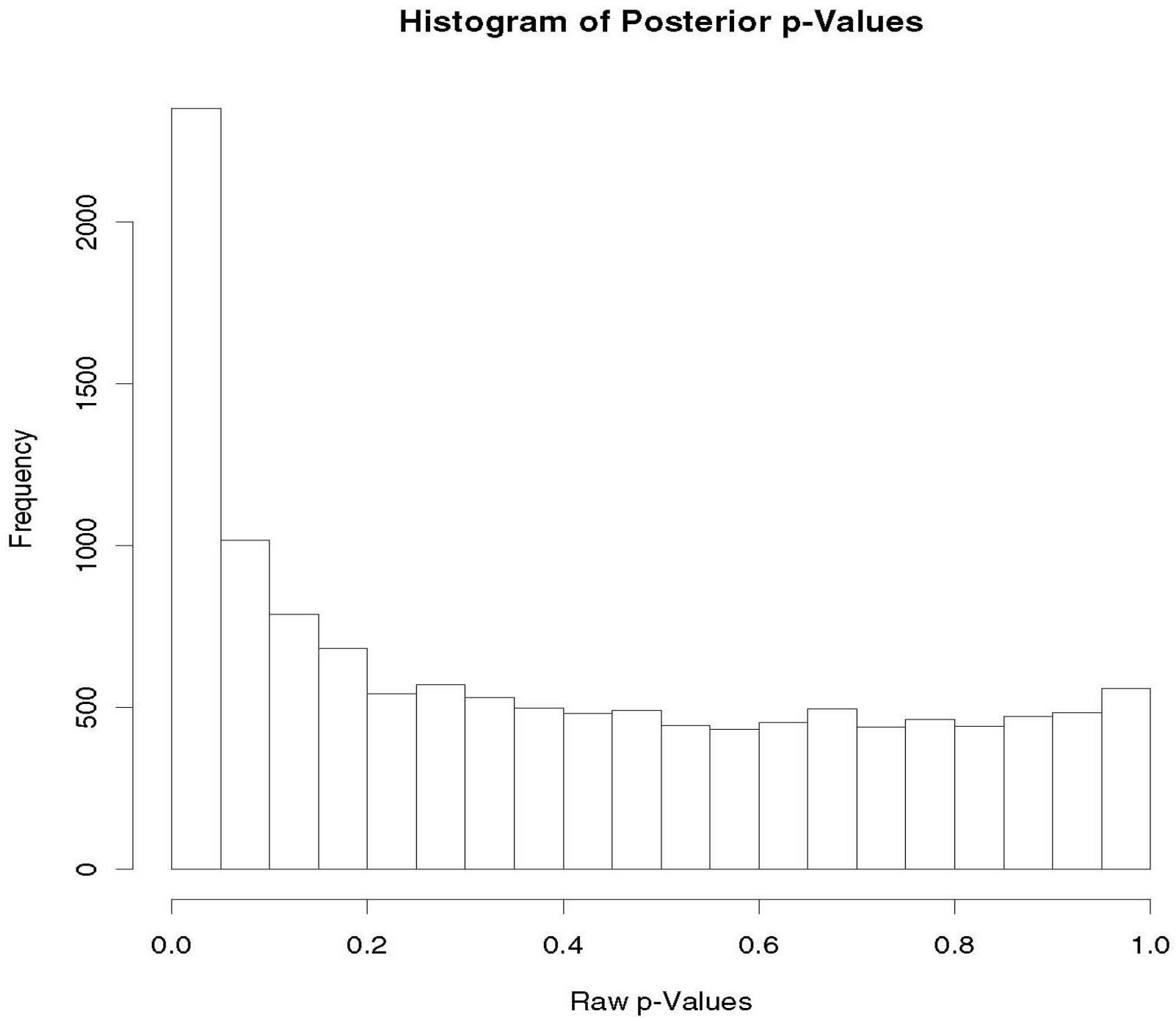
- The F statistics should be large if a significant effect exists, and near 1 if no significant effect exists.

- If very small F statistics occur, it means something is wrong.

- As an alternative, we can use a model that says that the variation in different genes is similar but not identical. The model that assumes the variation to be identical is not tenable in this data set (Wright and Simon 2003; Churchill 2003; Rocke 2003; Smyth 2004).

- Note that we have removed any trend in the variance with the mean. What is left is apparently random.

- The posterior best estimate MSE is a weighted average of the gene-specific MSE (with weight 4/8.6) and the global estimate (with weight 4.6/8.6) and has 8.6 degrees of freedom. The weights depend on the data set.


Histogram of Posterior p-Values
Frequency
0
500
1000
1500
2000
0.0
0.2
0.4
0.6
0.8
1.0
Raw p-Values

"5% Significant" Genes by Several Methods

MSE Source	TWER	FWER	FDR
Gene-Specific	2054	0	0
Global	4215	1029	2693
Posterior	2804	48	866

Metabolomics by NMR Spectroscopy

- Proton NMR spectroscopy produces a spectrum in which peaks correspond to parts of molecules.

- This can be used for single compounds to determine the structure.

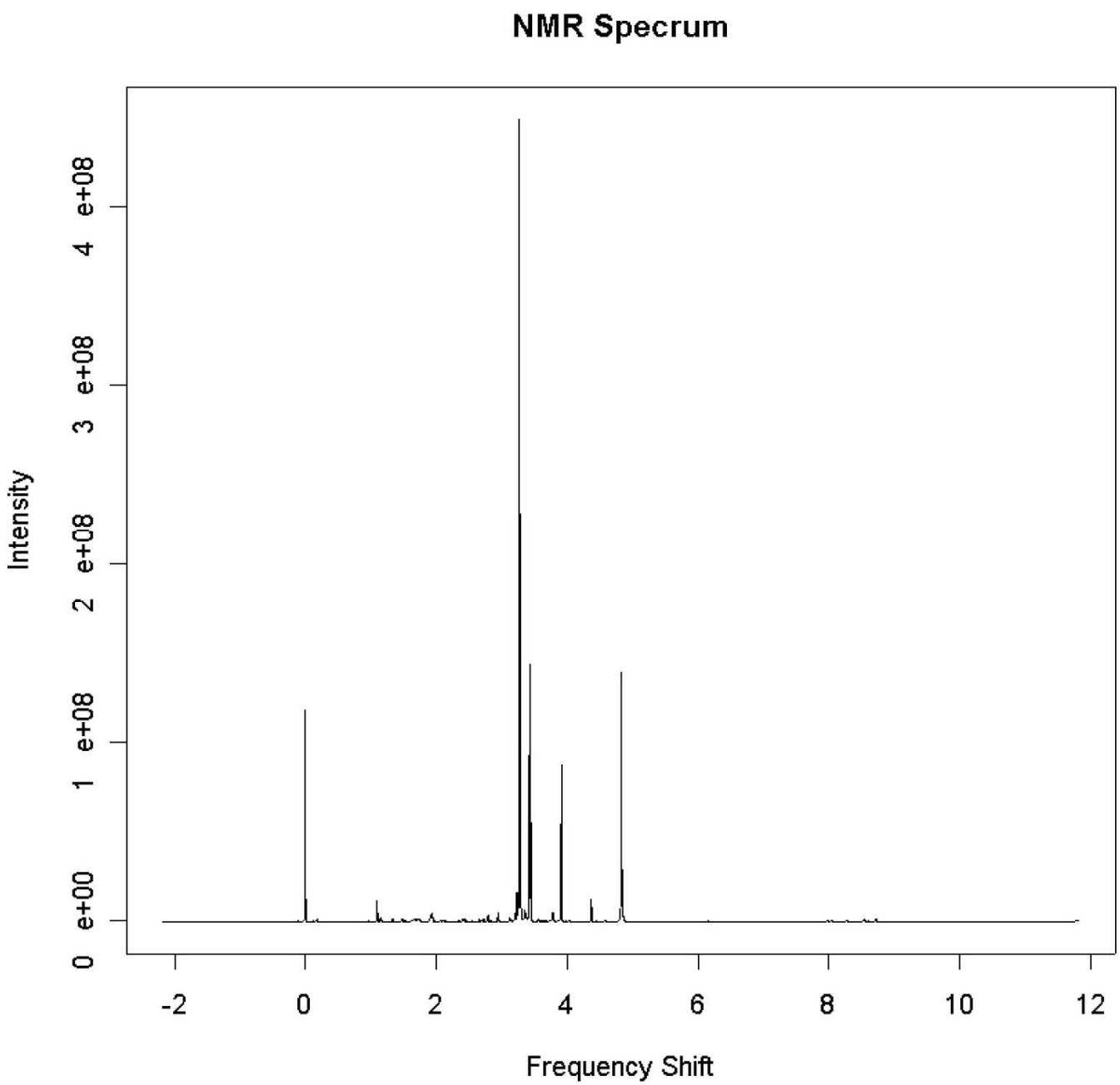
- Compounds often have specific signatures, so this can be used for compound identification (particularly by 2D NMR).

- For metabolomics work, one can use patterns in the spectra for discrimination/classification, and to identify regions of the spectrum which carry the discrimination information.

- Spectra need to be baseline corrected and peaks need to be aligned

- The peaks are of widely varying magnitudes, and some of the data are negative.

- The glog is a plausible transformation to help in the analysis of these data.


NMR Specrum
Intensity
0 e+00
1 e+08
2 e+08
3 e+08
4 e+08
-2
0
2
4
6
8
10
12
Frequency Shift

Variance Behavior of NMR Spectra

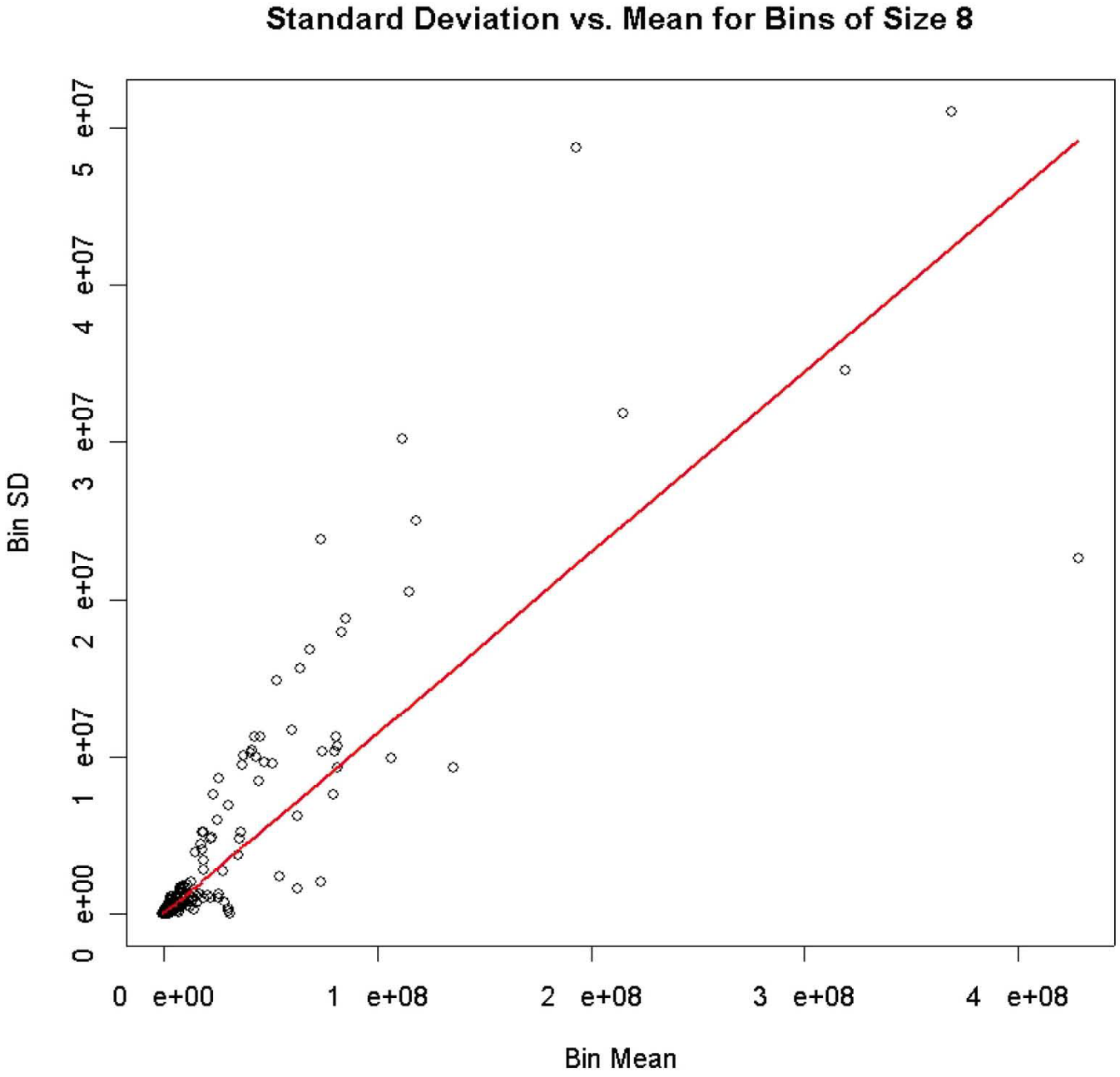
- We show an example spectrum of 65,536 points.

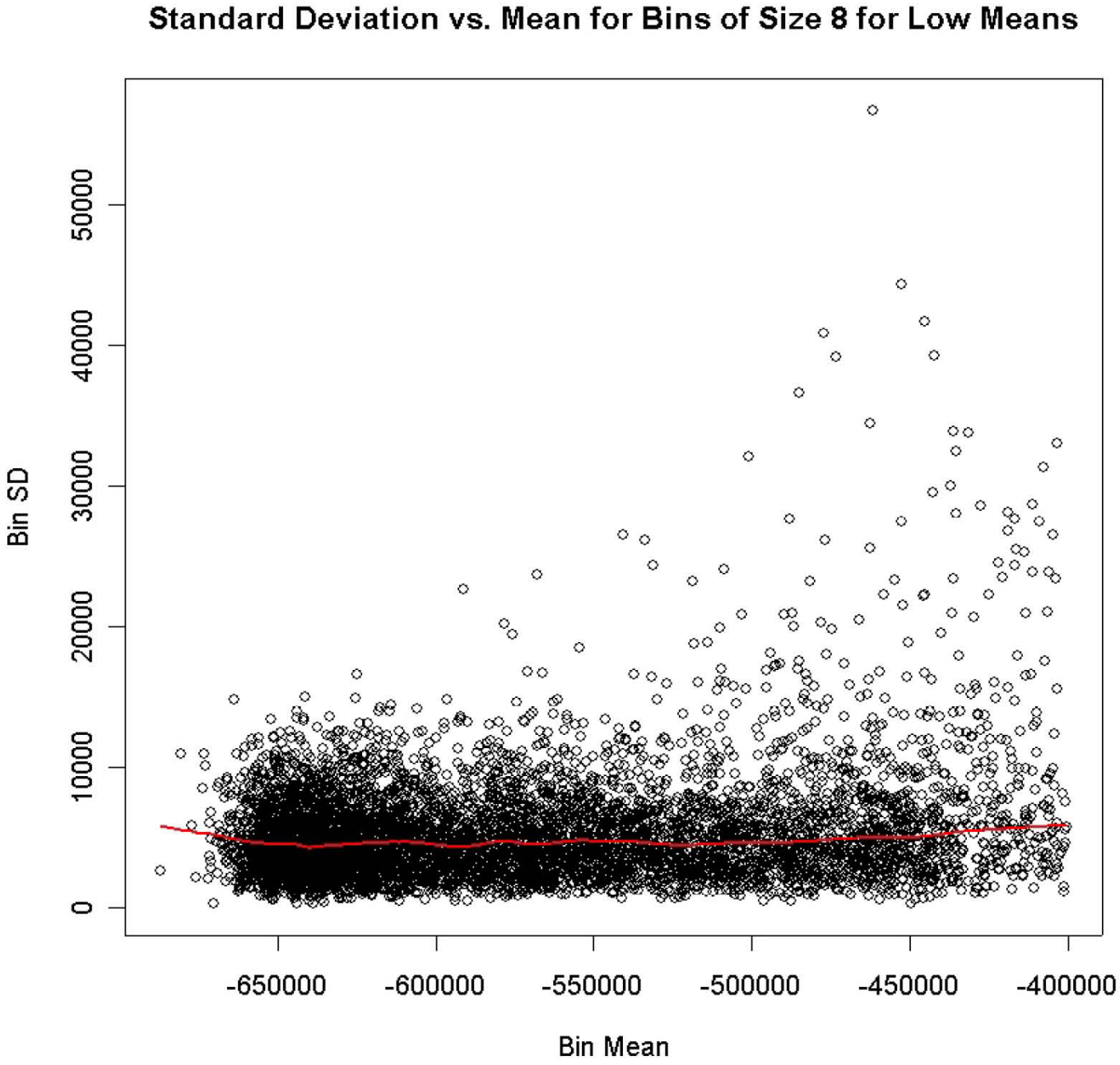
- We divide this into 8,192 bins of 8 points each and compute the mean and standard deviation within each bin.

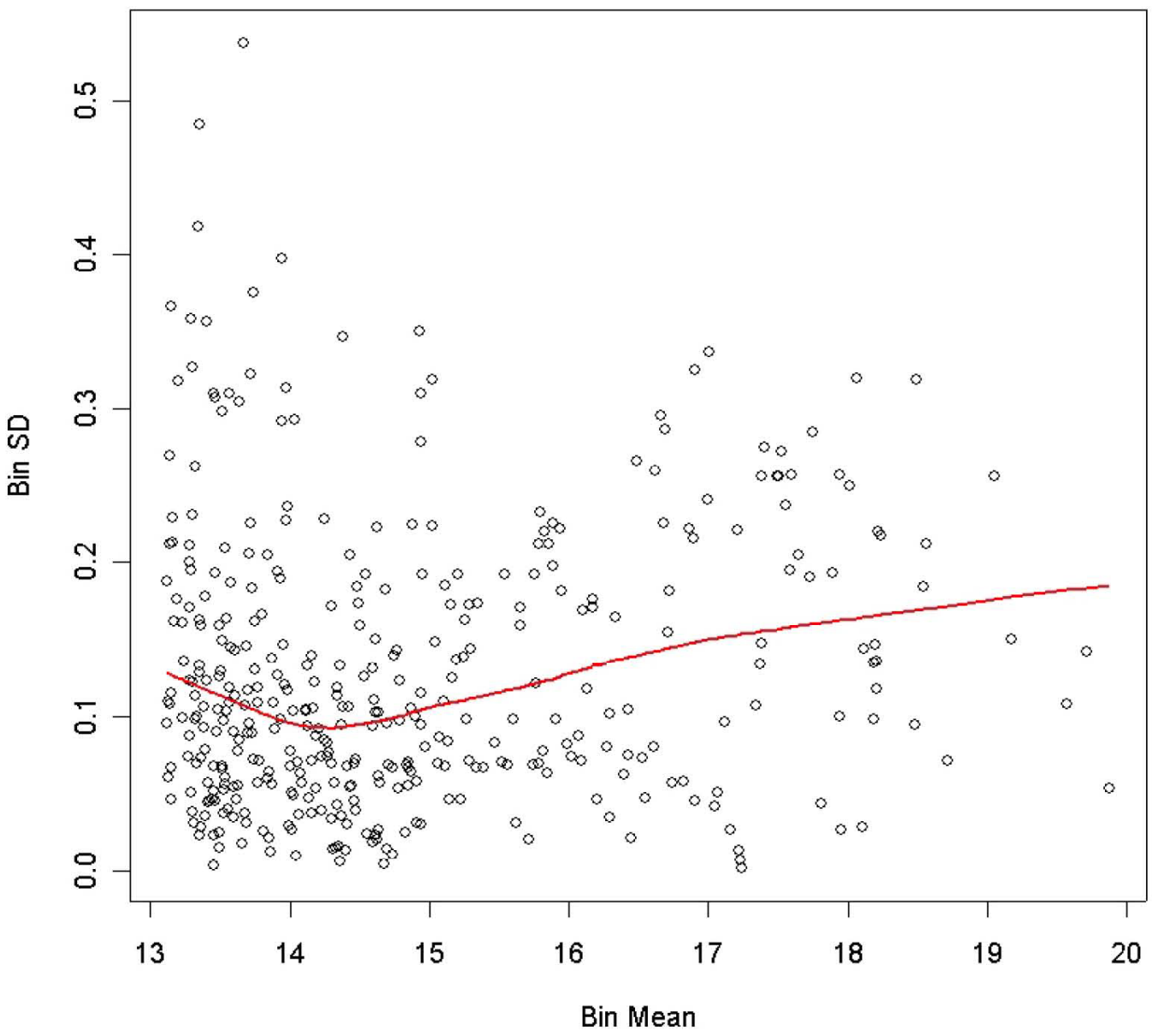
A model for the spectrum is

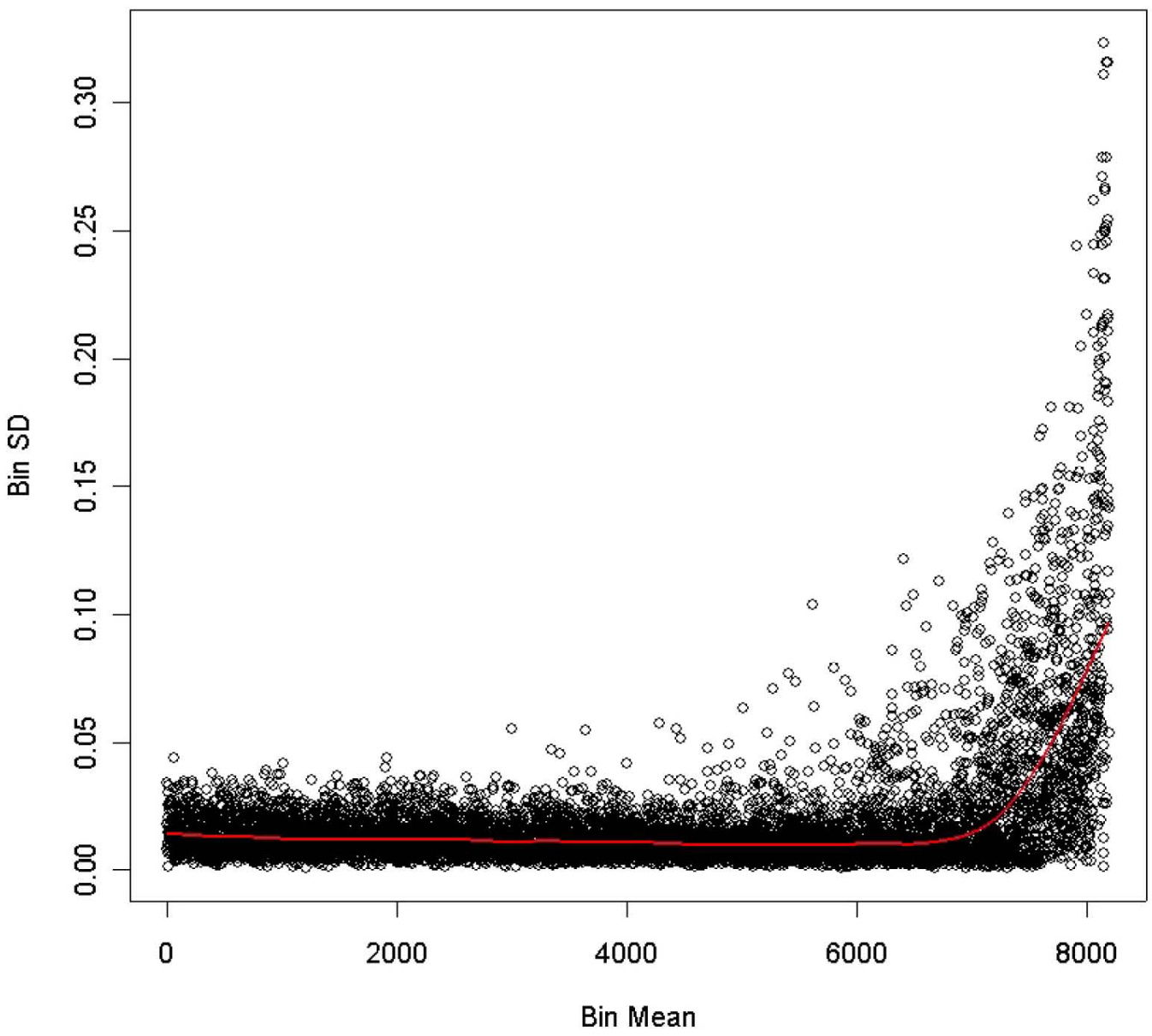
$$y_i = b_i + \mu_i e^{\eta} + \epsilon_i$$

Where b_i is the baseline, not presumed to be flat, μ_i is the true signal, and ϵ_i and η_i are measurement errors, not necessarily independent across nearby points.


Standard Deviation vs. Mean for Bins of Size 8
Bin SD
0 e+00
1 e+07
2 e+07
3 e+07
4 e+07
5 e+07
0 e+00
1 e+08
2 e+08
3 e+08
4 e+08
Bin Mean


Standard Deviation vs. Mean for Bins of Size 8 for Low Means
Bin SD
0
10000
20000
30000
40000
50000
-650000
-600000
-550000
-500000
-450000
-400000
Bin Mean

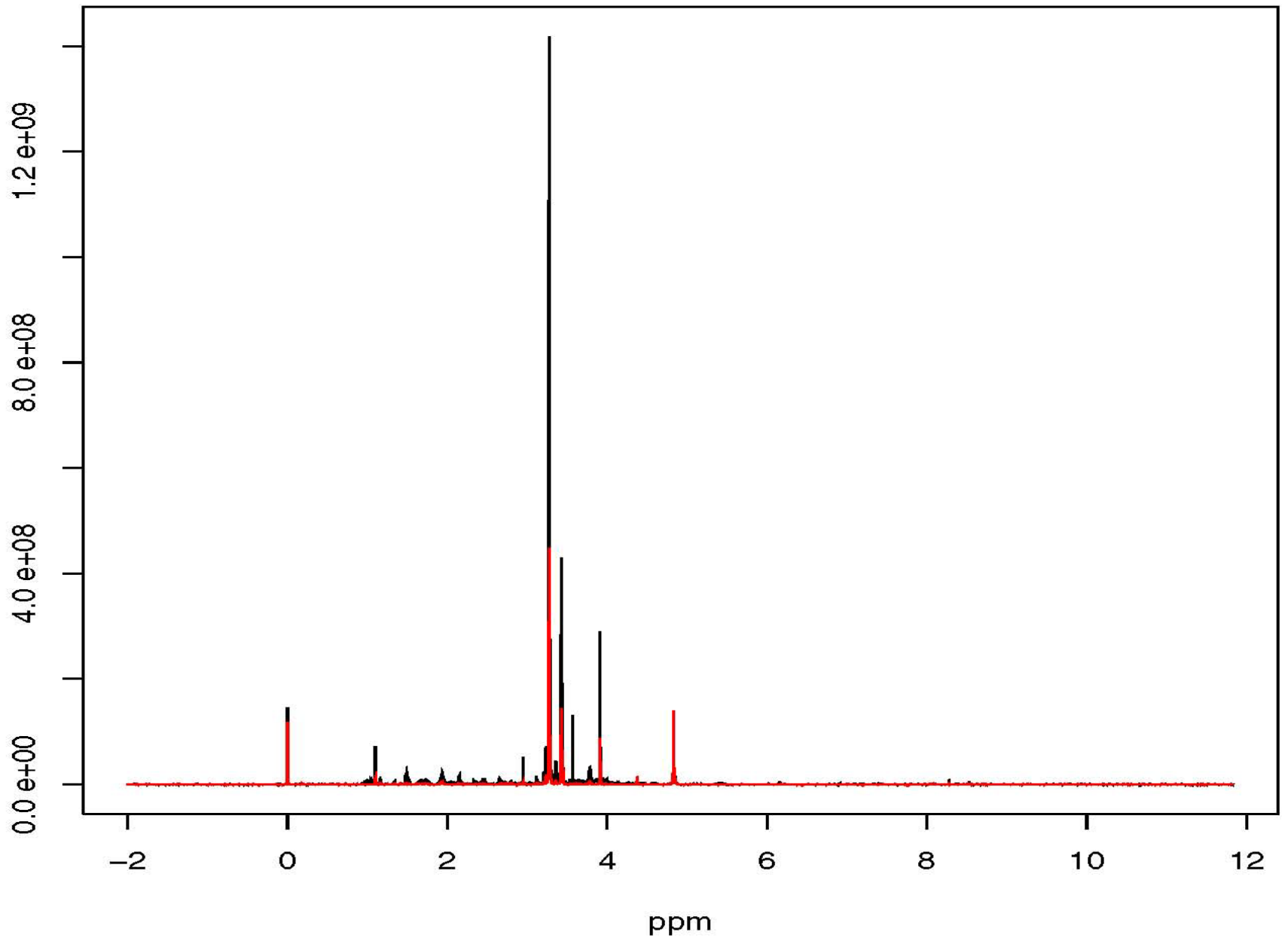

SD vs. Mean of Logs for Bins of Size 8 for High Means
Bin SD
0.0
0.1
0.2
0.3
0.4
0.5
13
14
15
16
17
18
19
20
Bin Mean

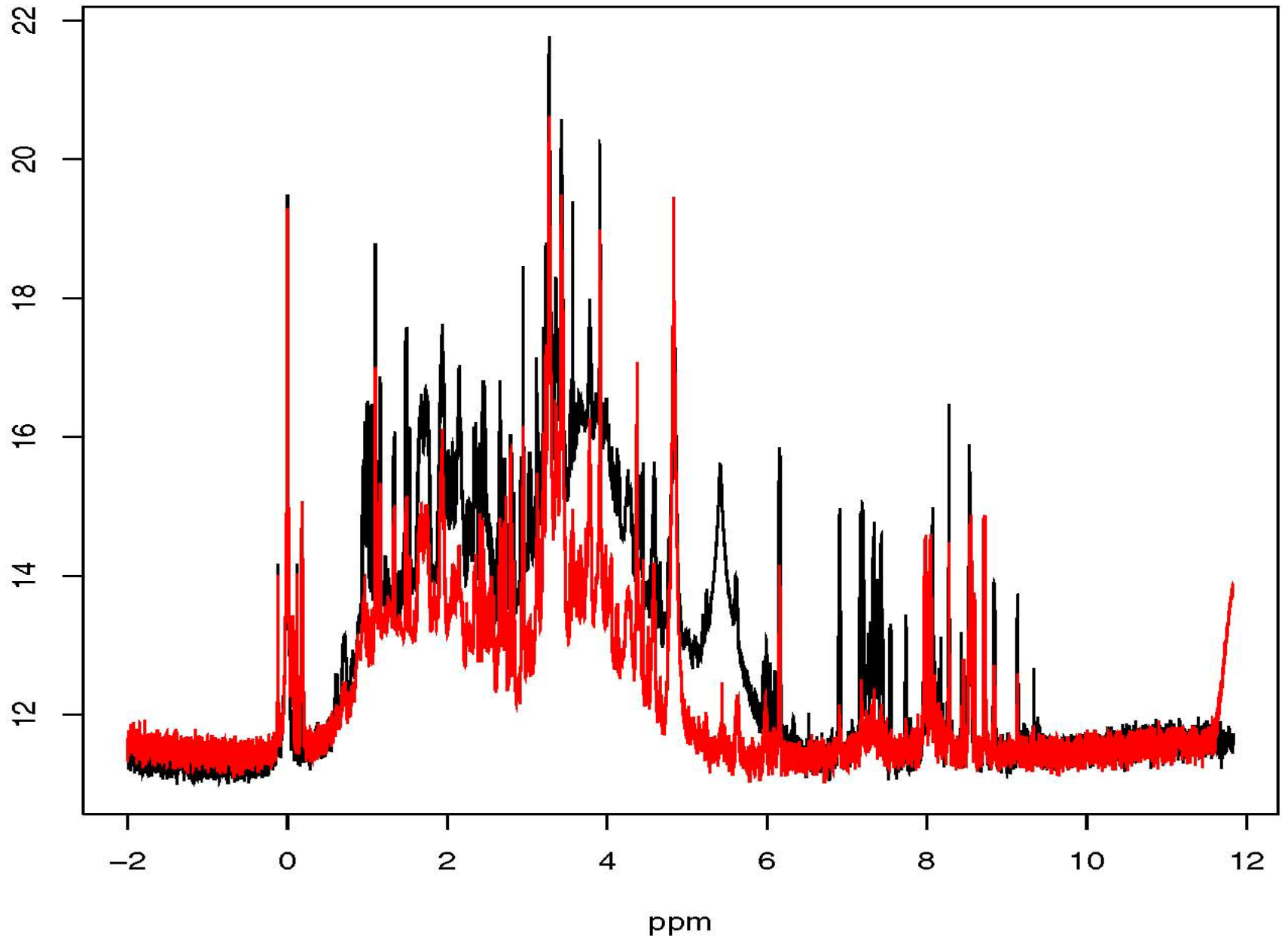

SD vs. Mean of Glogs for Bins of Size 8
Bin SD
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2000
4000
6000
8000
Bin Mean

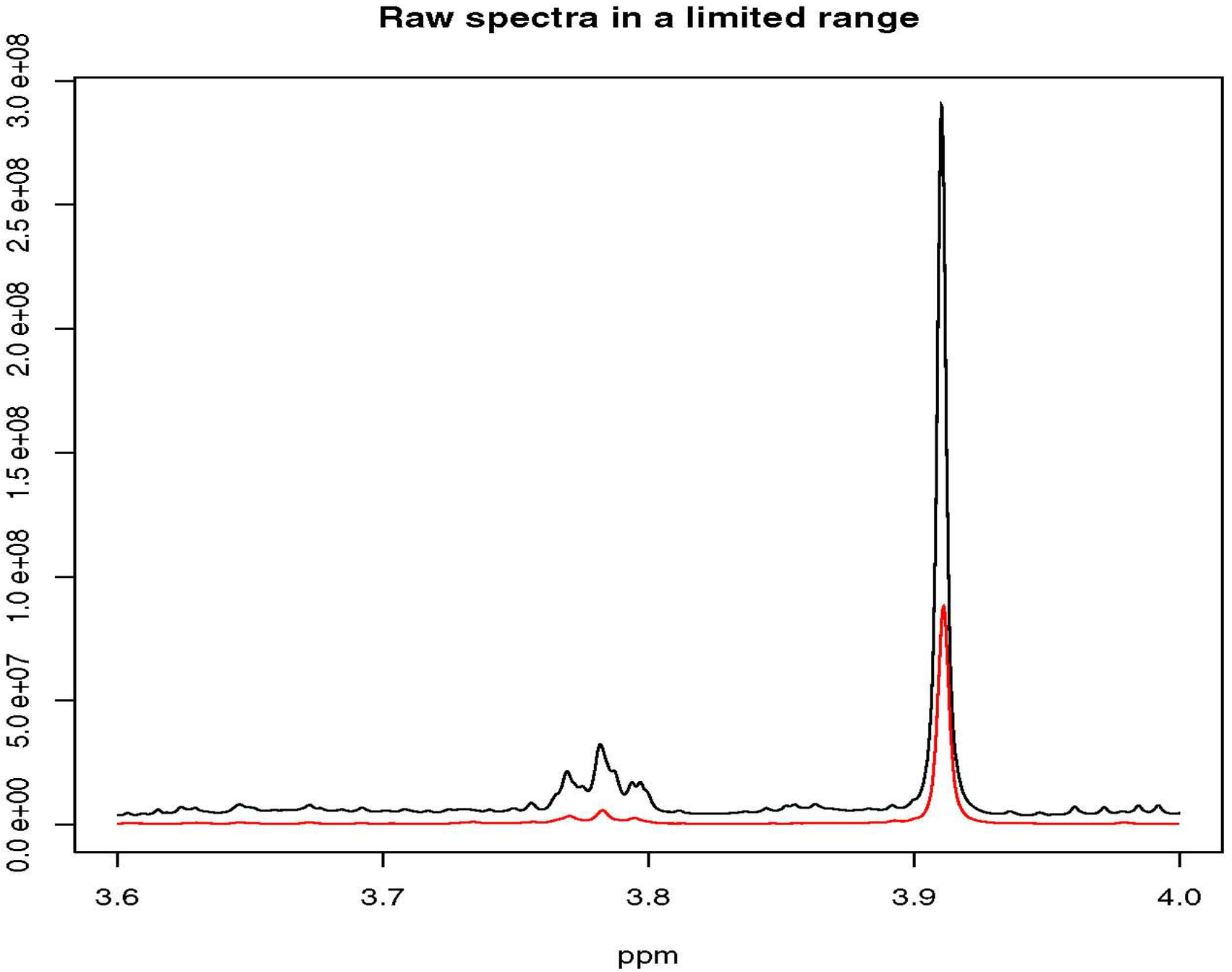
Baseline Estimation

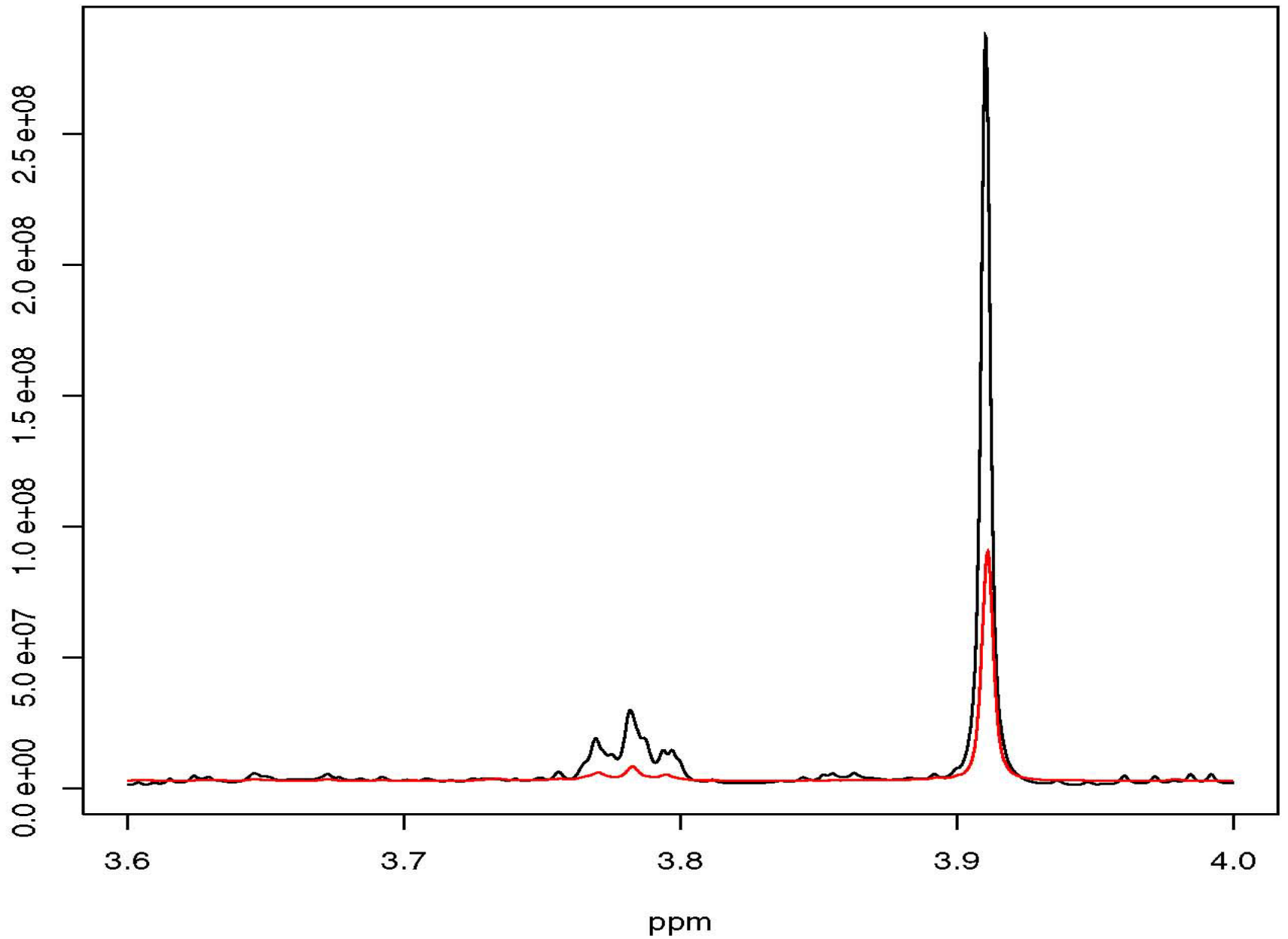
- The baseline in NMR is arbitrary and needs to be removed before analysis, just as in mass spec.

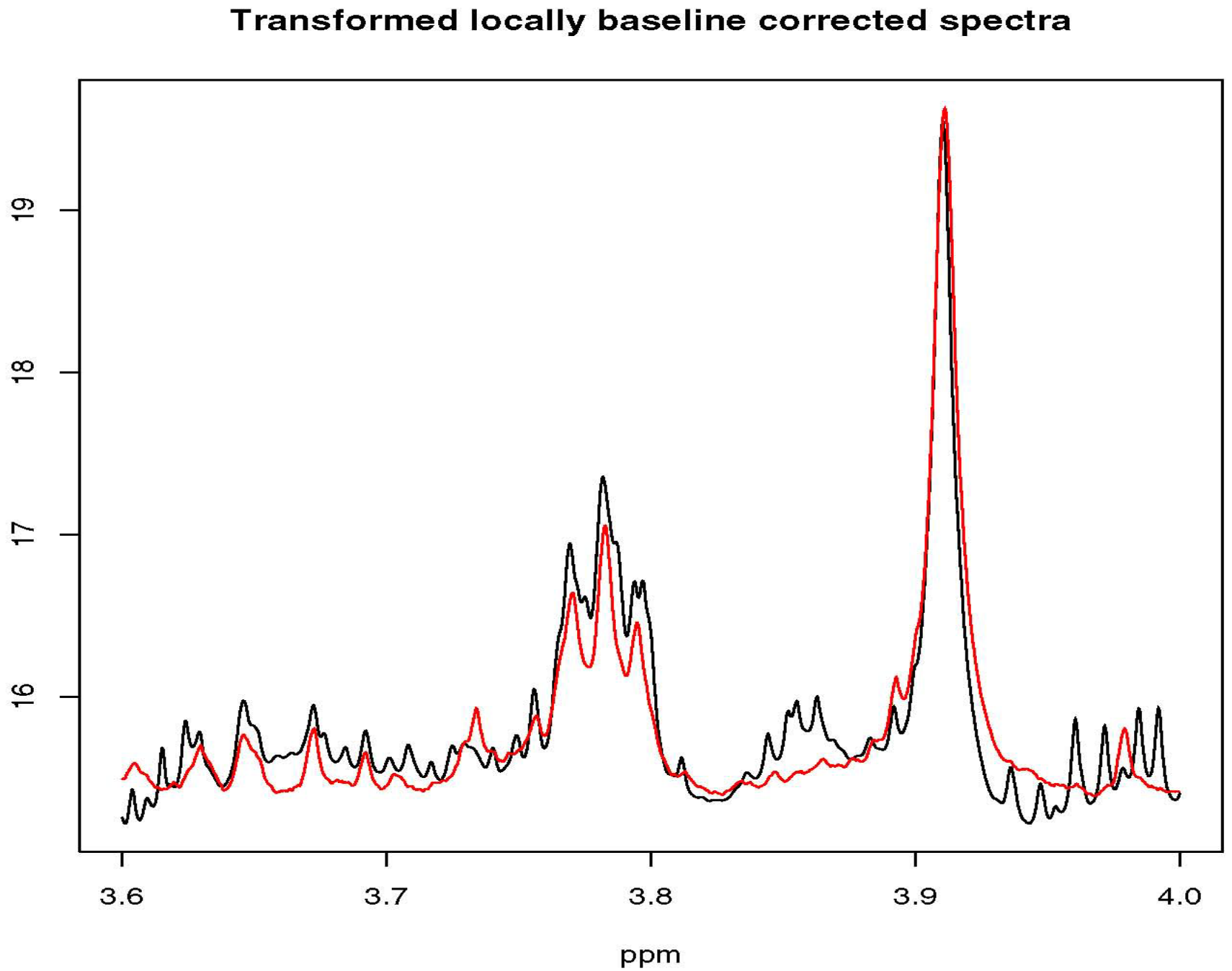
- The baseline is less well behaved than for mass spec

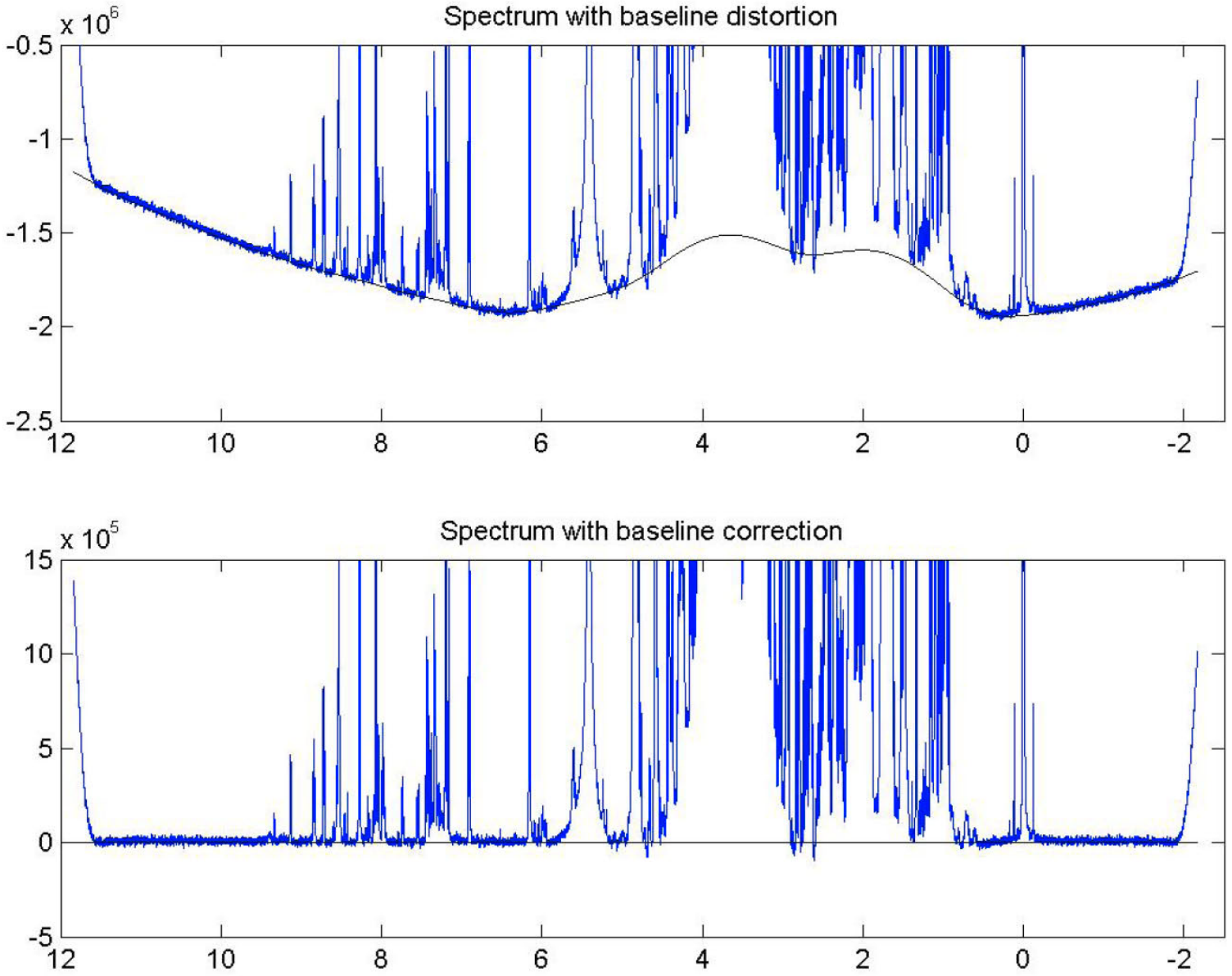

Raw baseline–corrected spectra
0.0 e+00
4.0 e+08
8.0 e+08
1.2 e+09
−2
0
2
4
6
8
10
12
ppm


One glog transform of whole spectrum
22
20
18
16
14
12
−2
0
2
4
6
8
10
12
ppm


Raw spectra in a limited range
0.0 e+00
5.0 e+07
1.0 e+08
1.5 e+08
2.0 e+08
2.5 e+08
3.0 e+08
3.6
3.7
3.8
3.9
4.0
ppm


Raw locally baseline corrected spectra
0.0 e+00
5.0 e+07
1.0 e+08
1.5 e+08
2.0 e+08
2.5 e+08
3.6
3.7
3.8
3.9
4.0
ppm


Transformed locally baseline corrected spectra
19
18
17
16
3.6
3.7
3.8
3.9
4.0
ppm


Spectrum with baseline distortion
x 10^6
-0.5
-1
-1.5
-2
-2.5
12
10
8
6
4
2
0
-2
Spectrum with baseline correction
x 10^5
15
10
5
0
-5
12
10
8
6
4
2
0
-2

Conclusion

- Gene expression microarray and other omics data present many interesting challenges in design and analysis of experiments.

- Statistical lessons from years of experience with laboratory, clinical, and field data apply with some modification to expression data, proteomics data, and metabolomics data.

- A properly chosen transformation can stabilize the variance and improve the statistical properties of analyses.

- Slide normalization and analysis of two-color arrays is made easier by this transformation.

- Other statistical calculations such as the analysis of variance that assume constant variance are also improved.

- After removal of systematic dependence of the variance on the mean, the remaining sporadic variation in the variance can be accounted for by a simple method.

- These methods can be applied to other types of data such as proteomics by mass spec and NMR spectroscopy metabolomics. The variables measured are a large number of peak heights or areas, or a large number of binned spectroscopic values

- "If your experiment needs statistics, you ought to have done a better experiment," (Ernest Lord Rutherford).

- Lord Rutherford to the contrary notwithstanding, if you need statistics, you may indeed be doing the right experiment.

- Papers are available at `www.cipic.ucdavis.edu/~dmrocke` or by mail and e-mail.

Acknowledgements

IDAV Faculty
Sue Geller (Texas A&M)
David Woodruff
Research Staff and Postdocs
Jian Dai
Lexin Li
Parul Purohit
John Tillinghast
Students and Former Students
Shagufta Aslam (UCD)
Blythe Durbin (UC Berkeley)
Wen-Ying Feng
Johanna Hardin (Pomona College)
Dan Li
Shuang Liu
Danh Nguyen (UC Davis)
Machelle Wilson (Univ. Georgia)
Yuanxin Xi
Jingjing Ye
Jufen Zhou
Lei Zhou
UC Davis Collaborators
Matt Bartosiewicz
Alan Buckpitt
Satya Dandekar
Jeff De Ropp
Bruce German
Dorothy Gietzen
Zelanna Goldberg
Jeff Gregg
Paul Hagerman
Bruce Hammock
Rivka Isseroff
Carlito Lebrilla
Kent Pinkerton
Bob Rice
Pam Ronald
Ralph deVere White
Outside Collaborators
Doug Hawkins (University of Minnesota)
Wolfgang Huber (German Cancer Research Center, Heidelberg)
Larry Kauvar (Trellis Bioscience)
Robert Nadon (McGill University)
Dan Solomon (Scripps Institute)
Mark Viant (University of Birmingham)
Martin Vingron (MPI/Molecular Genetics, Berlin)
Steve Watkins (Lipomics)
Funding
NSF
NIEHS
NIH
US EPA
UC Davis MIND Institute