# Some Principles for the Design and Analysis of Experiments using High-Throughput Assay Methods

BIM 283 Advanced Experimental Design for Biomedical Engineers

January 21, 2025

# The -Omics Revolution

Gene expression microarrays, RNA-Seq, proteomics by Luminex and mass spectrometry, and metabolomics by mass spectrometry and NMR spectroscopy presents enormous opporunities for fundamental biological research and for applications in medicine and biology.

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

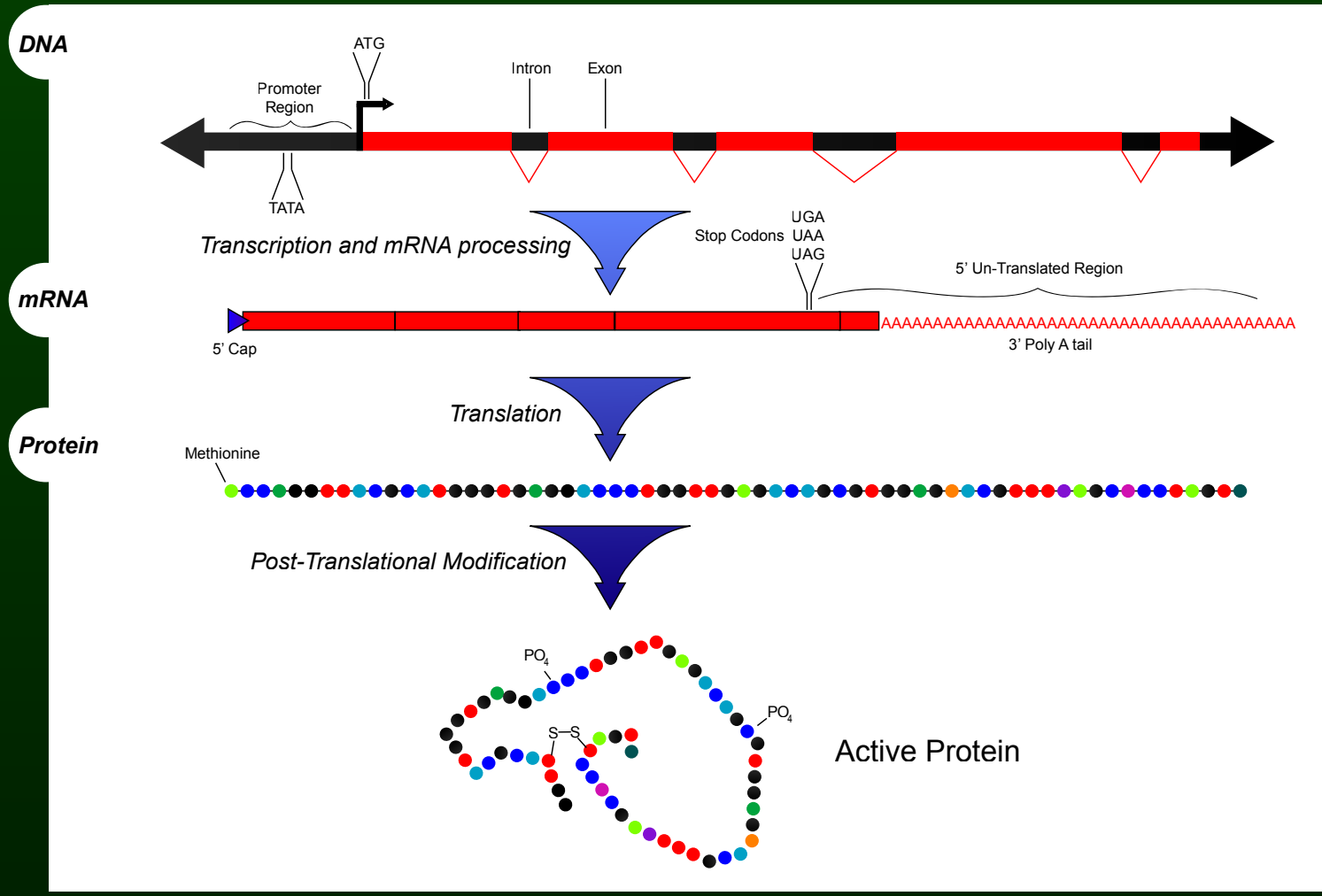
**Genome** Complement of all genes, or of all components of genetic material in the cell (mostly static).

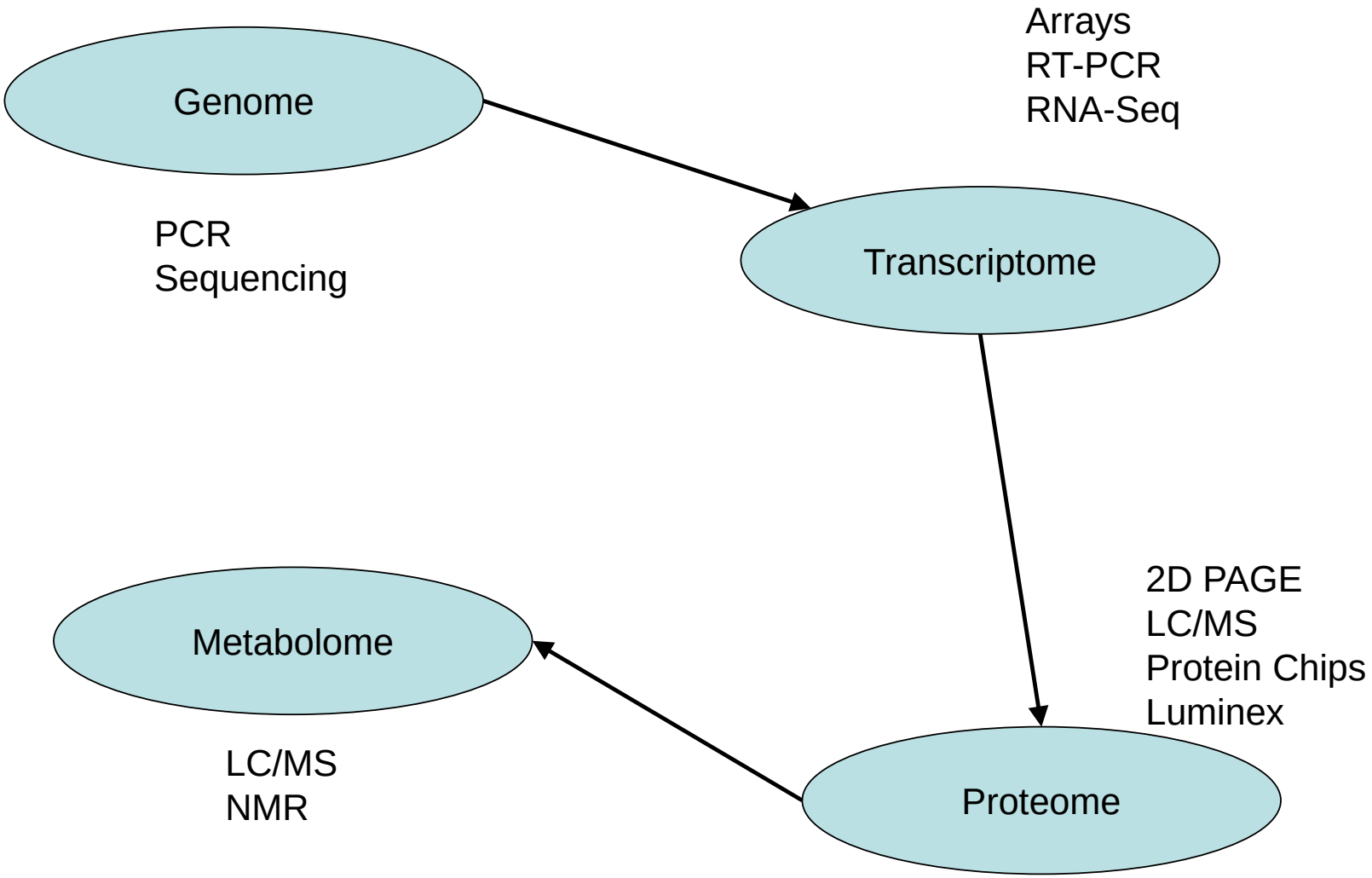
**Transcriptome** Complement of all mRNA transcripts produced by a cell (dynamic).

**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).

# Central Dogma of Molecular Biology : Eukaryotic Model




Genome
Arrays
RT-PCR
RNA-Seq
PCR
Sequencing
Transcriptome
2D PAGE
LC/MS
Protein Chips
Luminex
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 50,000.

- We need more biological replicates than you can afford!

- 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 can be an 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. But this too can be variable and important, for example in cancer studies.

- 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 differences. This will clearly miss some large real differences.

- 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.

- We can say that the probability that each is a real difference is 95%.

# 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.

# 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 High Throughput Data

Some well known properties of measurement error in assays with many measurements include the following:

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

- For low level measurements, the CV is much higher.

- Analysis 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 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 as in gene expression, proteomics, and metabolomics.

- The fold change is the ratio of two responses.

- 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 $\mu$, 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 because measurement error always exists.

- So a reasonable model for the variance of such assay 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 $\alpha$ 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

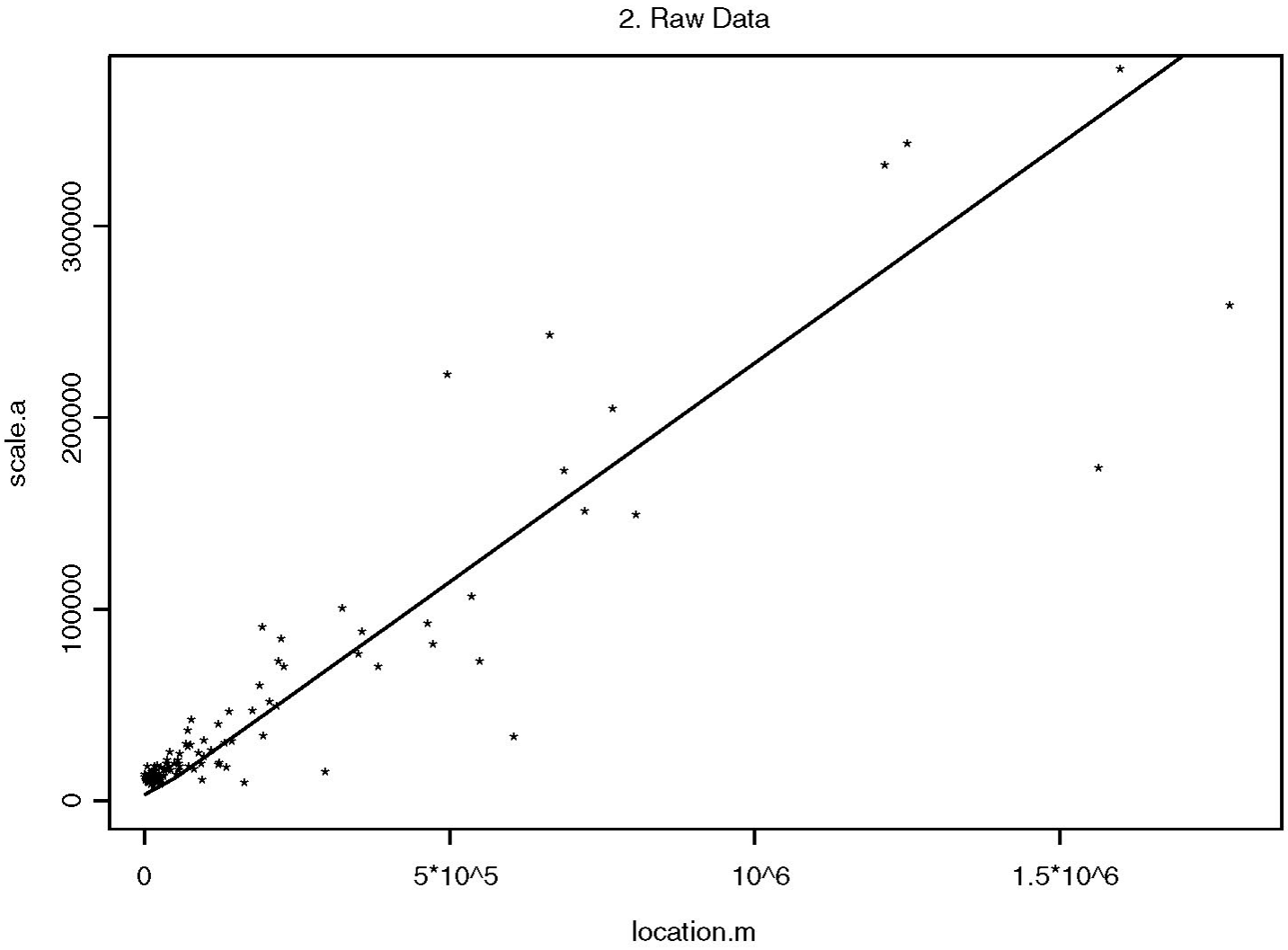
$$\begin{aligned} E(z) &= \mu \\ V(z) &= a^2 + b^2\mu^2 \end{aligned}$$

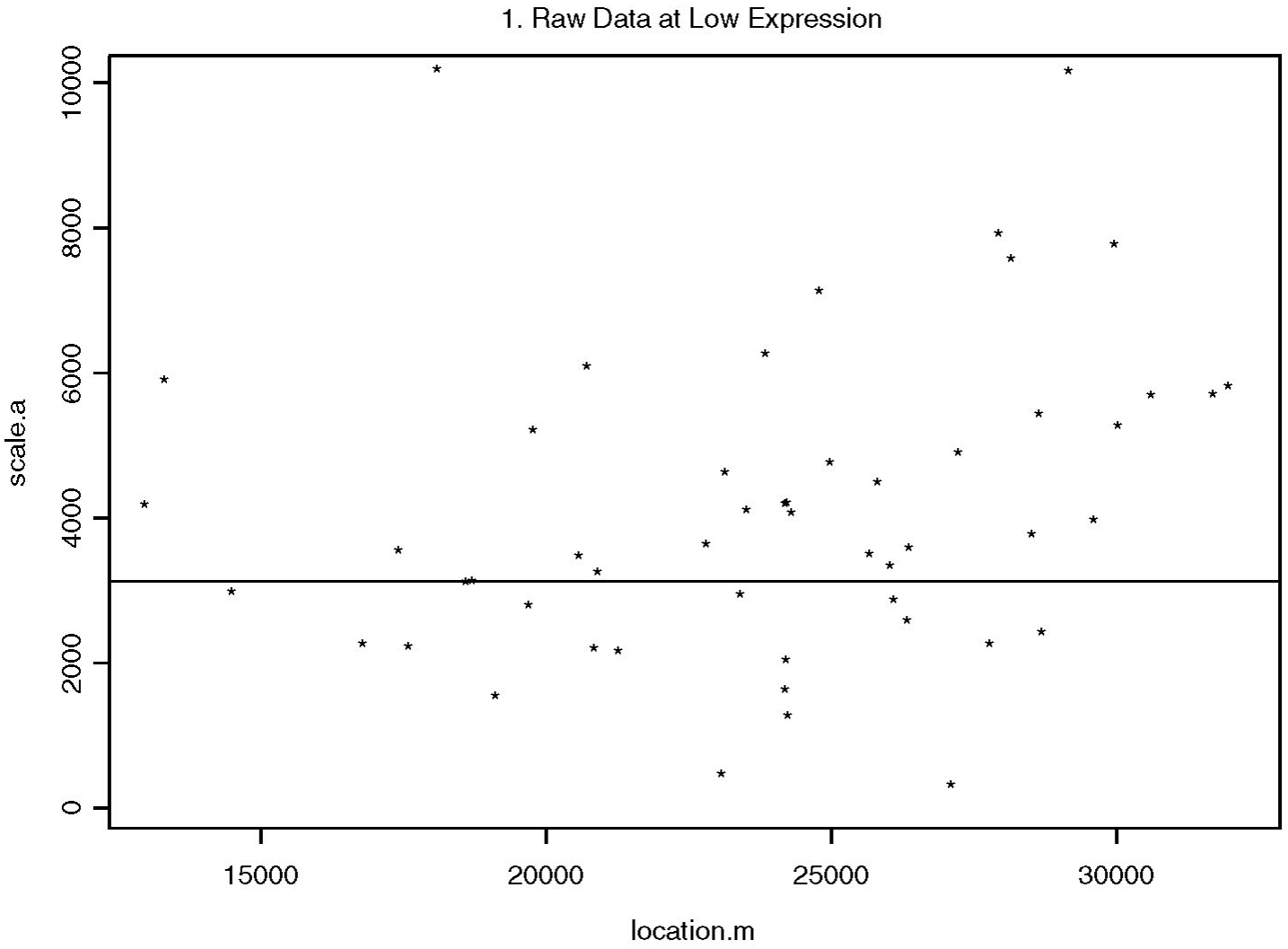
It can be shown that

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

# 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
location.m
0
2000
4000
6000
8000
10000
15000
20000
25000
30000

# 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 measured on slides. Counted data as in RNA-Seq requires some alteration to be described later.

- 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)

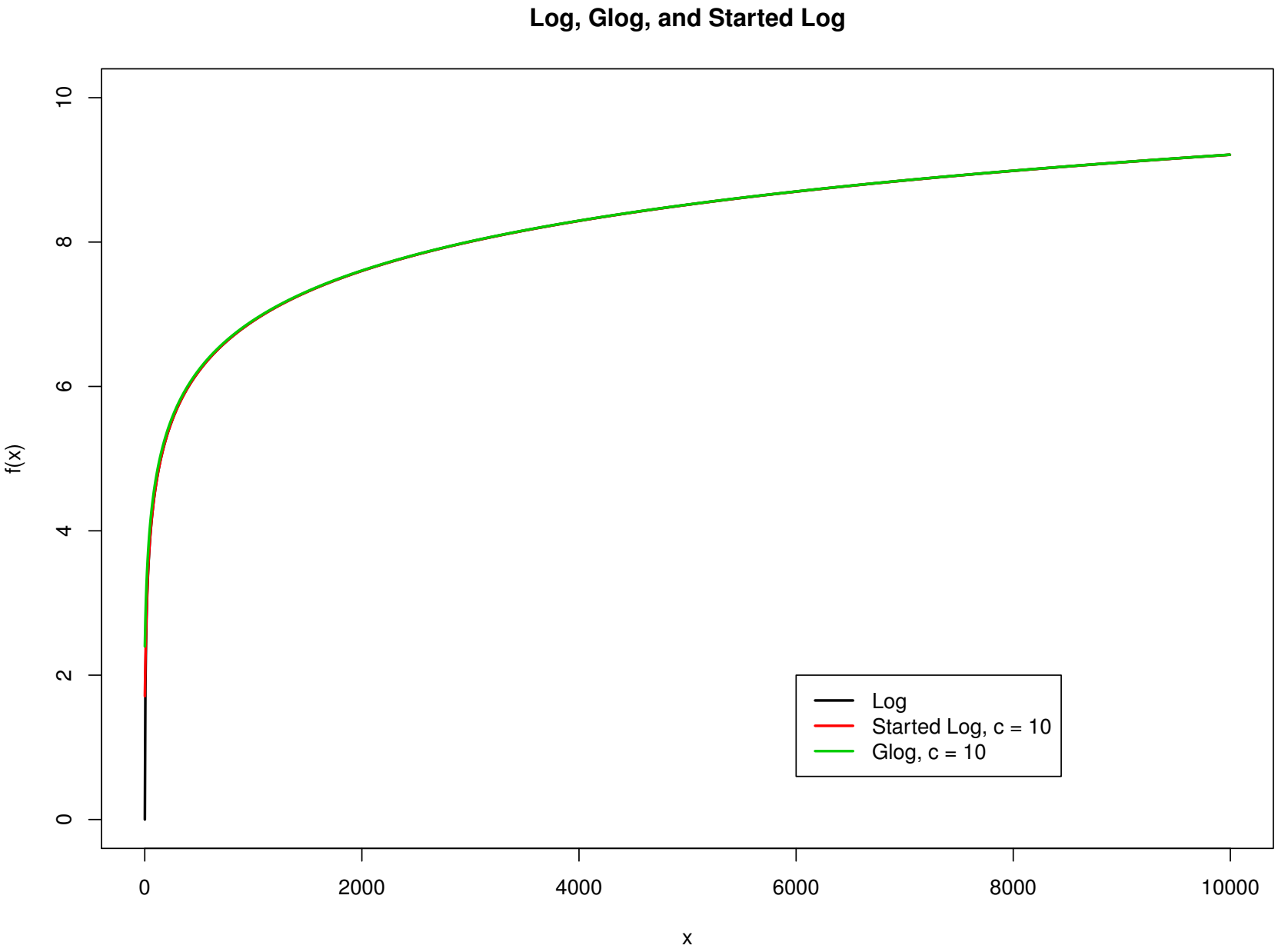
$$f(z) = \mathrm{ln}(z + \sqrt{z^2 + a^2/b^2})$$

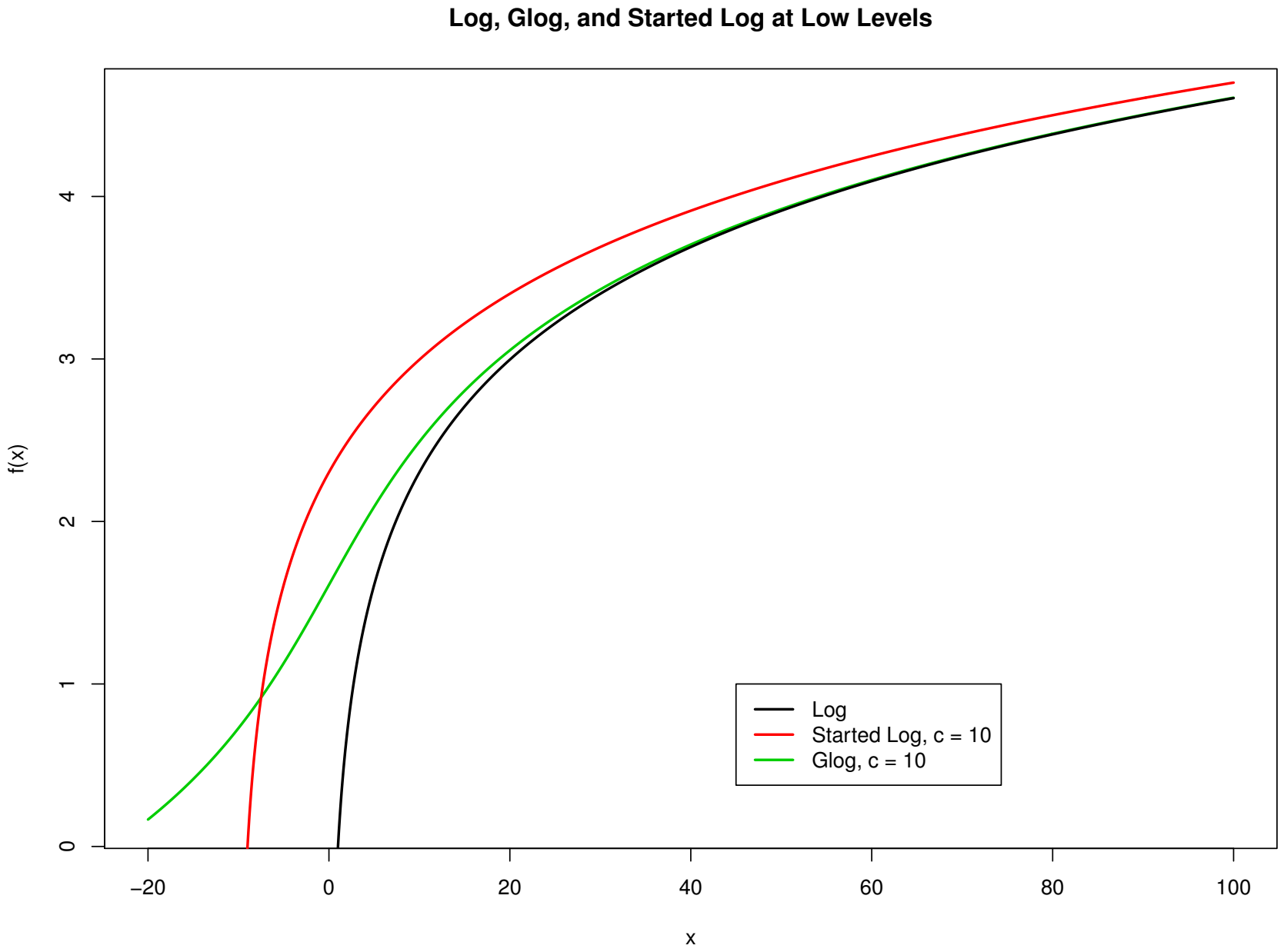
- $f(z) \sim \mathrm{ln}(z)$ for large $z$.
- $f(z)$ is approximately linear for $z = 0$.
- $f(z)$ is monotonic (does not change the order of size of data).

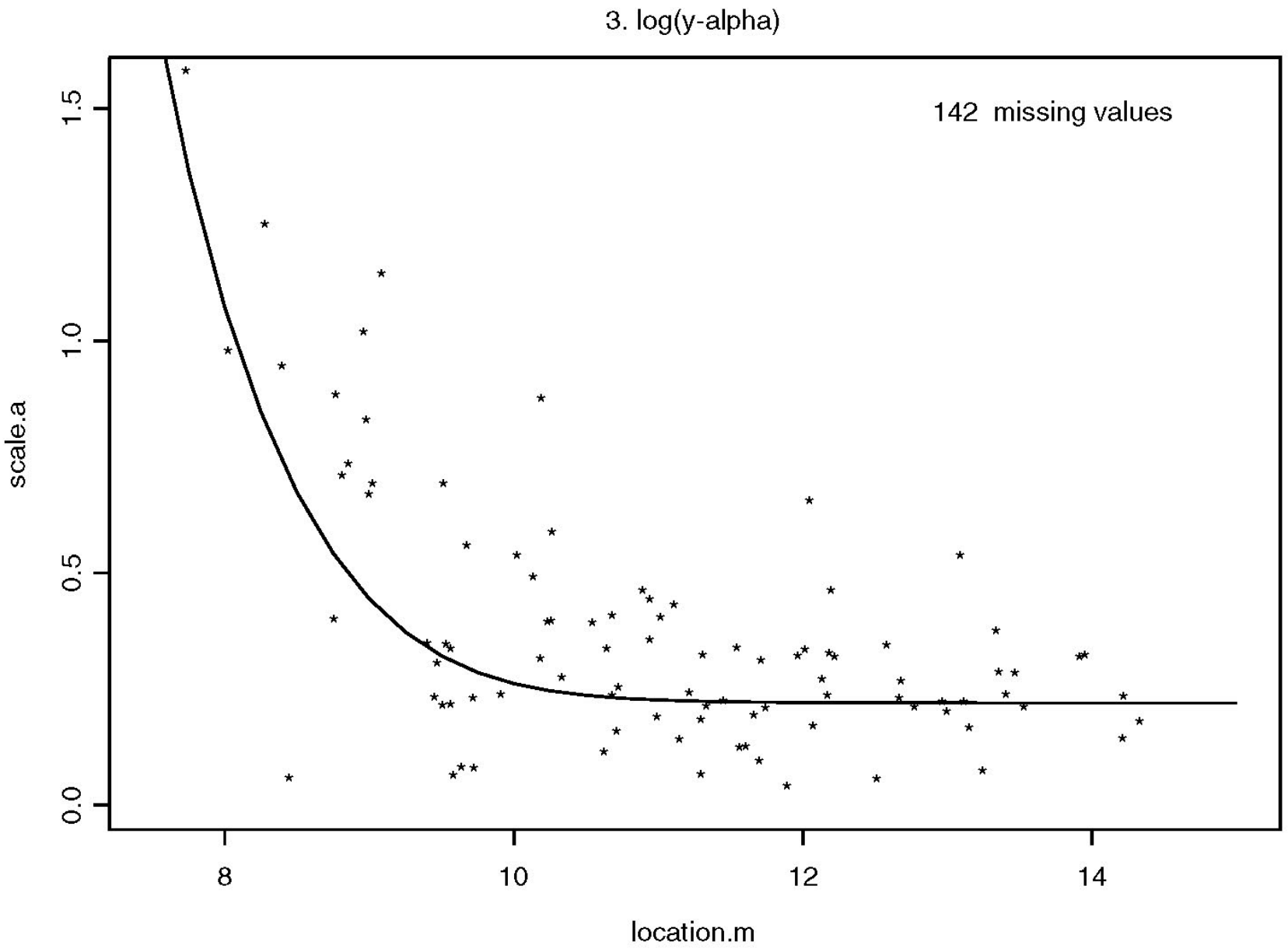
Another transformation family that has similar properties is the started log, defined by

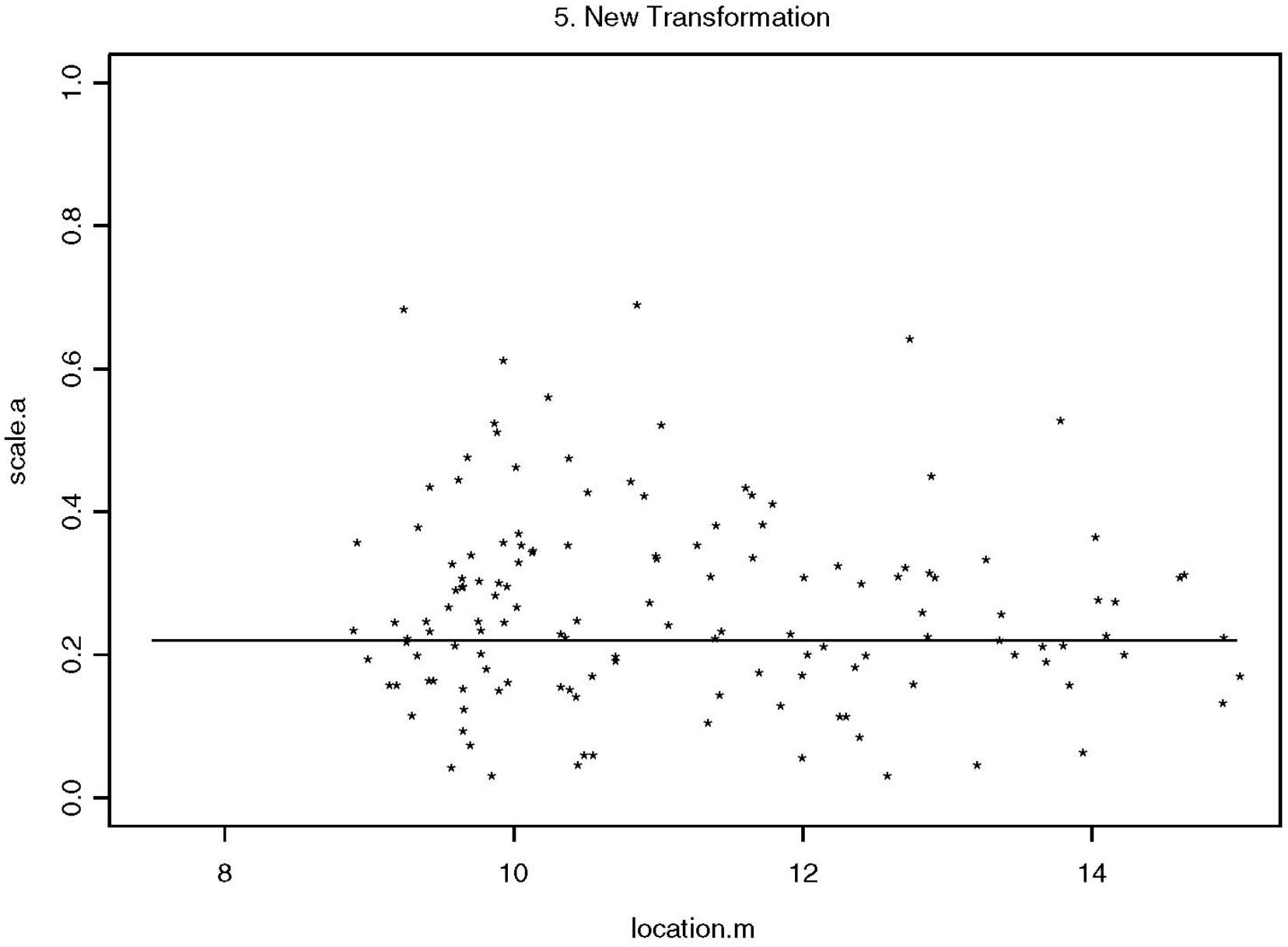
$$g(z) = \ln(z + c)$$

This is often easier to handle, though as with the glog, the parameters must be chosen wisely.


Log, Glog, and Started Log
f(x)
x
0
2
4
6
8
10
0
2000
4000
6000
8000
10000
Log
Started Log, c = 10
Glog, c = 10


Log, Glog, and Started Log at Low Levels
f(x)
0
1
2
3
4
−20
0
20
40
60
80
100
x
Log
Started Log, c = 10
Glog, c = 10


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
1.0
0.8
0.6
0.4
0.2
0.0
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.

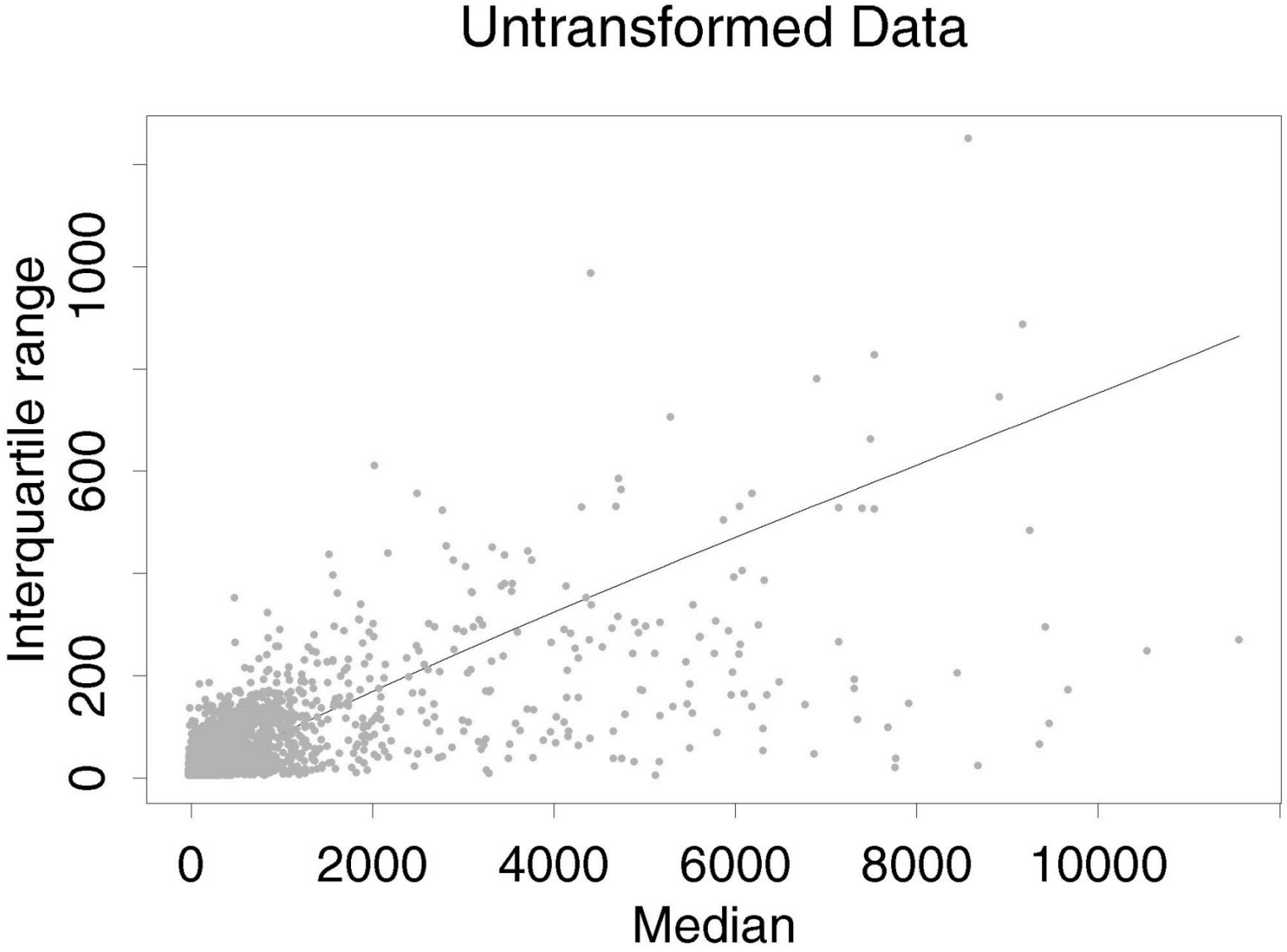
## Normalization and Transformation of Arrays

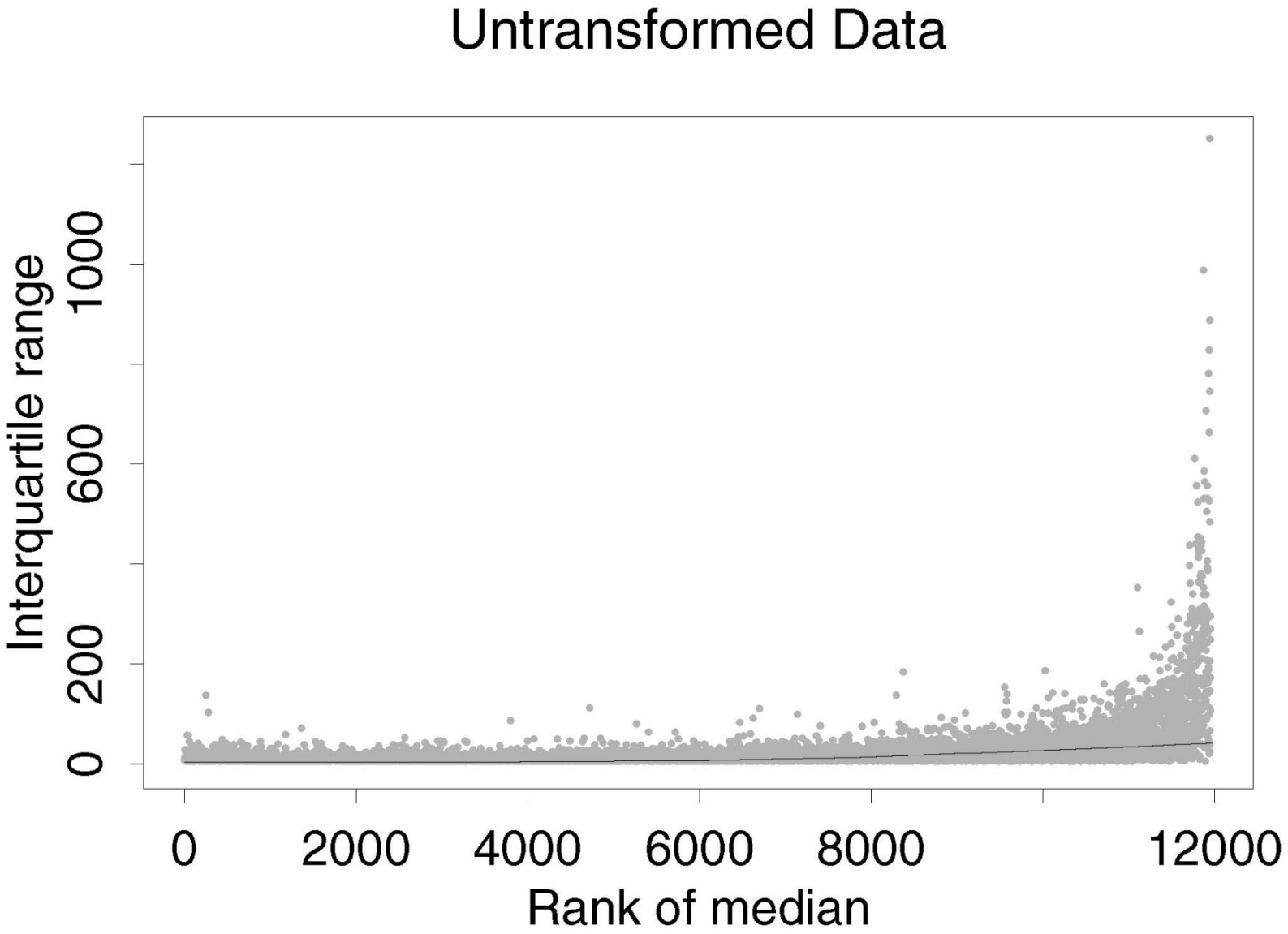
Given a set of replicate chips from the same biological sample, we can simultaneously determine the transformation parameter and the normalization.

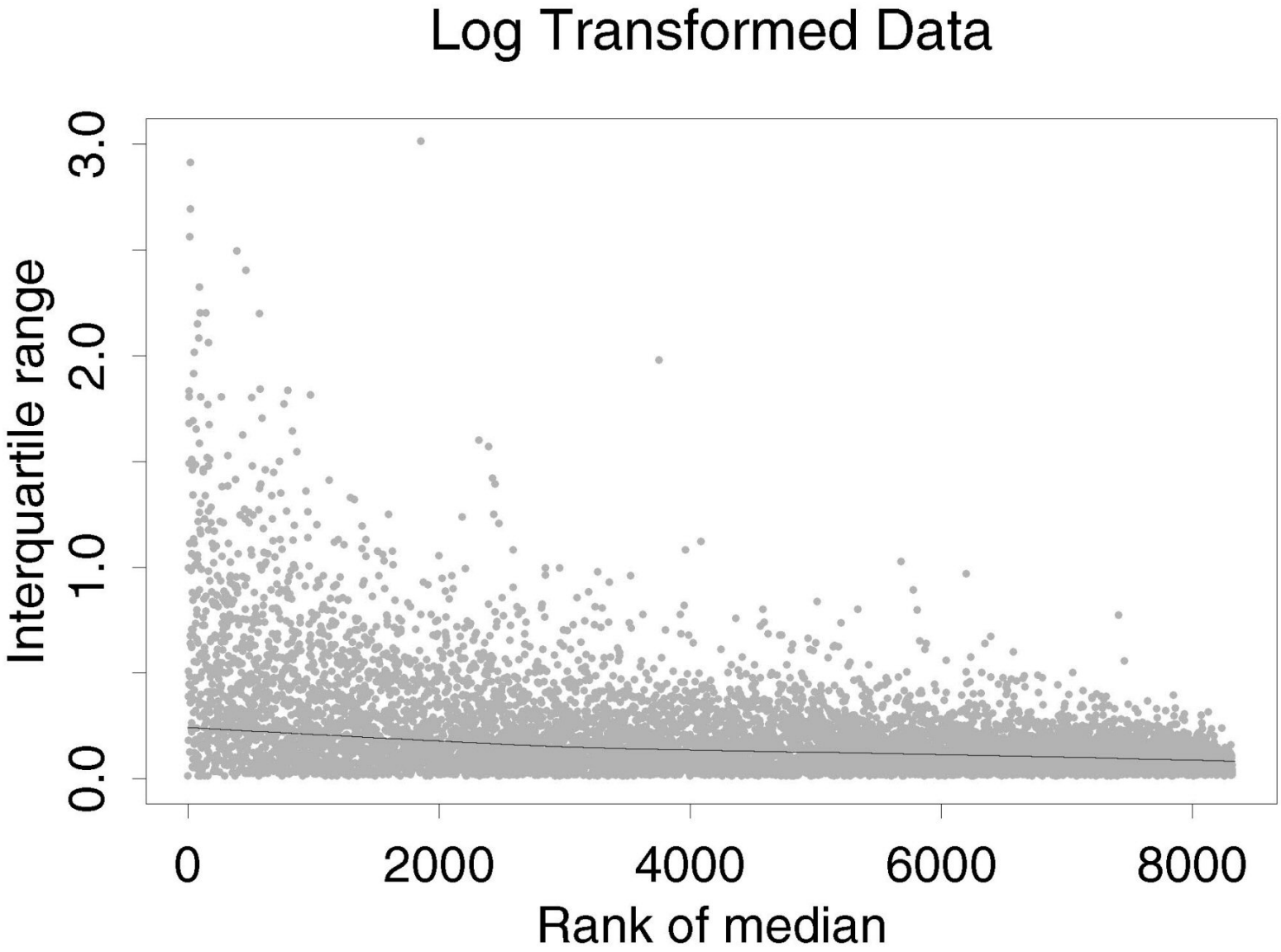
The statistical model used is

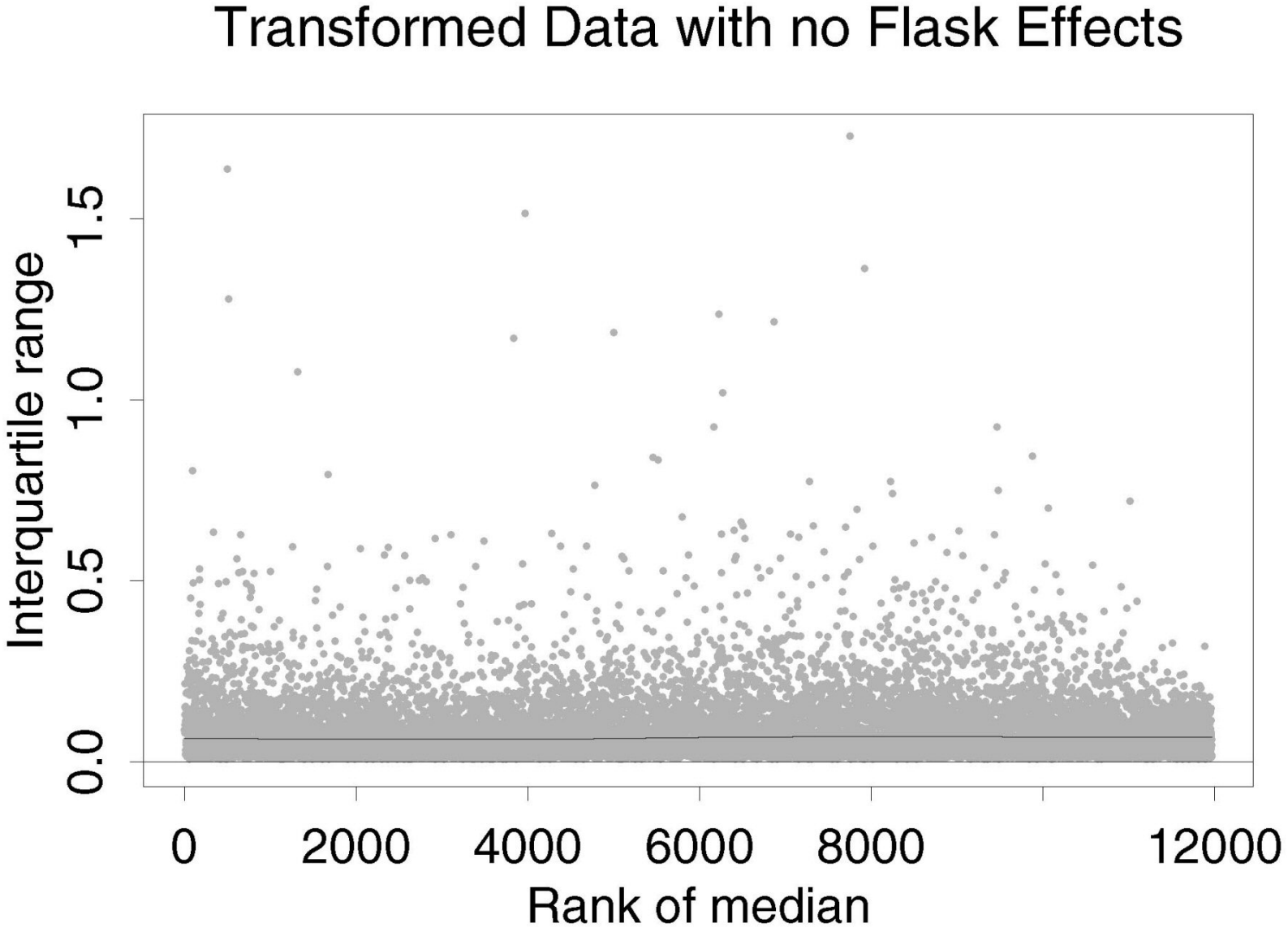
$$h_{\lambda,\alpha}(\text{intensity}) = \text{gene} + \text{chip} + \text{error}$$

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


Untransformed Data
Interquartile range
0
200
600
1000
0
2000
4000
6000
8000
10000
Median


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
Interquartile range
0.0
0.5
1.0
1.5
0
2000
4000
6000
8000
12000
Rank of median