"Time Perturbation" Studies of Somitogenesis Microarray Time Series

Earl F. Glynn

Scientific Programmer

Bioinformatics

Stowers Institute for Medical Research

5 Dec 2007

"Time Perturbation" Studies of Somitogenesis Microarray Time Series

- Purpose
- Background Concepts
- Microarray Analysis Overview
- Numerical Experiments
- Somitogenesis Experiments
- Summary

Purpose

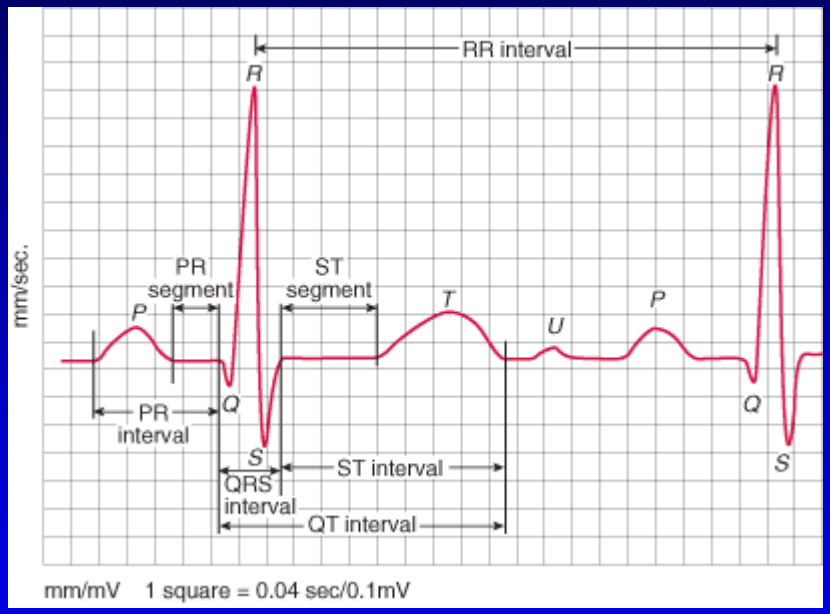
- Find "best" periodic genes from microarrary time series experiments exploring somitogenesis in mouse, chick and zebrafish.
- Develop general methodology that can be applied across all species.

Background Concepts

- Periodic Time Series in Biology
- Lomb-Scargle Analysis
- Hypothesis Testing with Lomb-Scargle
- Boxplots and "Notches"
- Rank Order and Rank Product

Periodic Time Series in Biology

Electrocardiogram (ECG): QRS Complex



Time

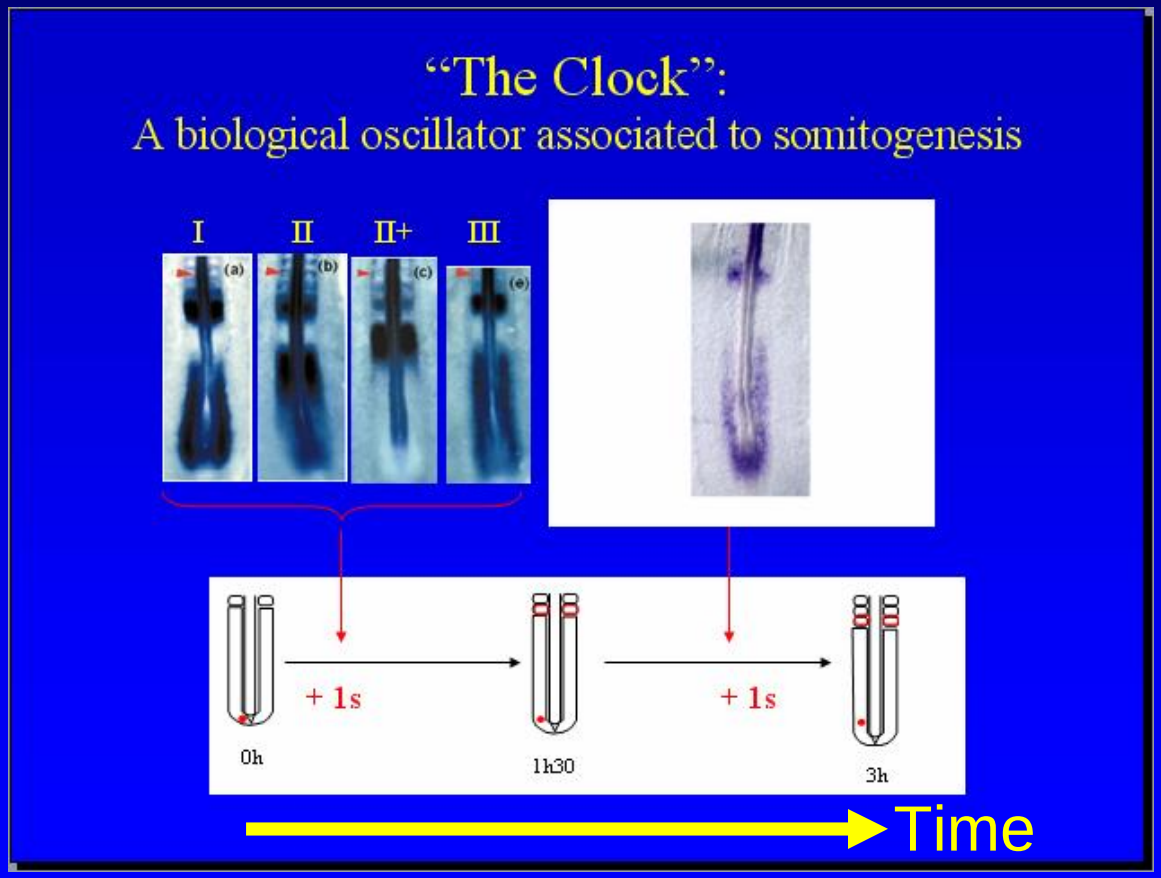
Source: http://www.merck.com/mmpe/print/sec07/ch070/ch070e.html

Continuous: $x(t + T) = x(t)$

Discrete: $x[n + N] = x[n]$

"Periodic" Time Series in Biology

Segmentation is Established During Somitogenesis

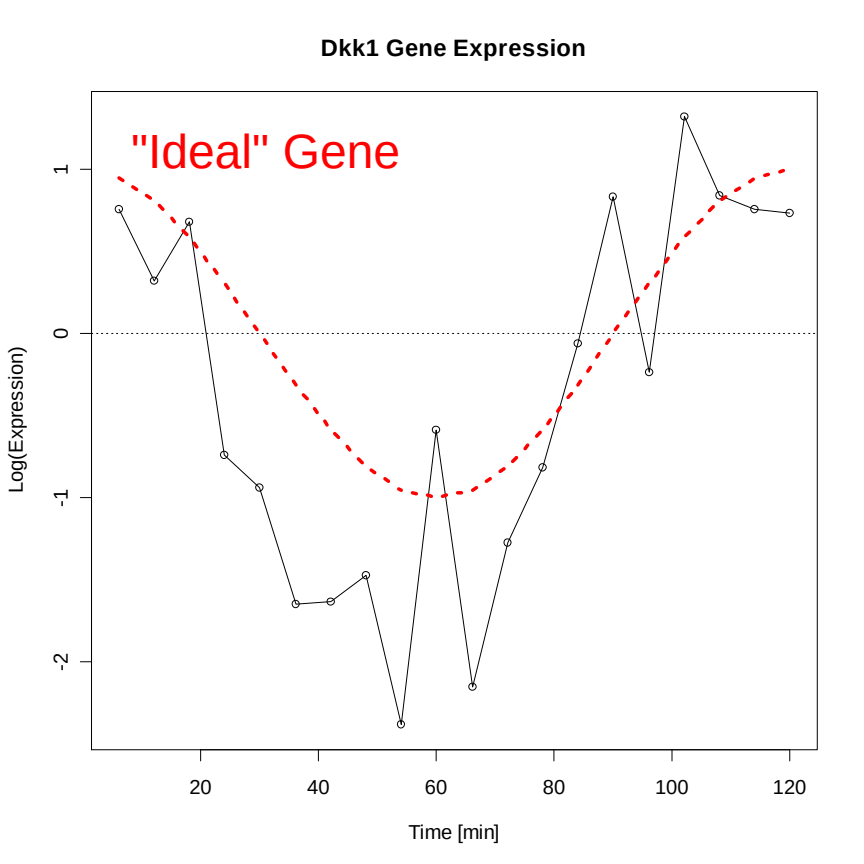


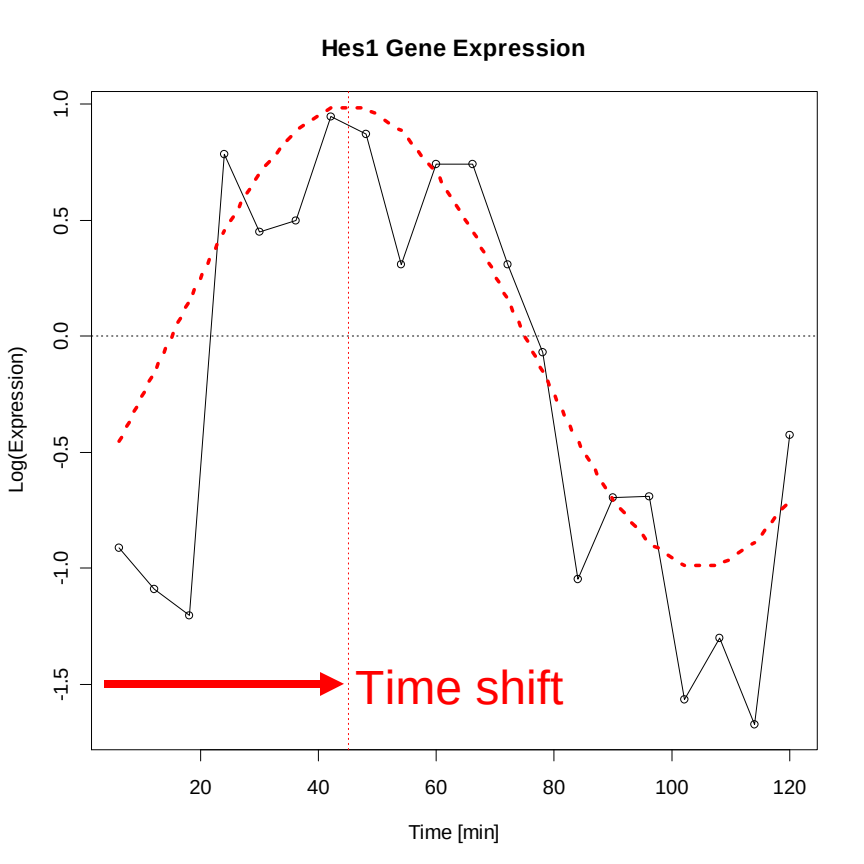
Mouse
Chick
Zebrafish
...

Source: Mary-Lee Dequéant, Stowers Science Club, May 23, 2003

"Periodic" Time Series in Biology

Somitogenesis





Source: Mary-Lee Dequéant, "Mouse 20" dataset, Nov 2004

Lomb-Scargle Periodogram Mathematical Details

$$h_i \equiv h(t_i), \quad i = 1, \ldots, N$$

$$\overline{h} \equiv \frac{1}{N}\sum_{1}^{N} h_i \qquad \sigma^2 \equiv \frac{1}{N-1}\sum_{1}^{N}(h_i - \overline{h})^2 \tag{13.8.3}$$

Now, the Lomb *normalized periodogram* (spectral power as a function of angular frequency $\omega \equiv 2\pi f > 0$) is defined by

$$P_N(\omega) \equiv \frac{1}{2\sigma^2}\left\{\frac{\left[\sum_j (h_j - \overline{h})\cos\omega(t_j - \tau)\right]^2}{\sum_j \cos^2\omega(t_j - \tau)} + \frac{\left[\sum_j (h_j - \overline{h})\sin\omega(t_j - \tau)\right]^2}{\sum_j \sin^2\omega(t_j - \tau)}\right\} \tag{13.8.4}$$

Here τ is defined by the relation

$$\tan(2\omega\tau) = \frac{\sum_j \sin 2\omega t_j}{\sum_j \cos 2\omega t_j} \tag{13.8.5}$$

$$P(>z) \equiv 1 - (1 - e^{-z})^M \tag{13.8.7}$$

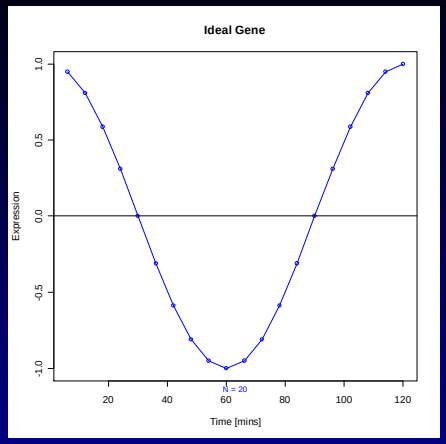
is the false-alarm probability of the null hypothesis, that is, the *significance level* of any peak in $P_N(\omega)$ that we do see.

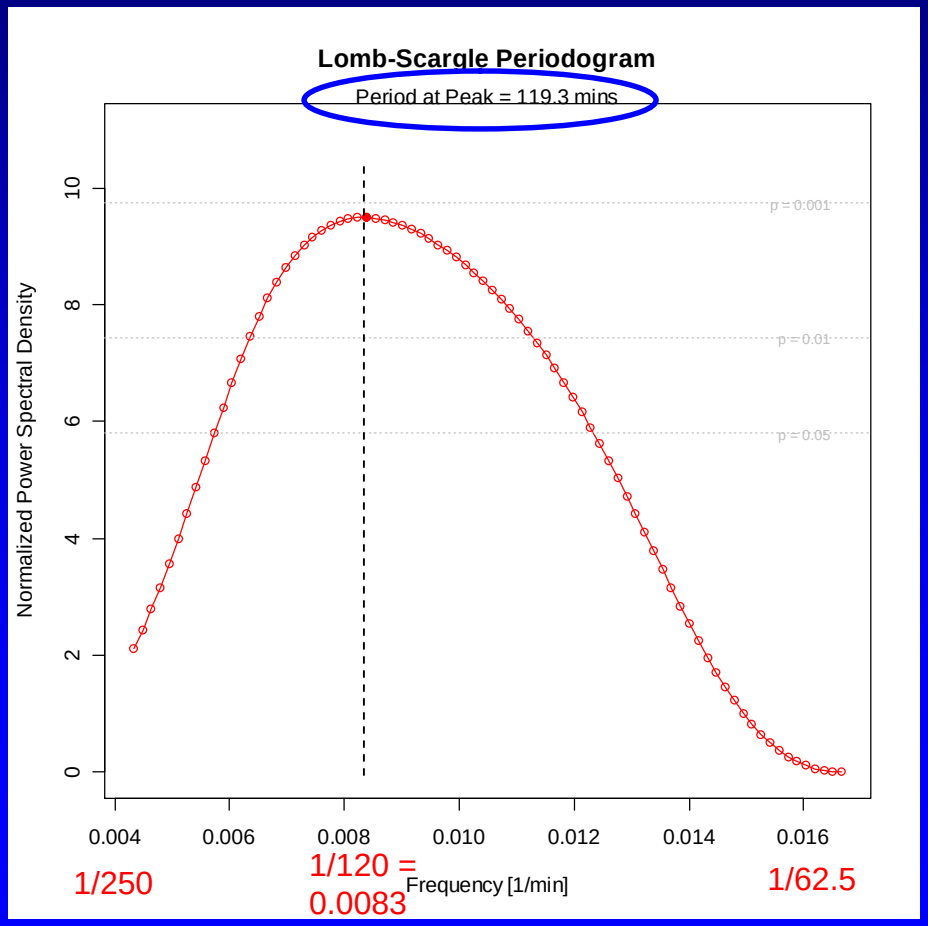
$P_N(\omega)$ has an exponential probability distribution with unit mean.

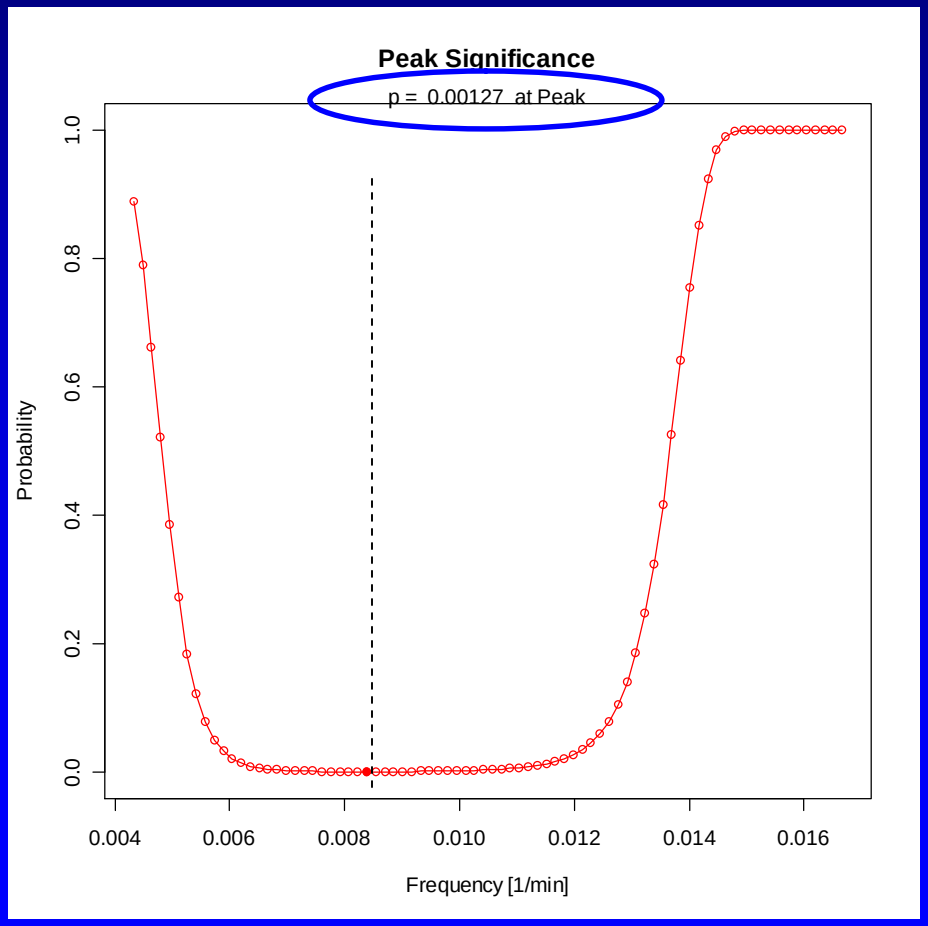
Source: *Numerical Recipes in C* (2nd Ed), 1992, p. 577

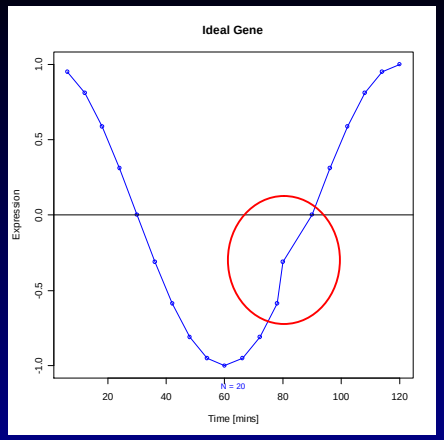
Lomb-Scargle Analysis:

"Ideal Gene"





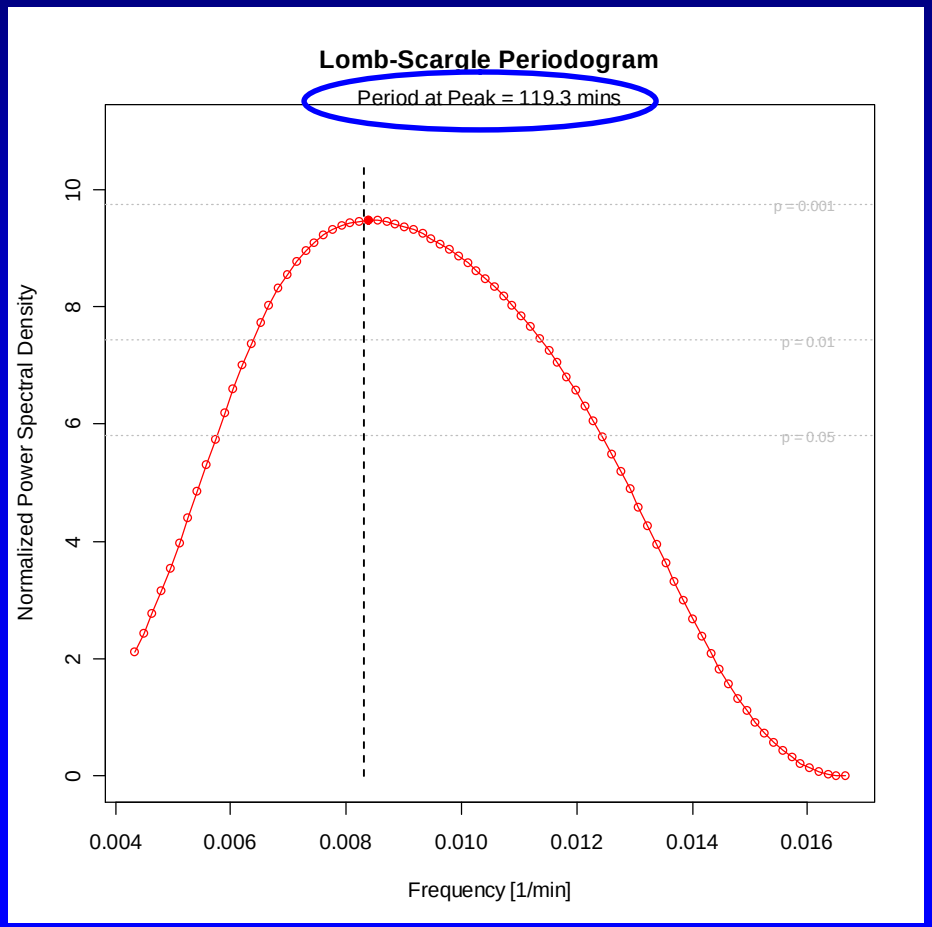


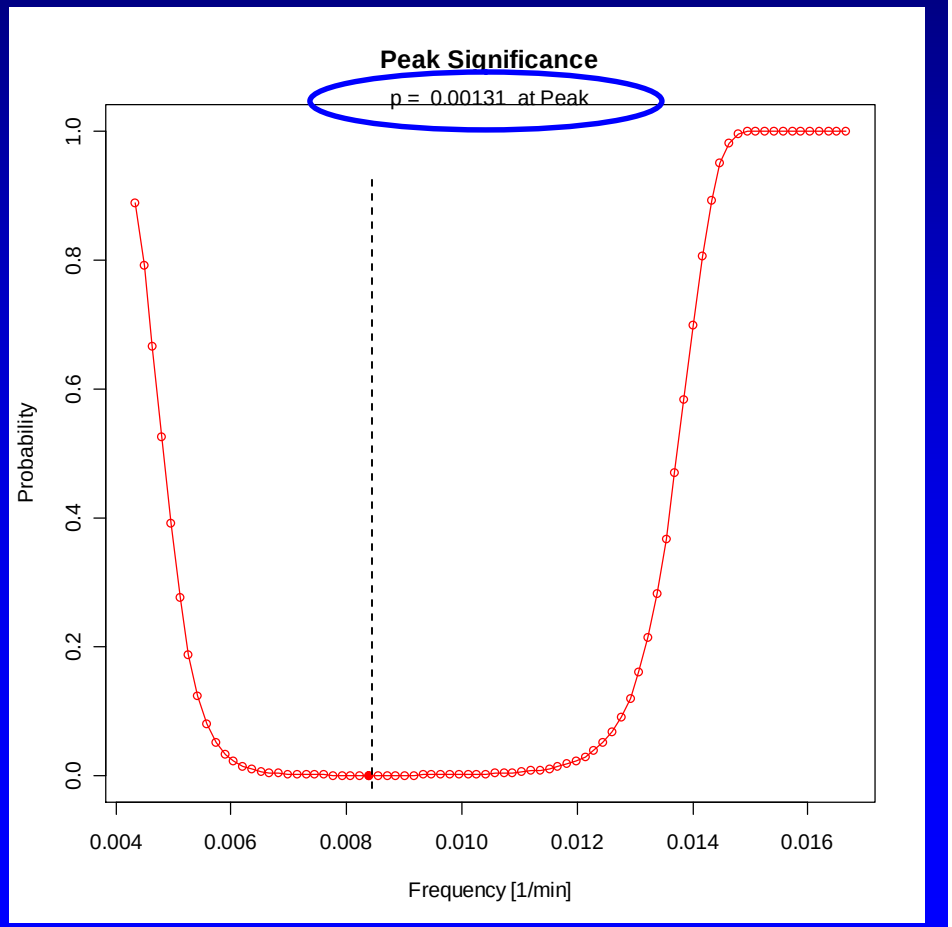


Lomb-Scargle Analysis:

"Ideal Gene" with one "bad" time point

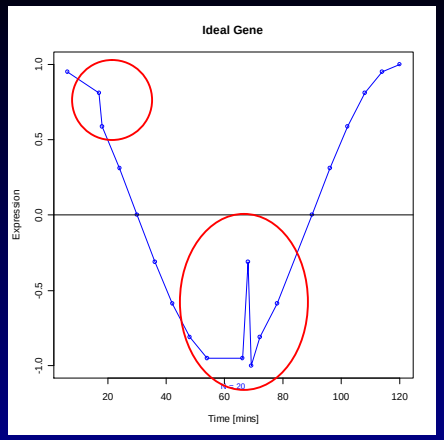
p =0.00127 for "perfect"



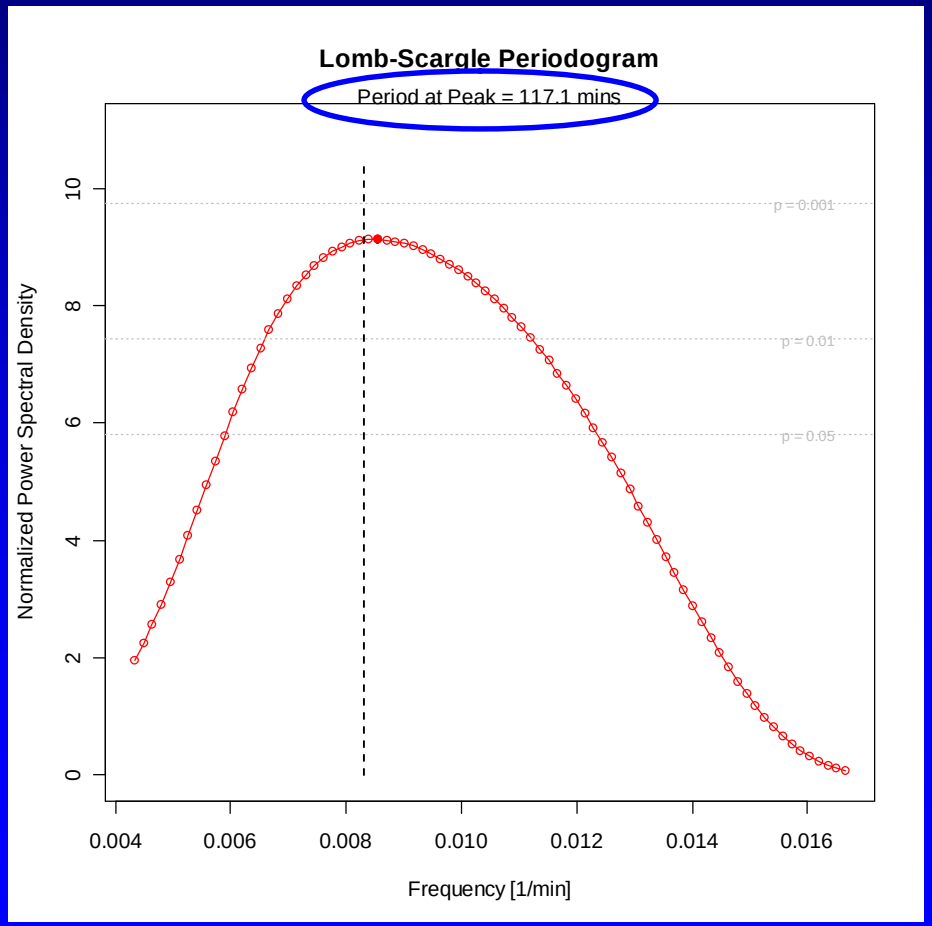


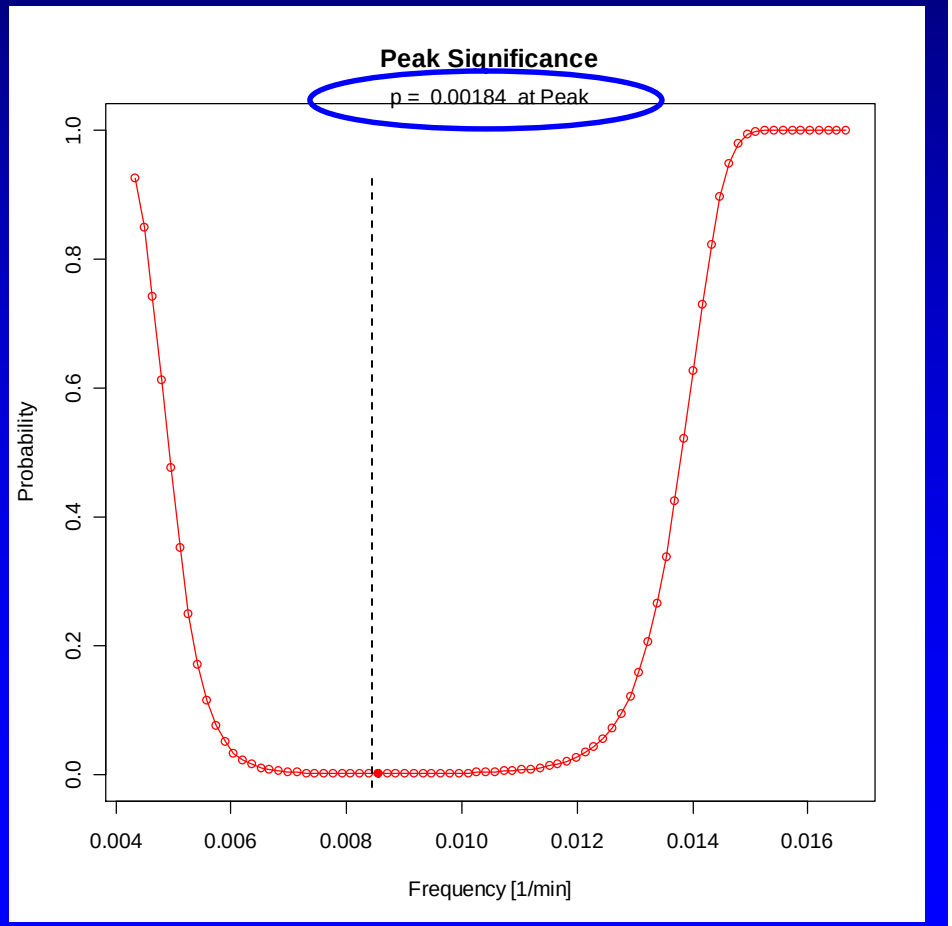
Lomb-Scargle Analysis:

"Ideal Gene" with three "bad" time points



p =0.00127 for "perfect"





Lomb-Scargle Analysis

Hypothesis Testing

SIGNIFICANCE TESTING OF PERIODOGRAM ORDINATES

CHRIS KOEN
Institute of Physics and Astronomy, National Central University, Republic of China

THE ASTROPHYSICAL JOURNAL, 348:700–702, 1990 January 10

III. HYPOTHESIS TESTING

The theory developed above can now be used to test hypotheses. Two hypotheses are relevant:

H_1: the observations do not constitute noise

H_2: the data contain a specific periodicity .

In this paper only a special case of H_2 will be considered, namely,

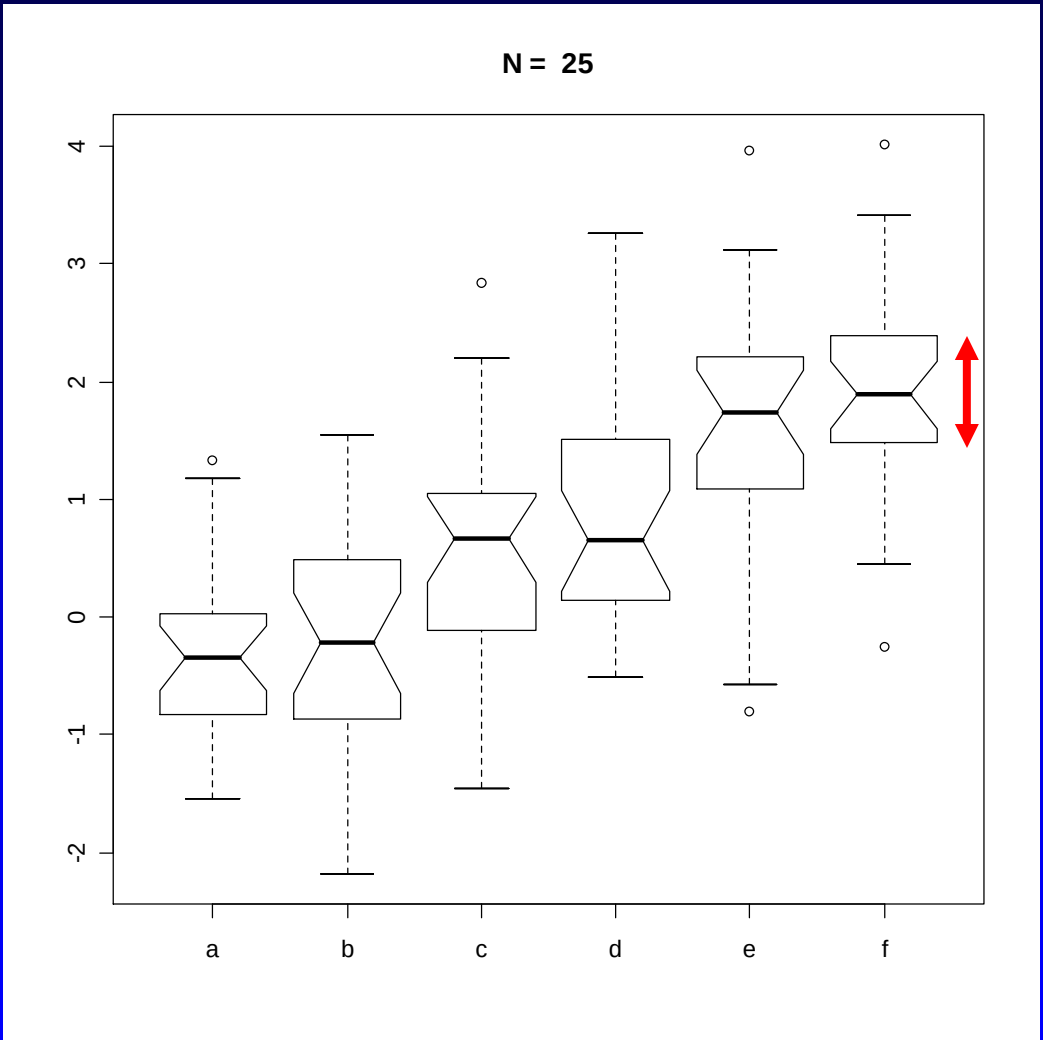
H_2: the largest periodogram ordinate is significant .

Koen suggests Lomb-Scargle evaluates H_1, while a Fisher test is needed for H_2, but ...

Box Plots and Notches

Means
a = 0.0
b = 0.0
c = 0.5
d = 1.0
e = 1.5
f = 2.0

S.D. = 1.0



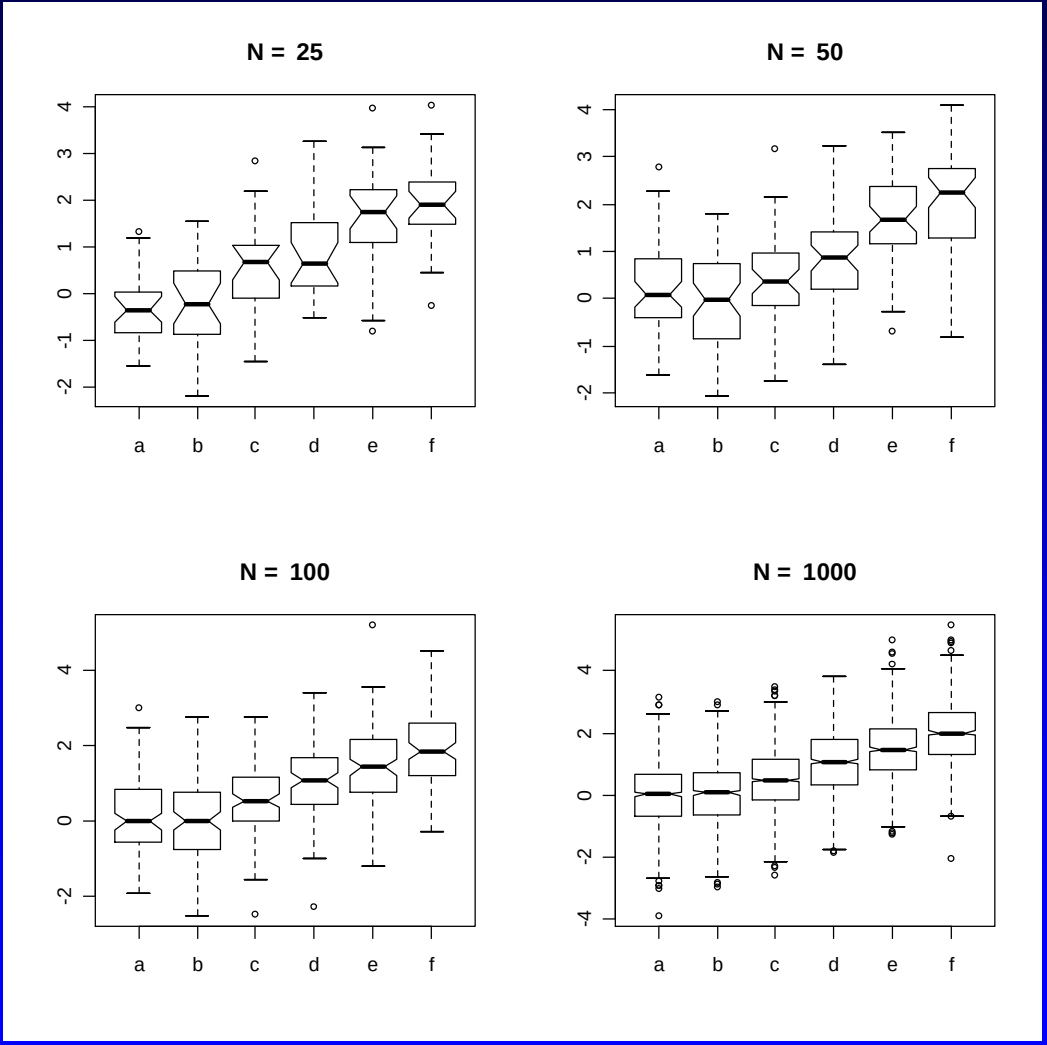
IQR:
Inter
Quartile
Range

Significant difference at 5% level when notches do not overlap

Box Plots and Notches

Means
a = 0.0
b = 0.0
c = 0.5
d = 1.0
e = 1.5
f = 2.0

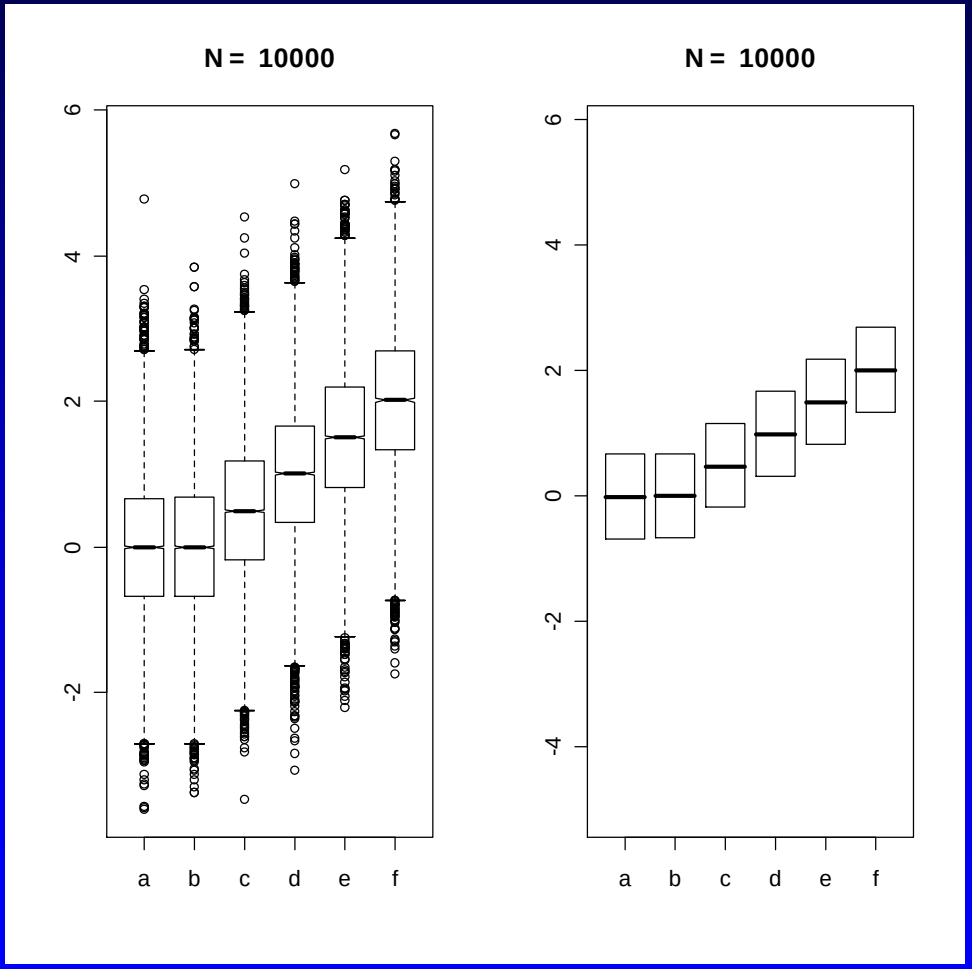
S.D. = 1.0



Box Plots

Means
a = 0.0
b = 0.0
c = 0.5
d = 1.0
e = 1.5
f = 2.0

S.D. = 1.0



Most boxplots on slides to follow have N=10,000.
Notches, "whiskers," outliers will usually not be shown.

Rank Order and Rank Product

Gene	Trial 1 Rank	Trial 2 Rank	Trial 3 Rank	Rank Product	Overall Rank
gene 1	1	2	1	2	1
gene 2	2	3	3	18	3
gene 3	3	1	2	6	2

Ranks may be more robust with weak signals.

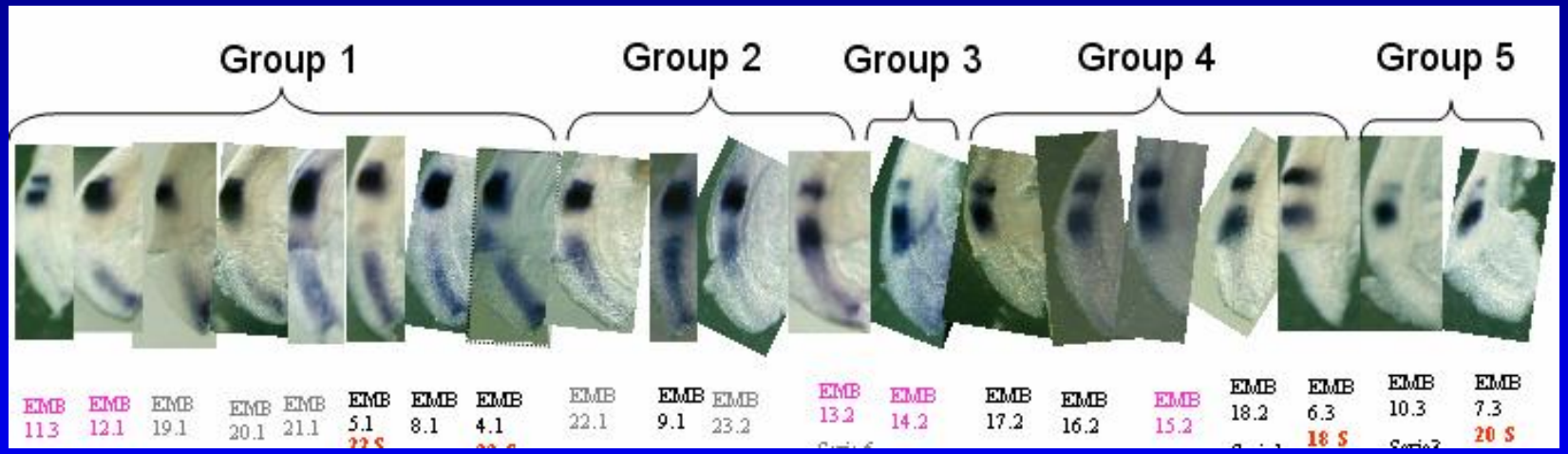
"Time Perturbation" Studies of Somitogenesis Microarray Time Series

- Purpose
- Background Concepts
- Microarray Analysis Overview
- Numerical Experiments
- Somitogenesis Experiments
- Summary

Microarray Analysis Overview

Somitogenesis Datasets: Mouse, Chick, Zebrafish

Time series from in situ hybridization images

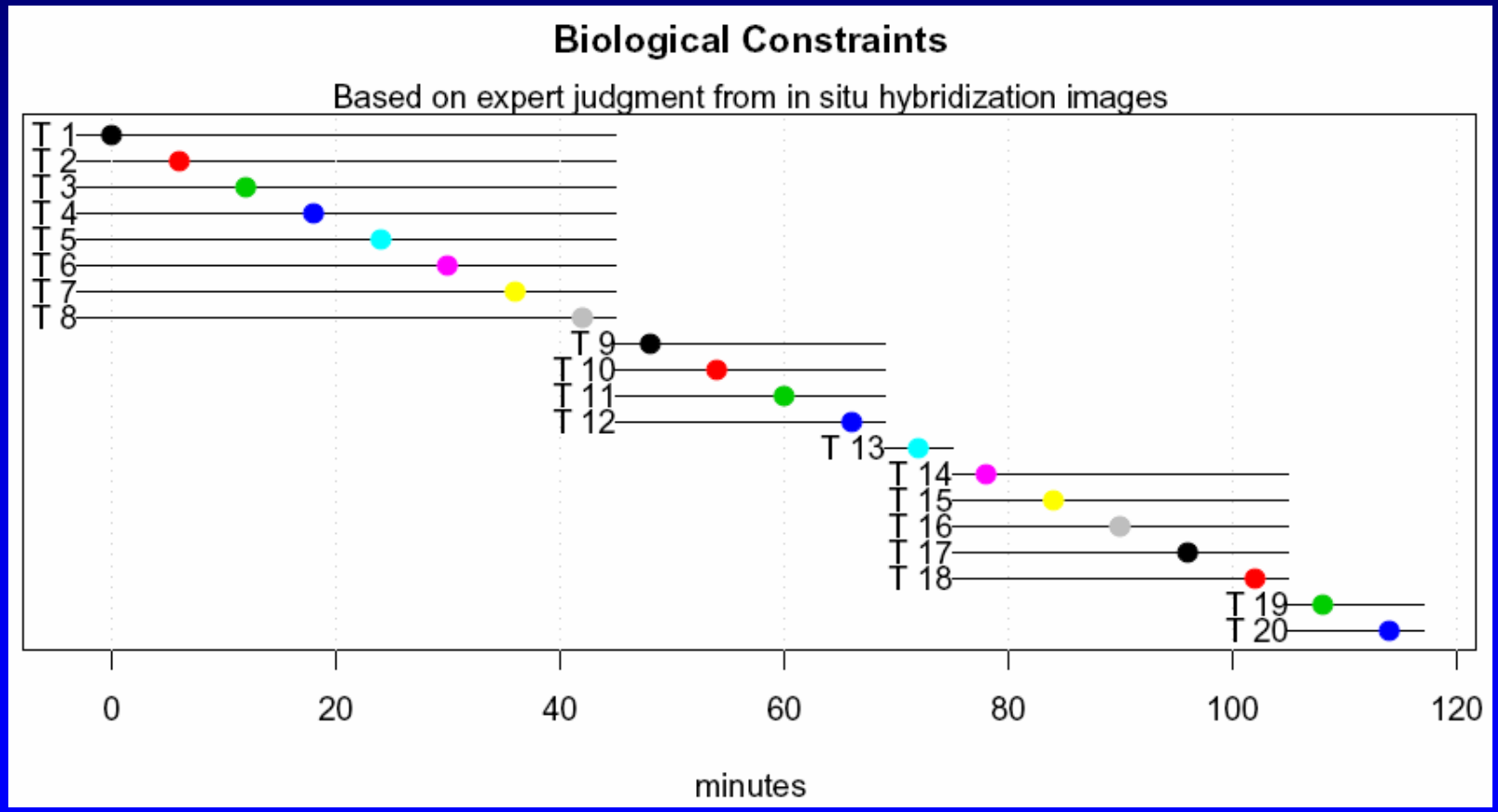


Time

- What time values should be used in time series?
- What biological constraints define groups?

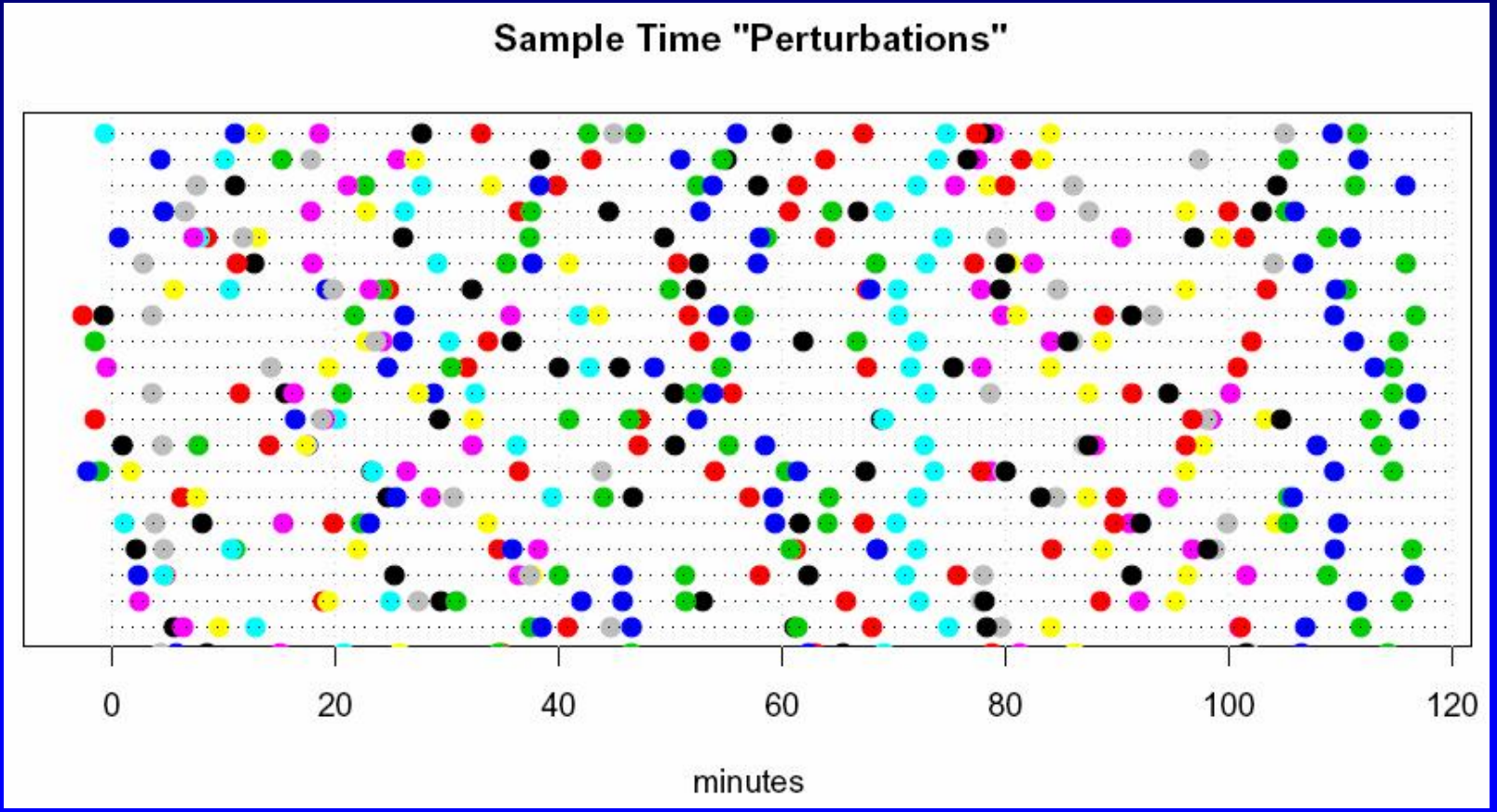
Source: Mary-Lee Dequéant, Order Nveau.ppt, circa Jan 2005

Time Perturbations



Dots show position of "baseline" times.

Time Perturbations



Each row of dots represents one time ordering.

Somitogenesis Datasets

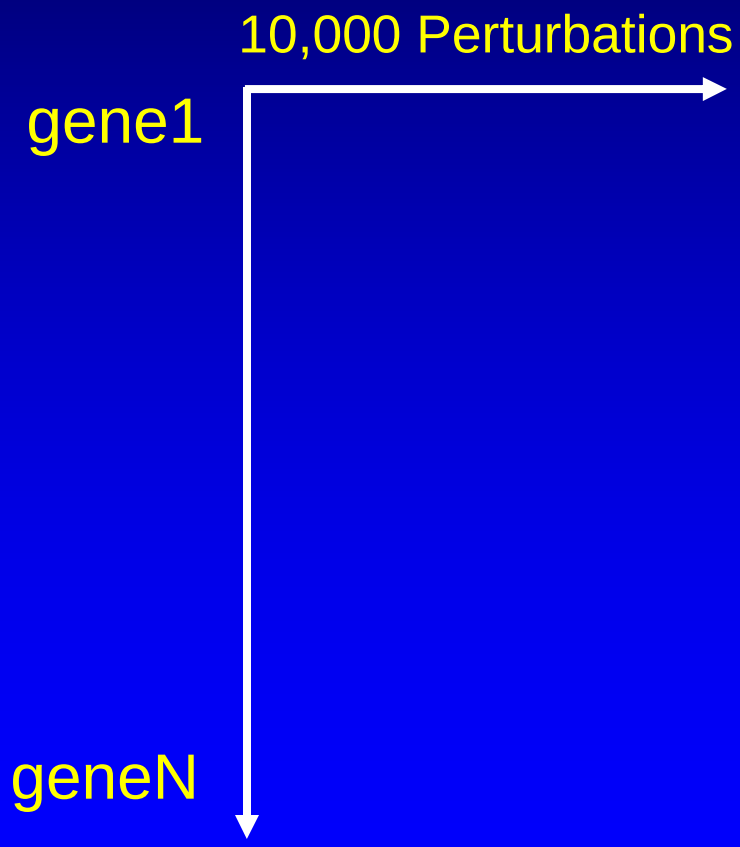
Time Perturbations

Affymetrix Chips

Mouse
N=22,690 or 45,101

Chick
N=38,535

Zebrafish
N=15,618



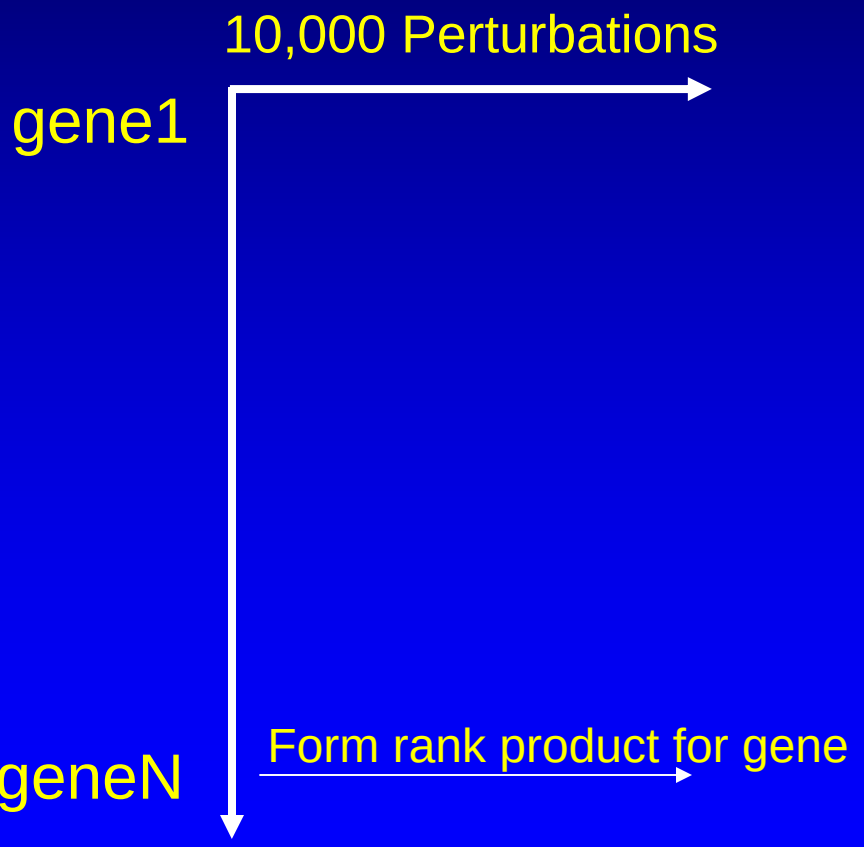
Somitogenesis Datasets

Time Perturbations

1. Perform Lomb-Scargle analysis on each of 10,000 perturbations for each gene. [Cluster problem]

2. For given perturbation, rank order genes based on Lomb-Scargle p-value.

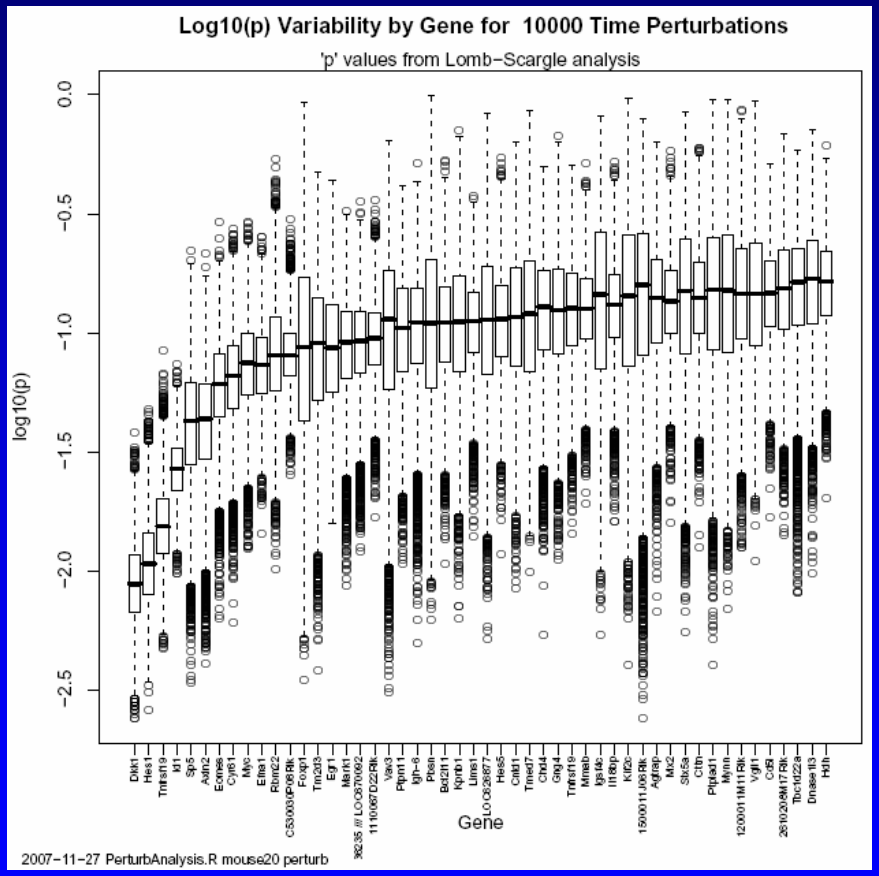
3. Form rank product for each gene for overall ranking.



Somitogenesis Datasets

Lomb-Scargle p-values

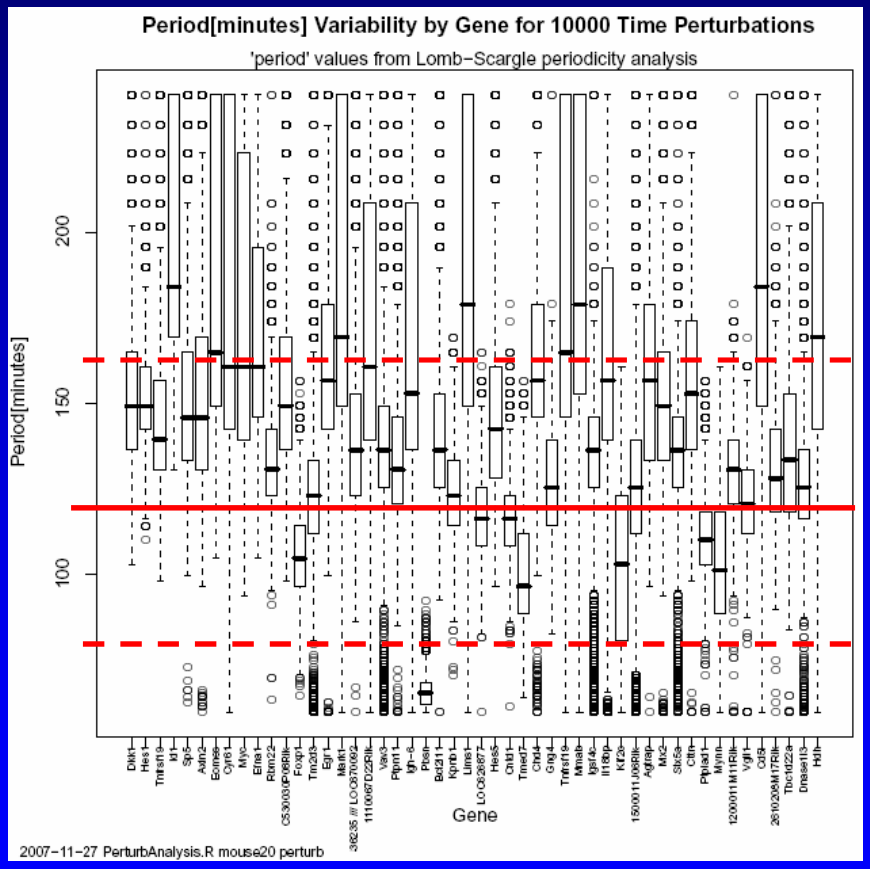
Ordered by "Overall Rank" (i.e., Rank product of 10,000 ranks from Lomb-Scargle p-values.)



Looks "good" but need to know how to interpret results.

Somitogenesis Datasets

Lomb-Scargle Periods



Mouse somite period is assumed to be 120 minutes

Need to filter results for period of biological interest.

"Time Perturbation" Studies of Somitogenesis Microarray Time Series

- Purpose
- Background Concepts
- Microarray Analysis Overview
- Numerical Experiments
- Somitogenesis Experiments
- Summary

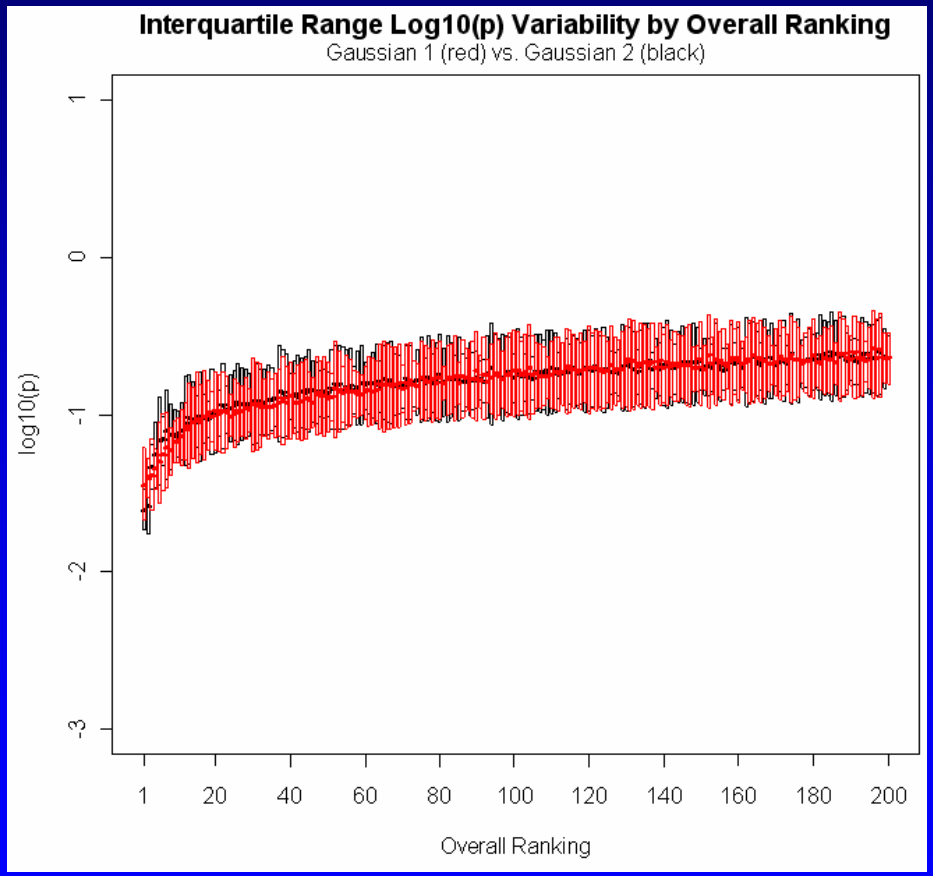
Numerical Experiments

"Controlled Experiments"

- Gaussian Noise vs. Gaussian Noise
- 2.5% "perfect" periodic genes
- 20 genes with signal-to-noise gradient
- 16 genes, half good, half not so good

Gaussian Noise vs. Gaussian Noise

Assume GeneChip experiments once normalized are roughly Gaussian.

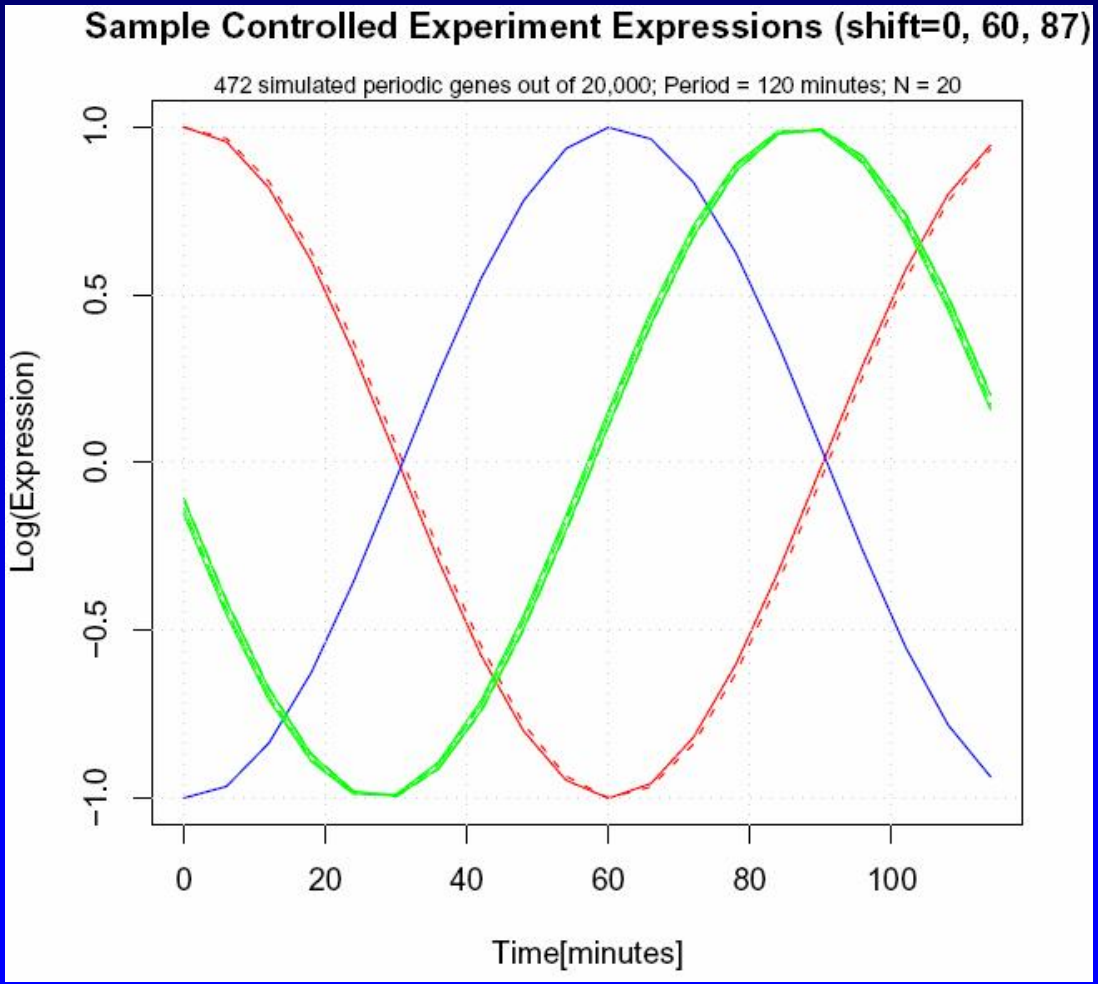


Compare signal rank with "noise" rank.

2.5% "perfect" periodic genes

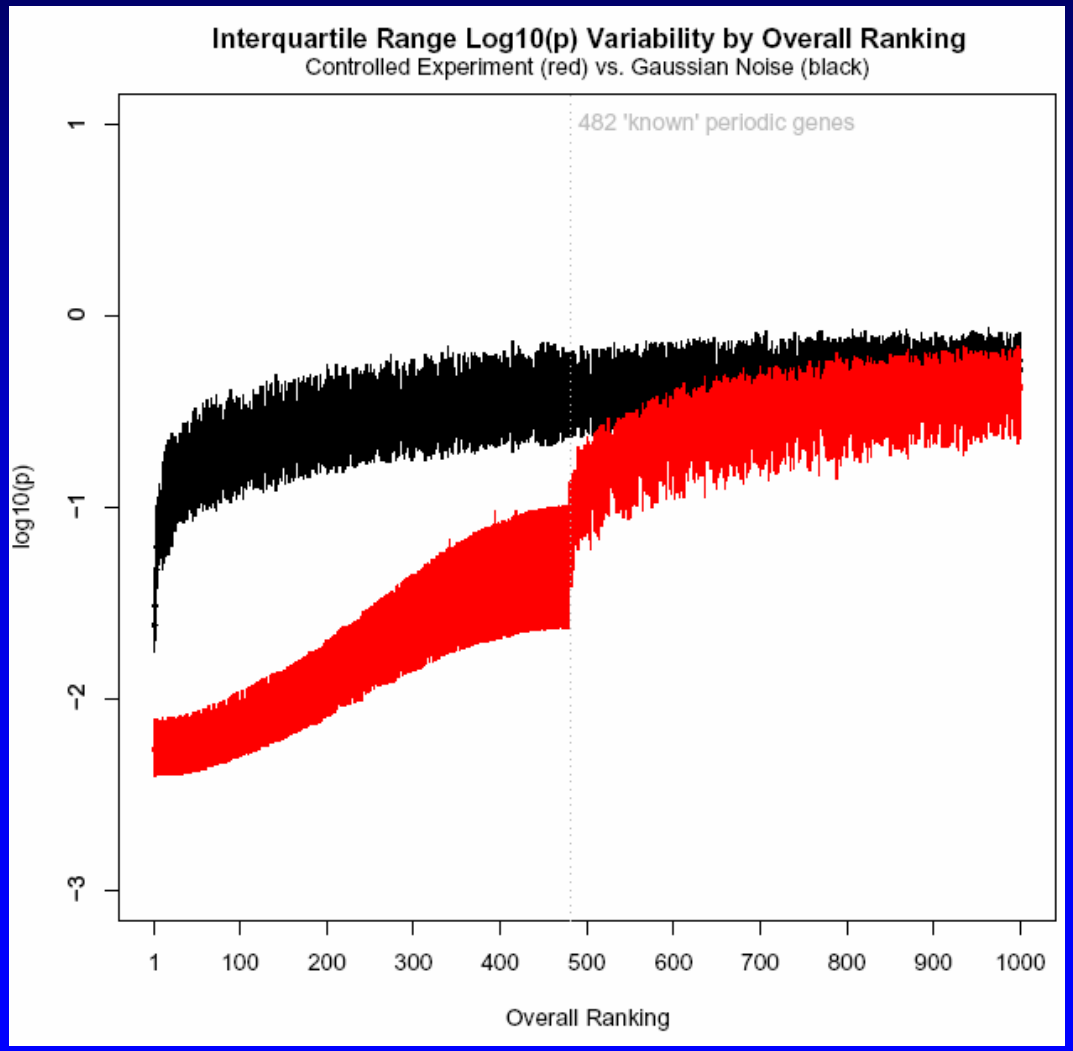
What if many strong periodic genes are involved in somitogenesis?

482 synthetic genes out of 20,000 were defined as sinusoids with various phase shifts



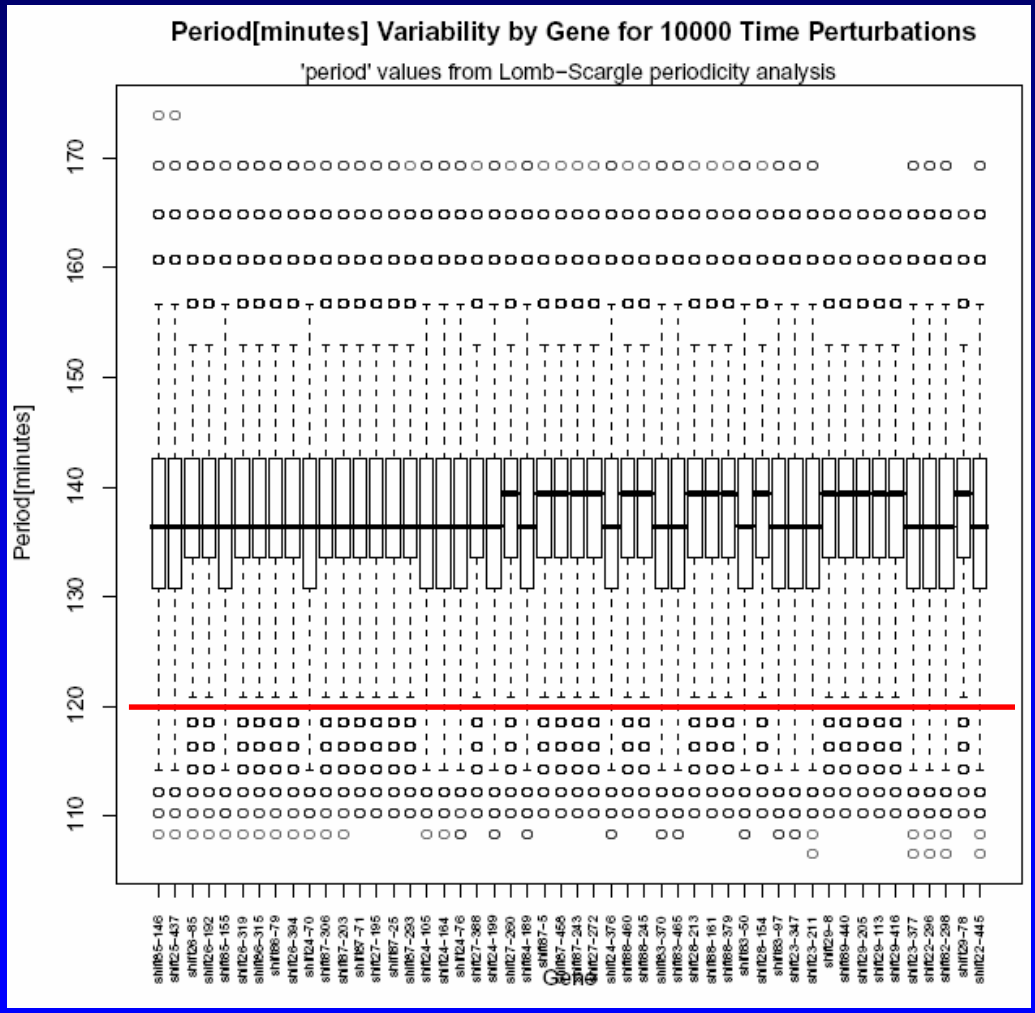
2.5% "perfect" periodic genes

What if many strong periodic genes are involved in somitogenesis?



2.5% "perfect" periodic genes

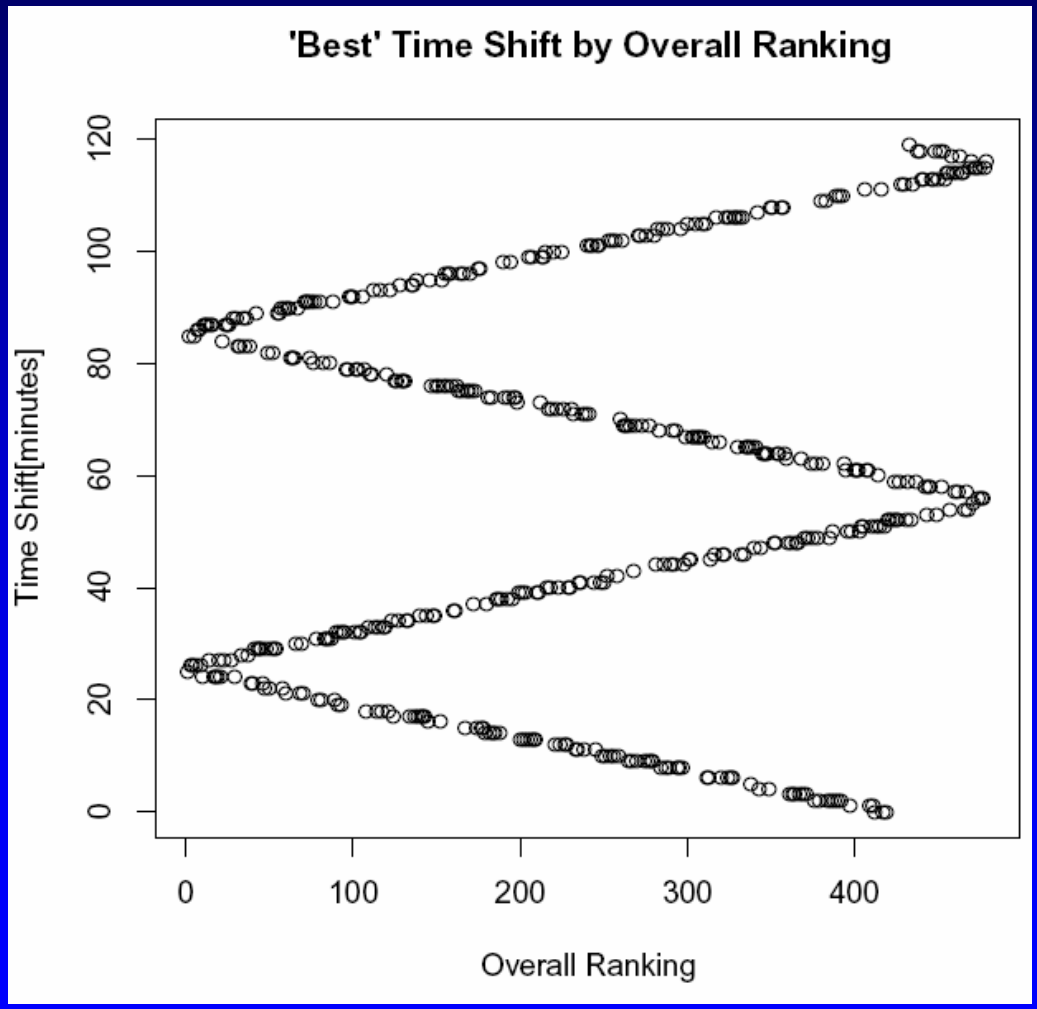
Are all periods this same?



Mouse somite period is assumed to be 120 minutes

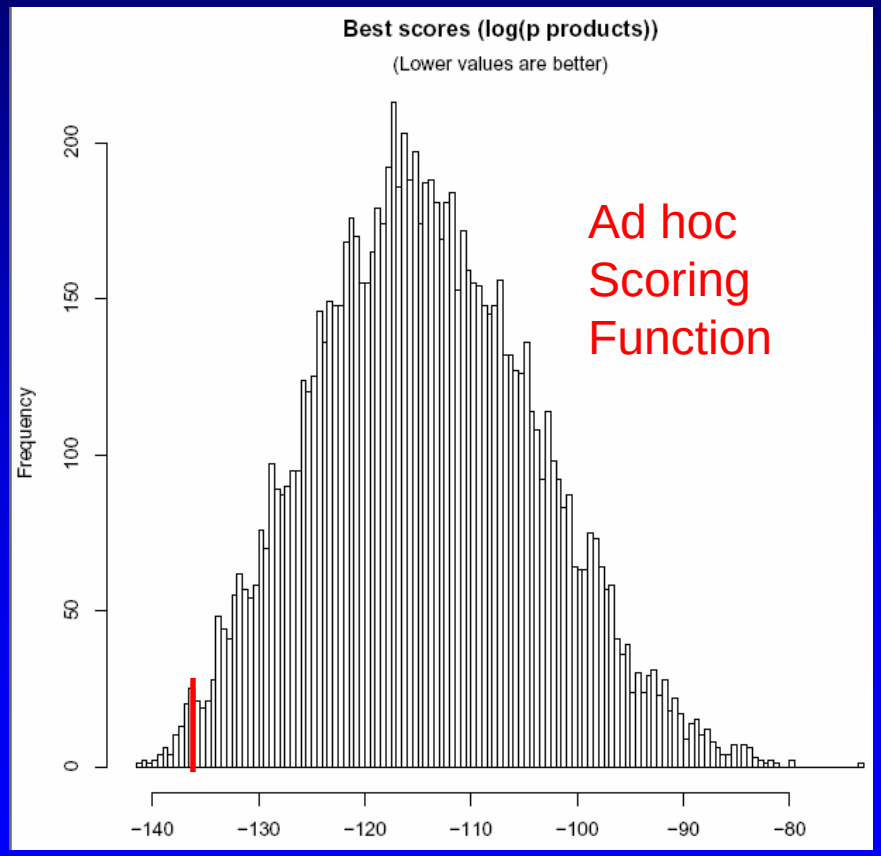
2.5% "perfect" periodic genes

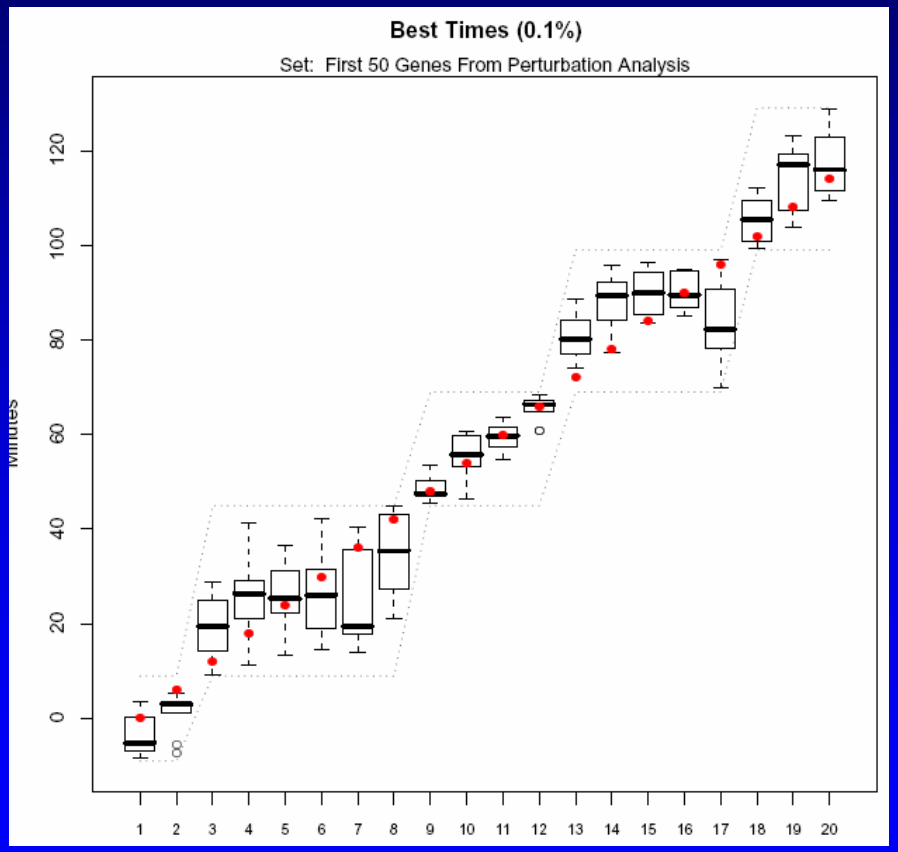
Are all time shifts "equal"?



2.5% "perfect" periodic genes

Can time series time points be estimated?



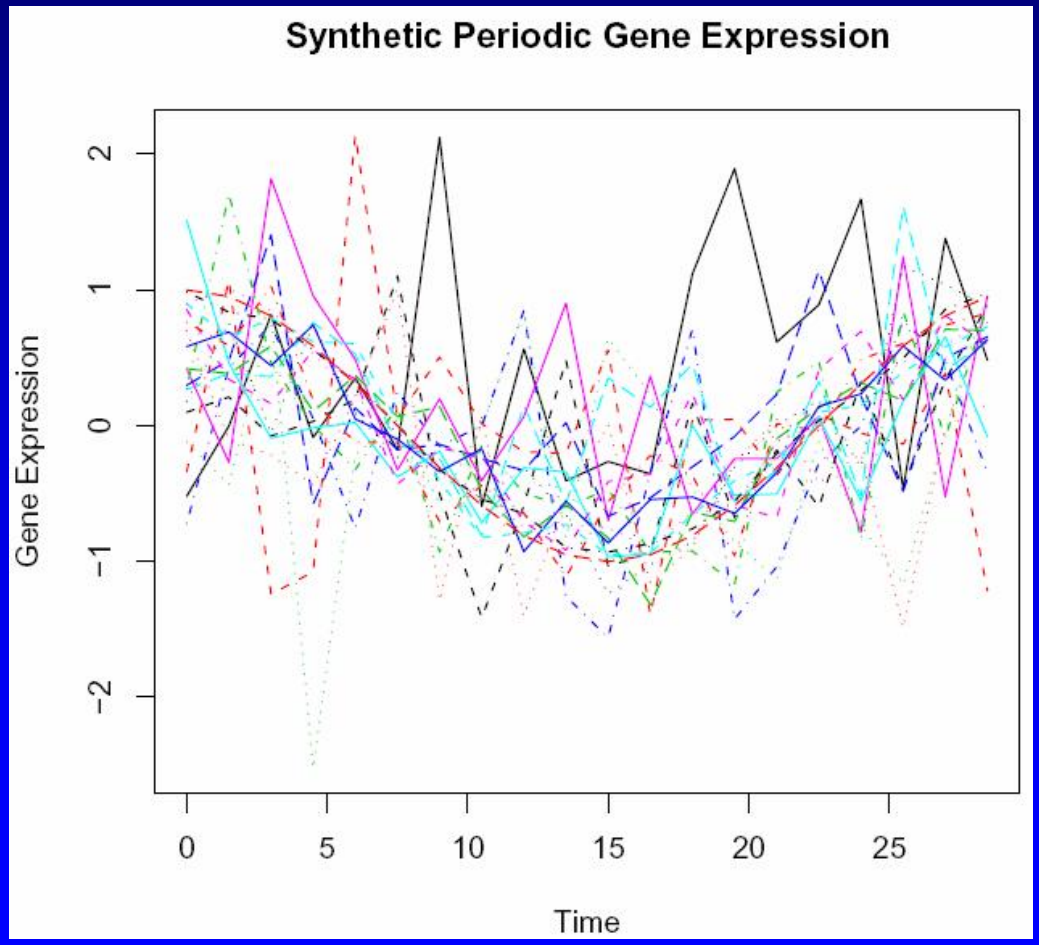


Maybe. Quite time consuming to explore alternatives.

20 genes with signal-to-noise gradient

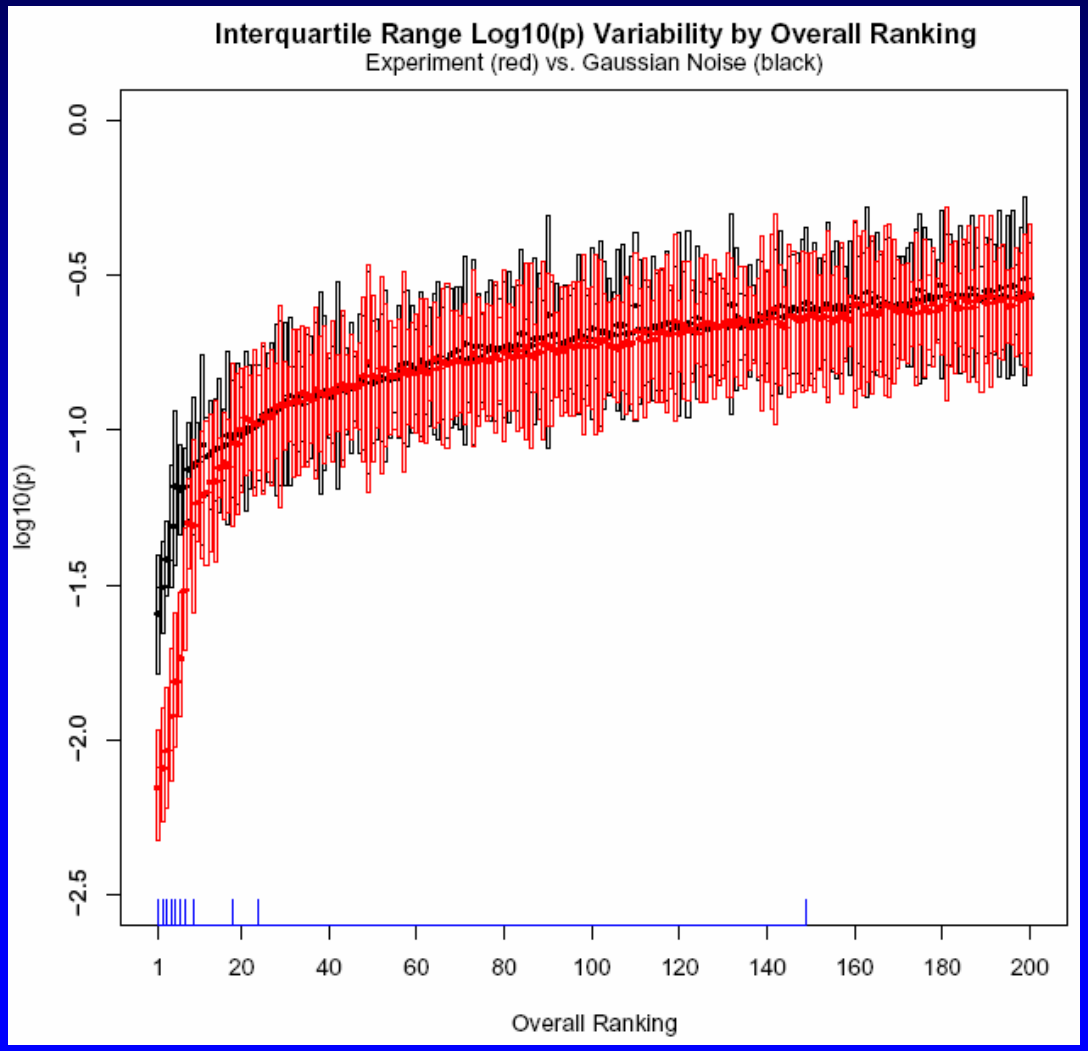
Expression = w * Cosine + (1-w) * Gaussian

w = 0.00 to 1.00
by 0.05



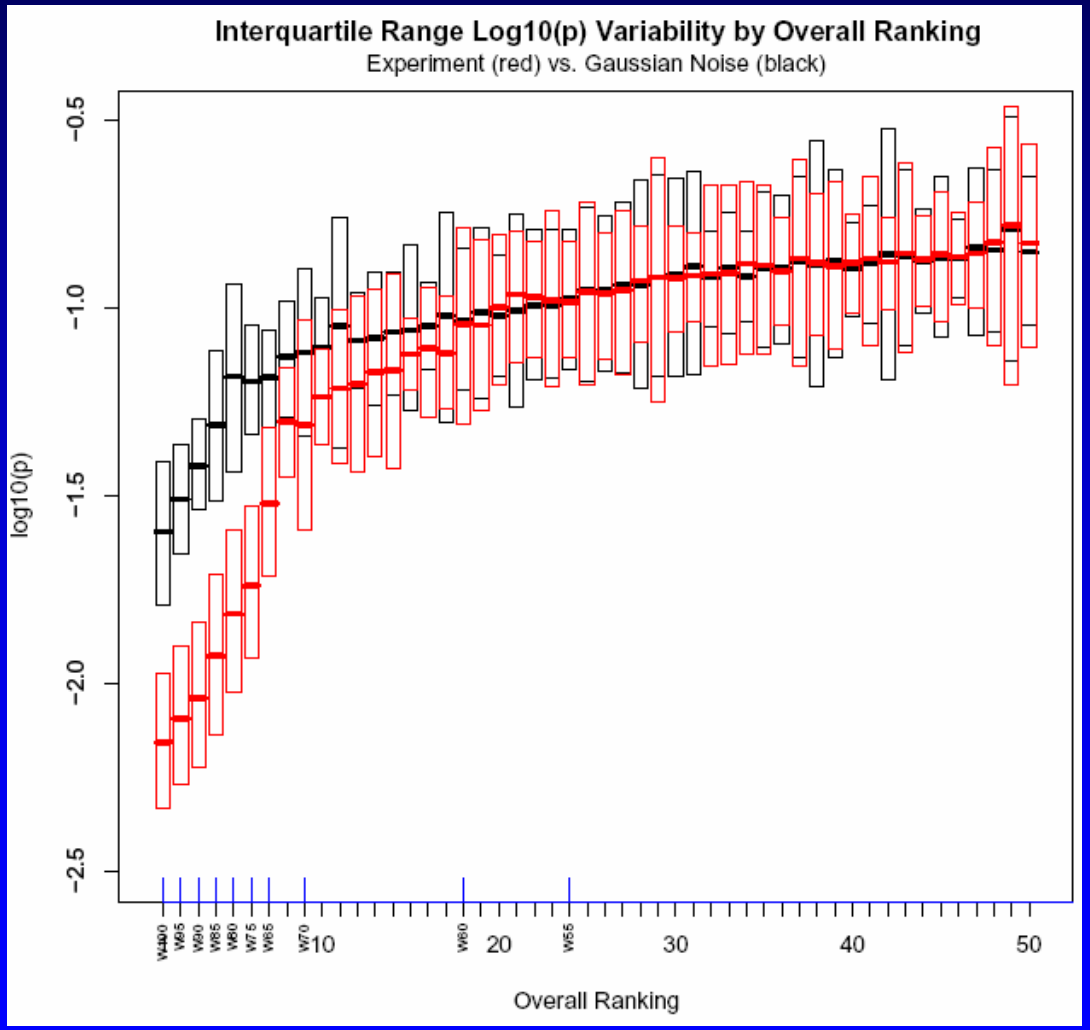
20 genes with signal-to-noise gradient

Rank	w
1	1.00
2	0.95
3	0.90
4	0.85
5	0.80
6	0.75
7	0.65
9	0.70
18	0.60
24	0.55
149	0.45

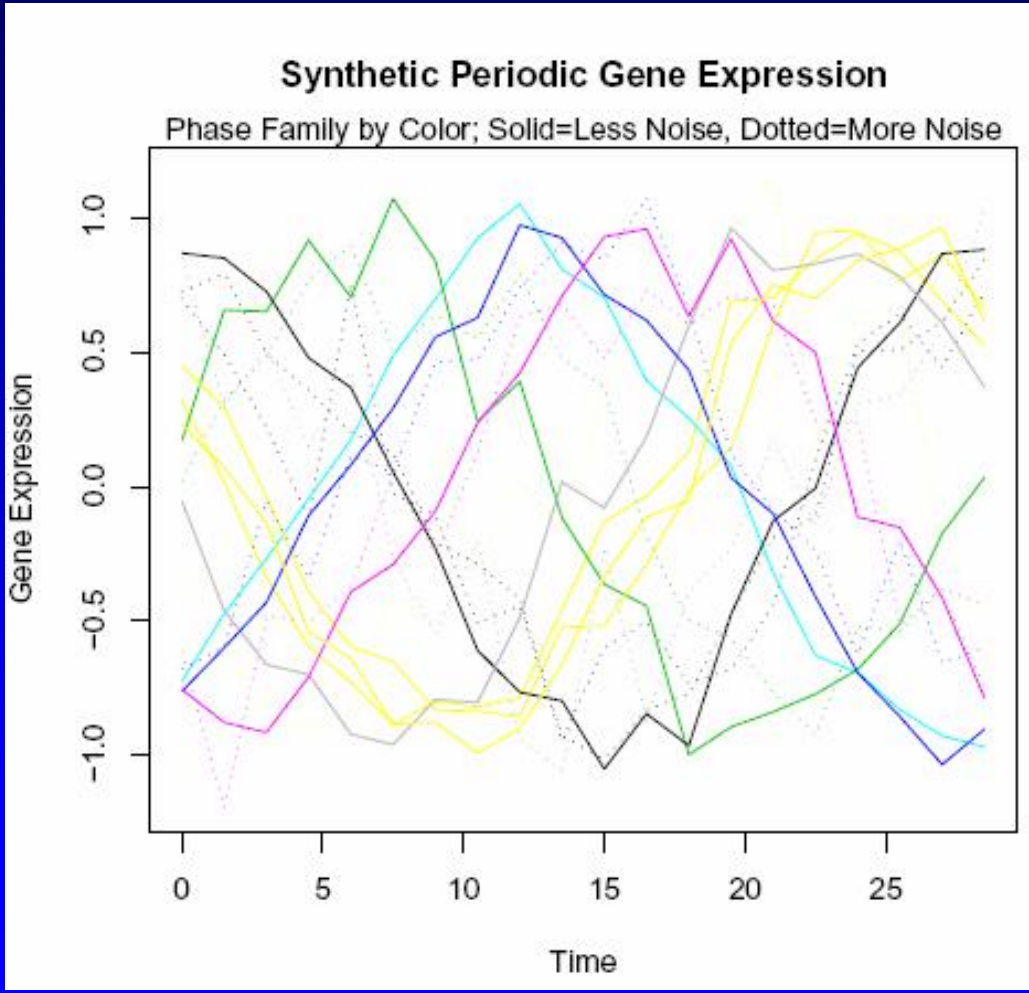


20 genes with signal-to-noise gradient

Rank	w
1	1.00
2	0.95
3	0.90
4	0.85
5	0.80
6	0.75
7	0.65
9	0.70
18	0.60
24	0.55
149	0.45



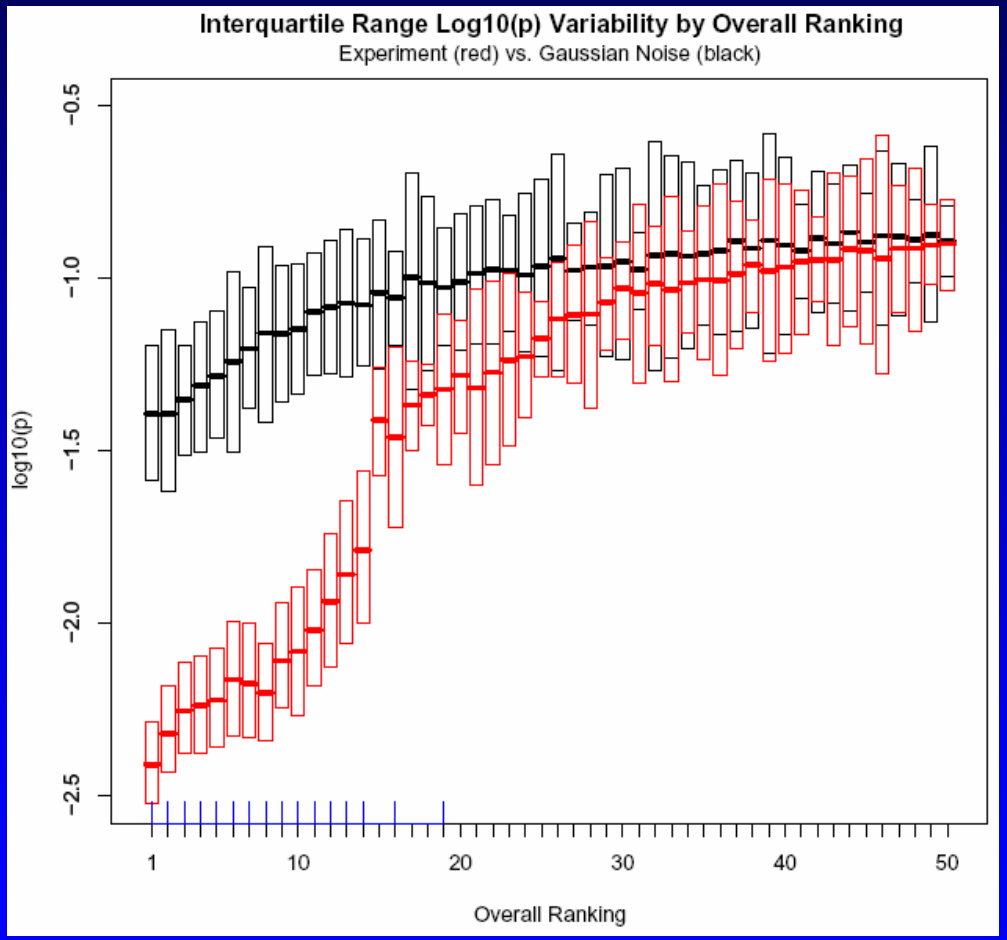
16 genes, half good, half not so good



16 genes, half good, half not so good

Rank	w	shift[min]
1	0.90	26.25
2	0.90	22.5
3	0.75	7.5
4	0.90	7.5
5	0.90	22.5
6	0.90	11.25
7	0.90	4
8	0.90	22.5
9	0.75	18.75
10	0.90	0
11	0.90	18.75
12	0.75	0
13	0.75	11.25
14	0.75	22.5
16	0.75	0
19	0.75	26.25

Zebrafish somite period is assumed to be 30 minutes

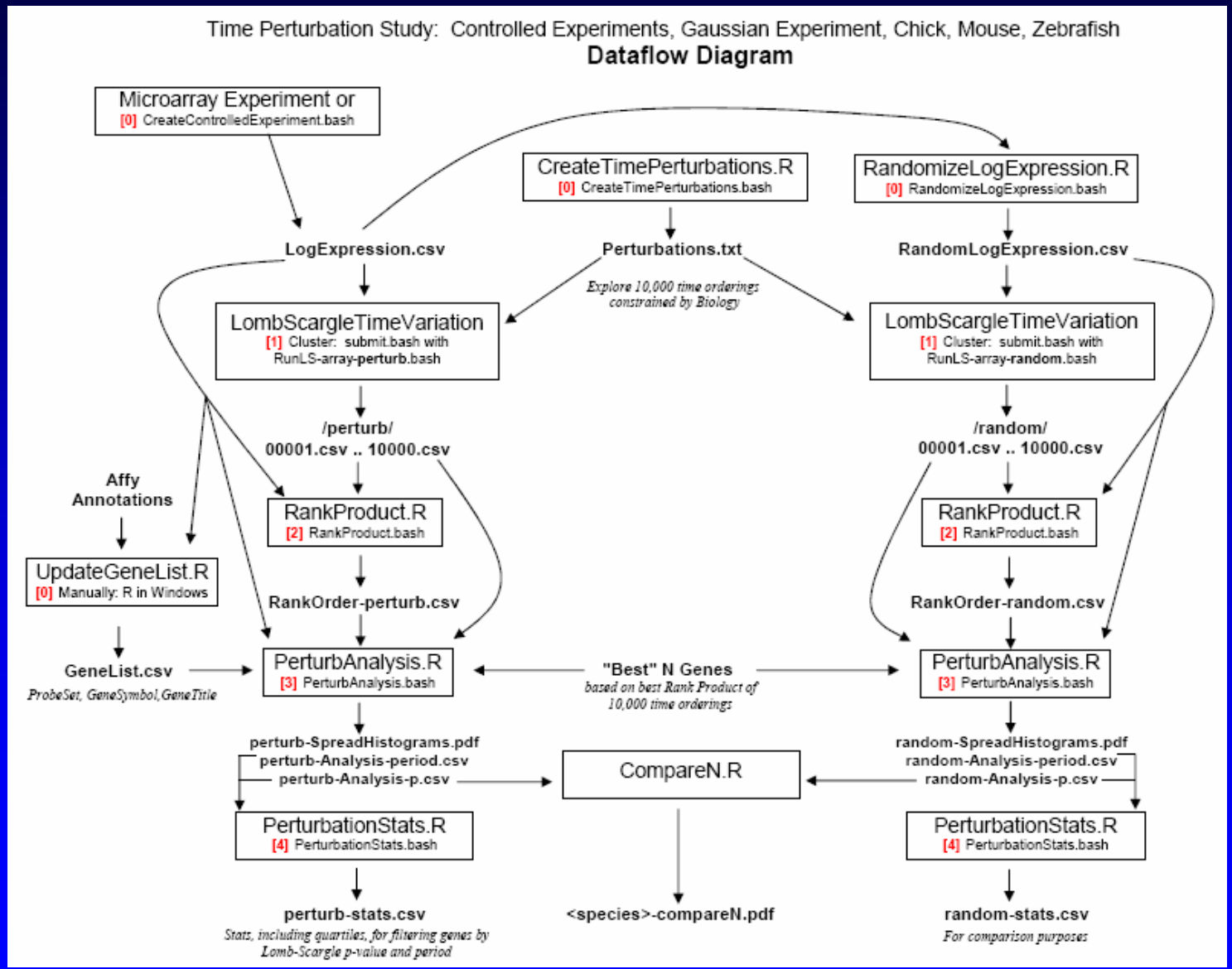


But is Gaussian noise the appropriate comparison for real experiments?

"Time Perturbation" Studies of Somitogenesis Microarray Time Series

- Purpose
- Background Concepts
- Microarray Analysis Overview
- Numerical Experiments
- Somitogenesis Experiments
- Summary

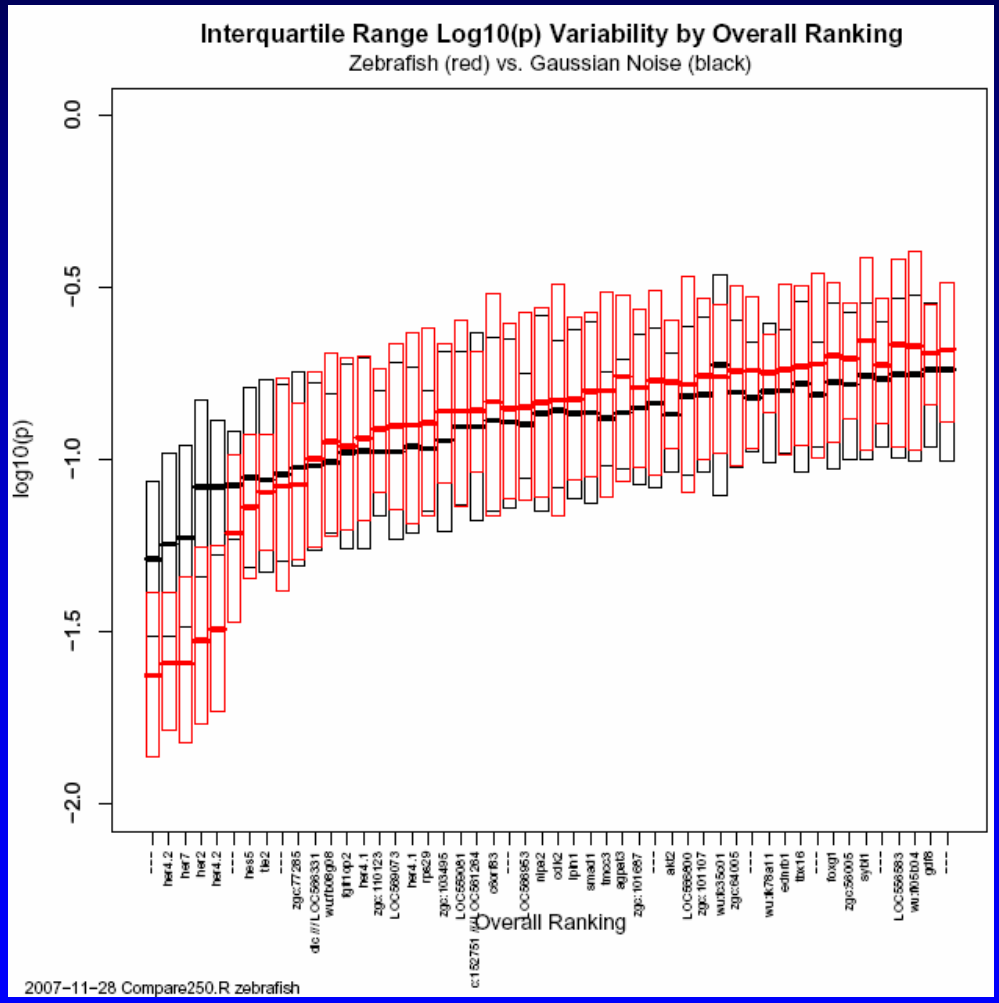
Time Perturbation Somitogenesis Studies



Time Perturbation Somitogenesis Studies

- Zebrafish
- Mouse
- Chick

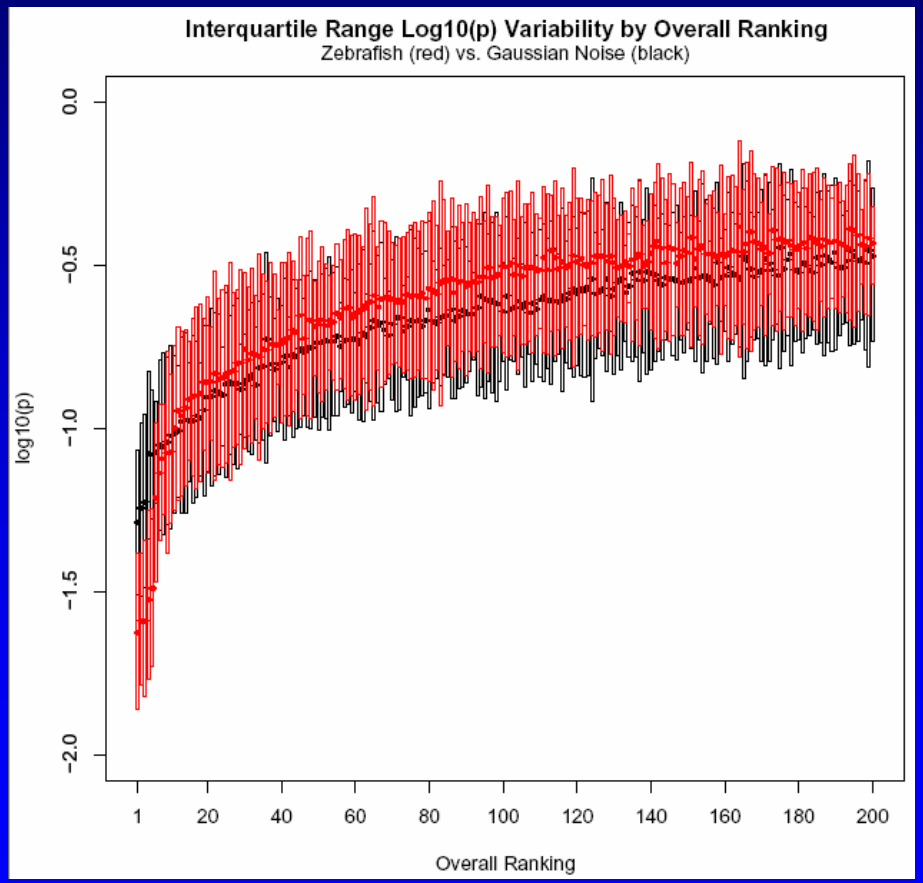
Zebrafish



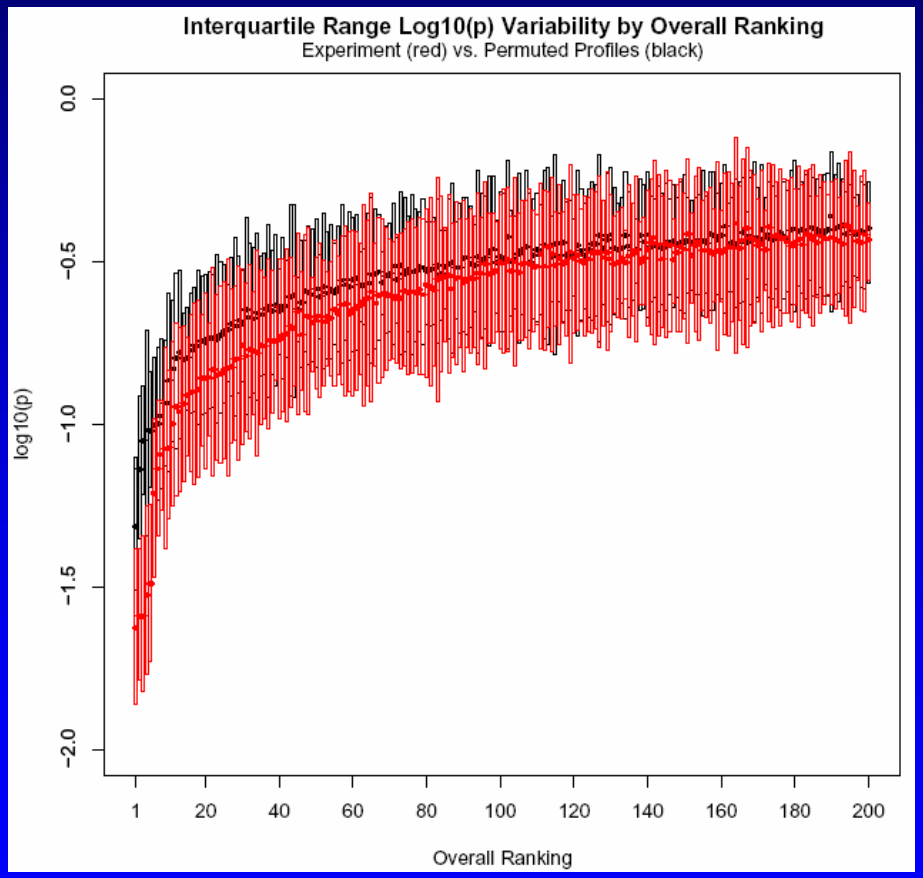
Is Gaussian assumption valid? Is Gaussian "too" random?

Zebrafish

Gaussian Noise



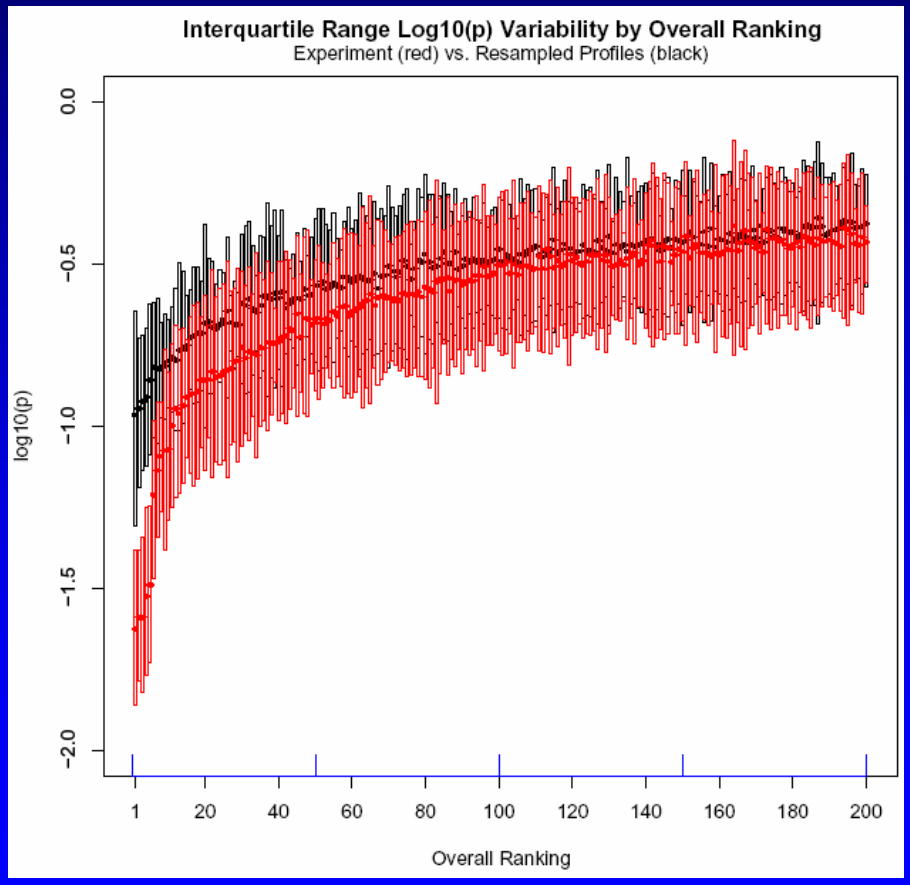
Permuted GeneChip Noise



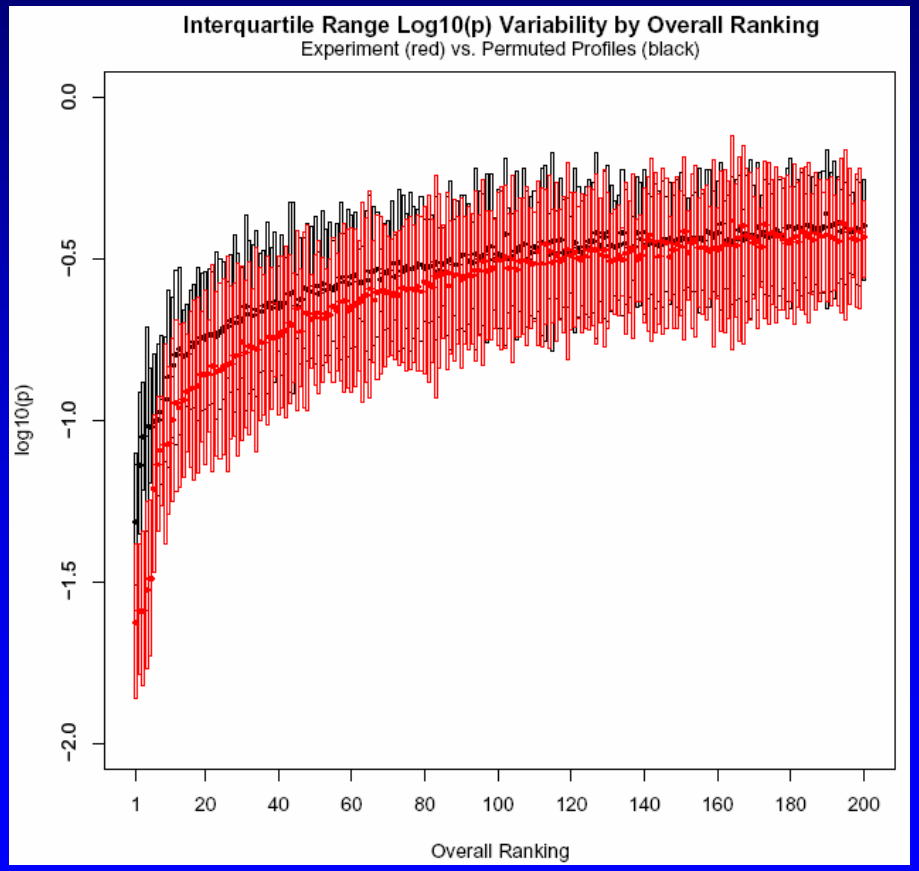
Is Gaussian assumption valid? Is Gaussian "too" random?

Zebrafish

Resampled GeneChip Noise



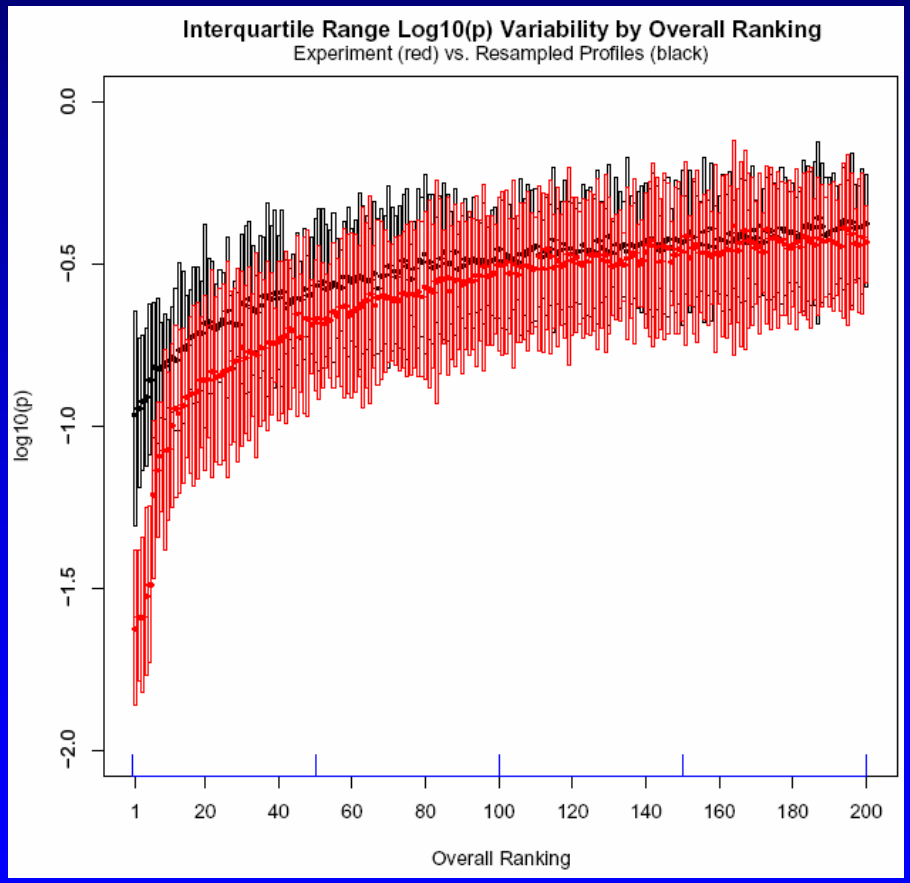
Permuted GeneChip Noise



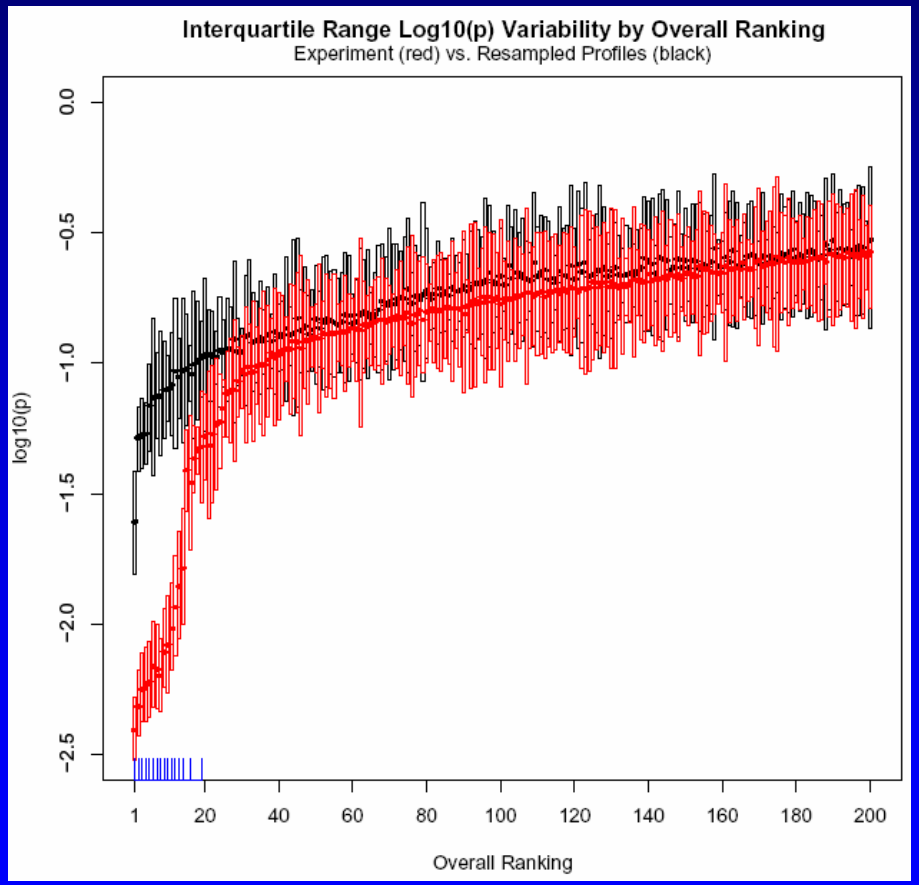
Resampled GeneChips seem to be best "noise" reference

Zebrafish

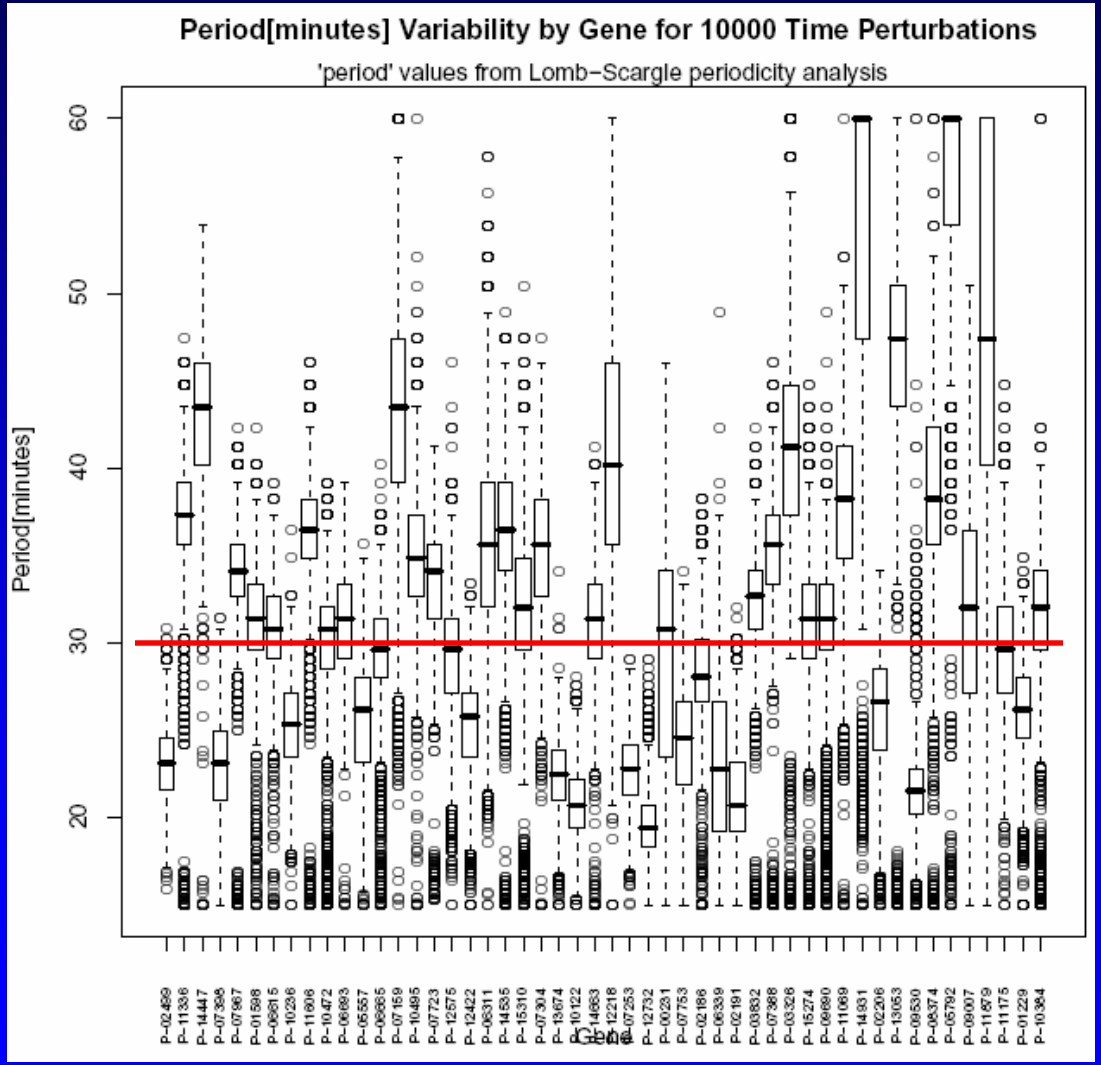
Actual Experiment



Simulated Experiment (16)



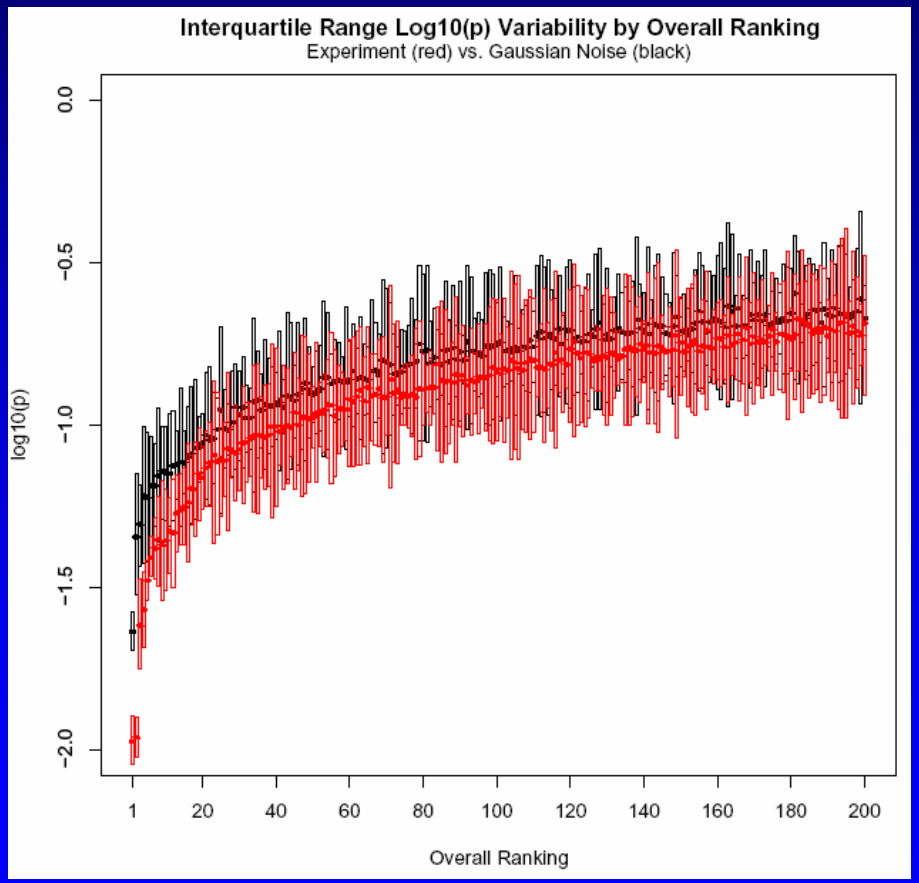
Zebrafish



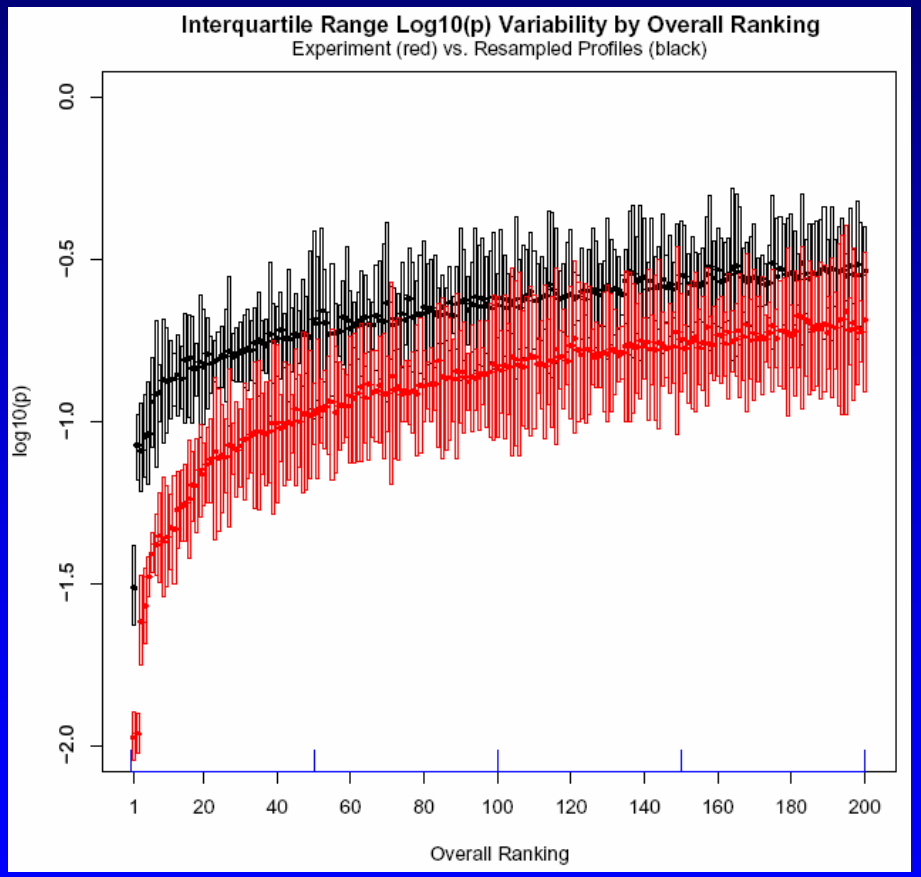
Zebrafish somite period is assumed to be 30 minutes

Mouse (17 points)

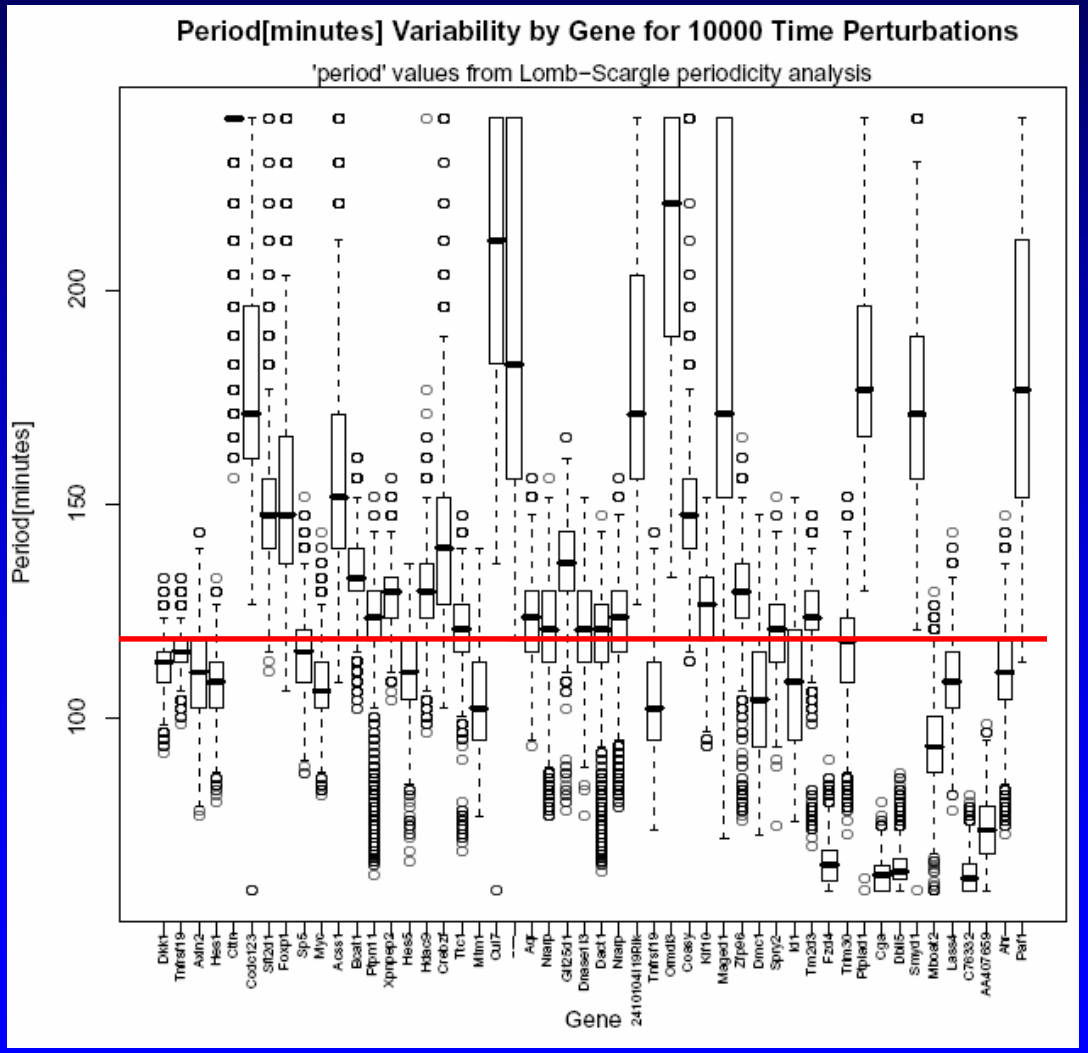
Gaussian Noise



Resampled GeneChip Noise



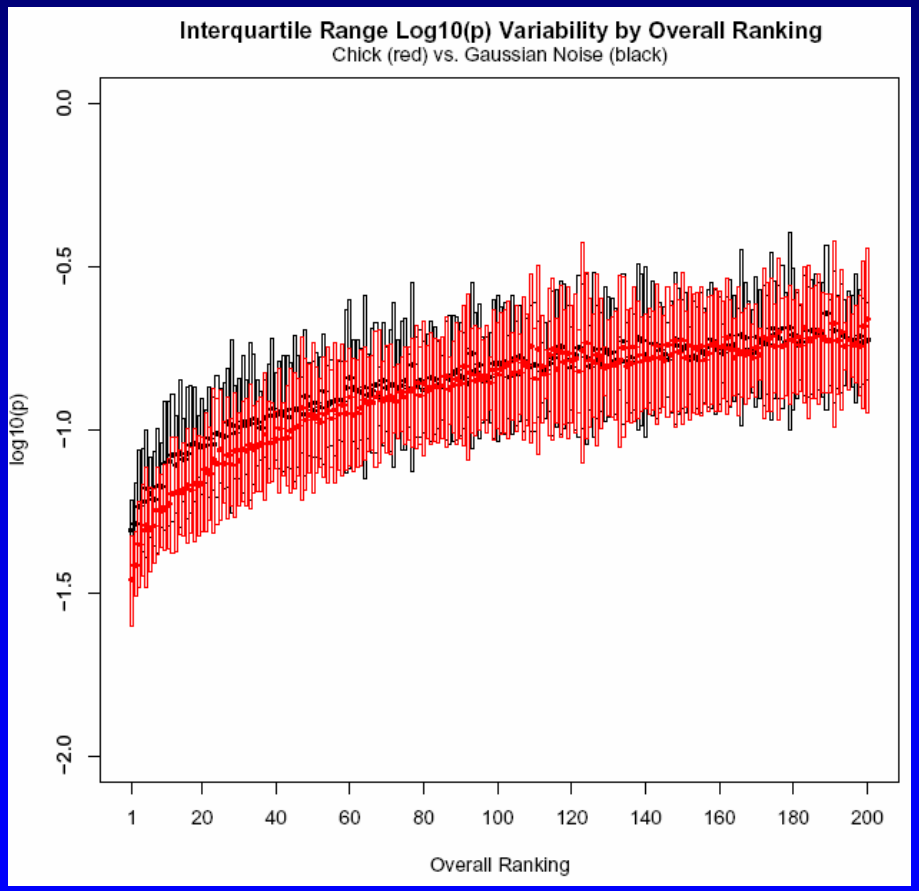
Mouse (17 points)



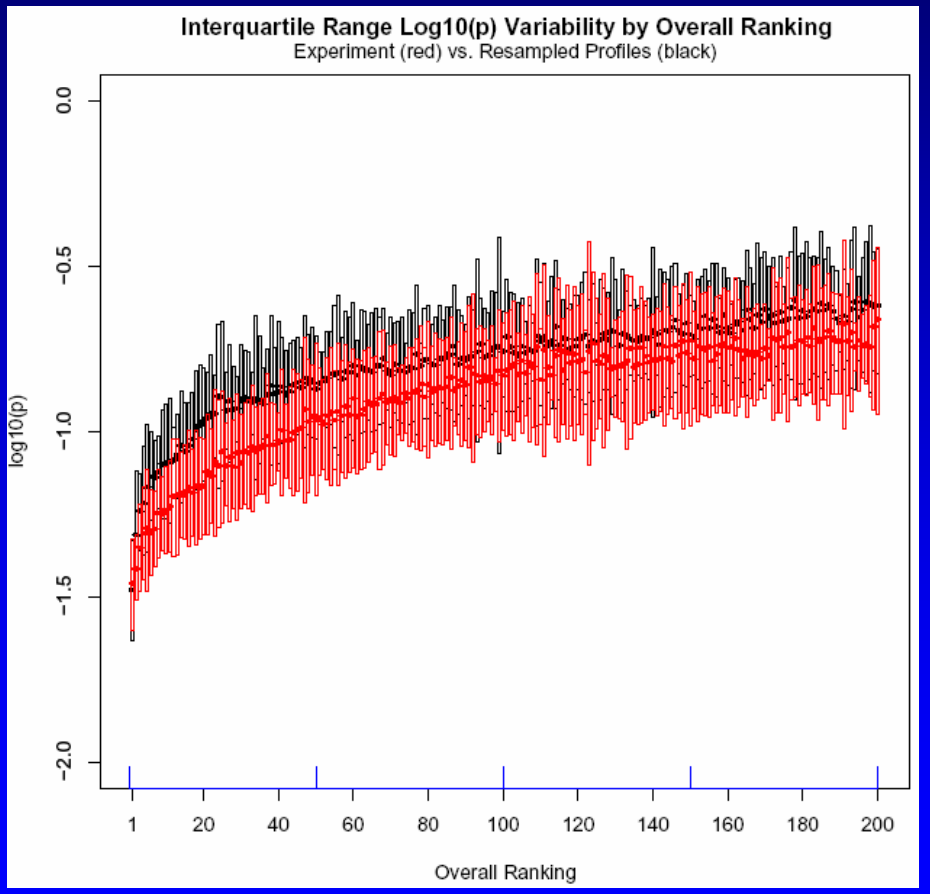
Mouse somite period is assumed to be 120 minutes

Chick

Gaussian Noise



Resampled GeneChip Noise



Summary

- Common "time perturbation" methodolgy seems to working for all species
- Simulated experiments can give insight into actual experiments
- Additional simulated experiments needed to understand mouse and chick results

Acknowledgements

Pourquié Lab
Olivier Pourquié
Mary-Lee Dequéant (Mouse)
Aurelie Krol (Chick)
Daniela Röllig (Zebrafish)
[Max Planck Institute]

Bioinformatics
Arcady Mushegian
Jie Chen
[University of Missouri, Kansas City]