GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

30 November 2006

Lecture 26: Time Series

- Defining the Question
- Synchronizing
- Sampling
- Defining Test Statistics
- Grouping
- More on Phase Shifts
- Open Questions

Why Time Series Studies?

Trying to assess periodicity (cell cycle, circadian rhythm)

Trying to assess response/causality (drug treatment, gene knockouts)

Trying to infer networks (modelling of ensembles)

These questions are all big and important. However, there are many fewer papers dealing with this type of data, in large part because deciding how to get, measure, and use time course data is often *hard.*

Cycles, always cycles

Canonical example here: Spellman et al (Mol Biol Cell, 1998), looking at cell cycle stuff in yeast.

A more recent example: Bozdech et al (PLoS Biology, 2003), cell cycle behavior of malaria in blood.

Understanding which genes are active in which parts of the cycle may be useful for guiding interventive therapies for malaria. (heal the yeast?)

Lining up the cells

One of the initial problems in looking at cyclic behavior is simply that most groups of cells are at equilibrium – some are at the beginning, middle, and end of the cell cycle, and when gene expression levels from this mixture are examined, the cyclic variation will not be visible.

Since we need RNA from a large number of cells in order to measure the system, we need to either extract cells at a specific part of the cell cycle, or coerce the population into lining up.

Synchronization

The Spellman et al study used both extraction and coercion methods.

In terms of extraction, the cells were filtered by size (elutriation), and the smaller ones were selected. Ideally these would be young cells just embarking on the G1 phase.

Synchronization

In terms of coercion, two methods were used:

1. α factor arrest

2. temperature-sensitive knockout arrest (cdc15; a related study by Cho et al used cdc28)

The hope was that the use of multiple methods would enable some type of redundancy; genes found to be cycling in all of the conditions would be interesting.

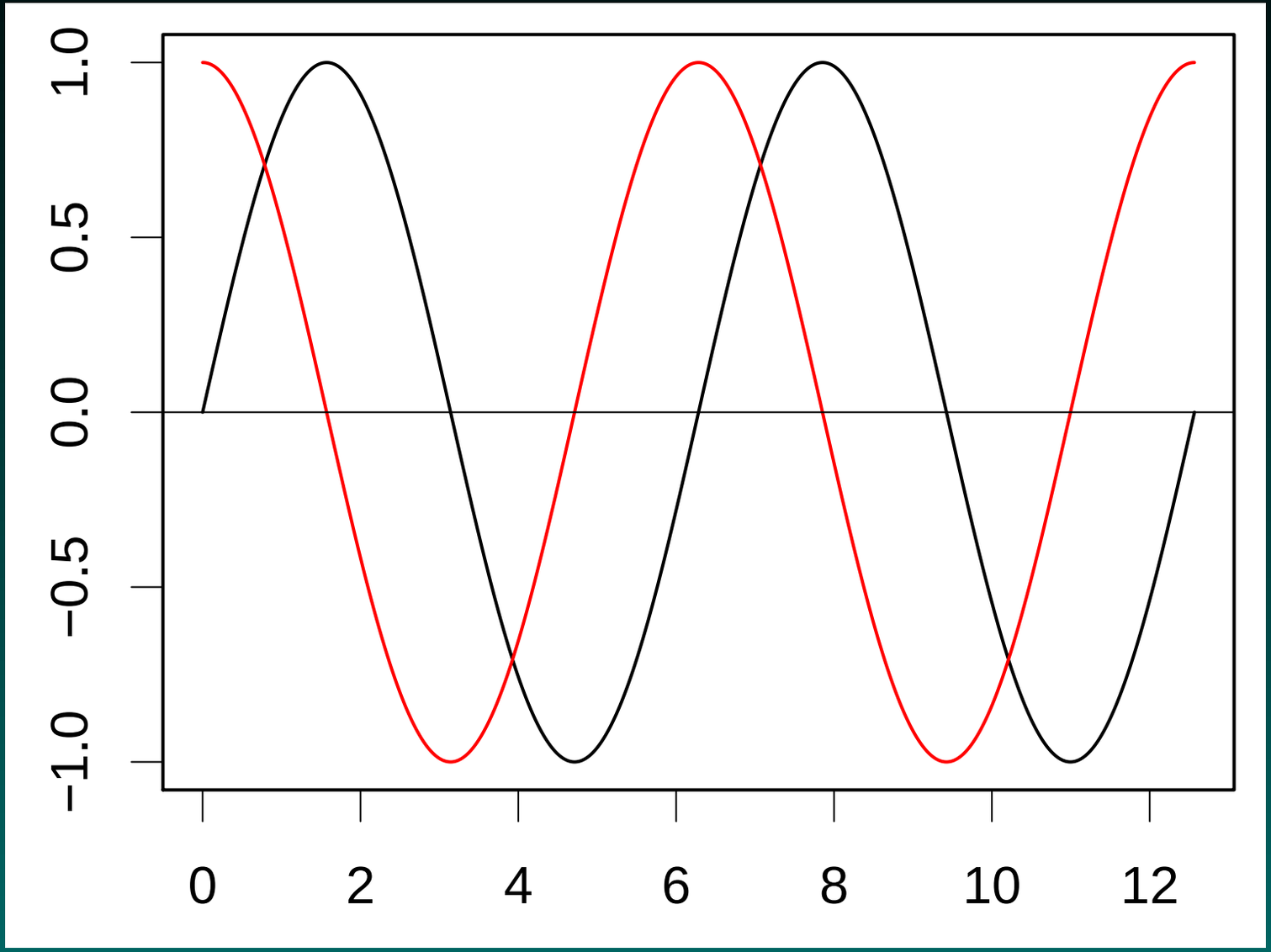
They're lined up. Now what?

Once we have cells all gathered at the same start line, and we let them surge forth, we are faced with two questions:

How long should we measure?

How many times should we measure?

Simple periodic behavior



How long?

In terms of how long, there are some limits imposed by diffusion – even though they started out lined up, the cells will not stay lined up with respect to their position in the cell cycle for very long. It is not clear to me whether this problem is best defined in terms of numbers of hours, numbers of cell cycles, or both.

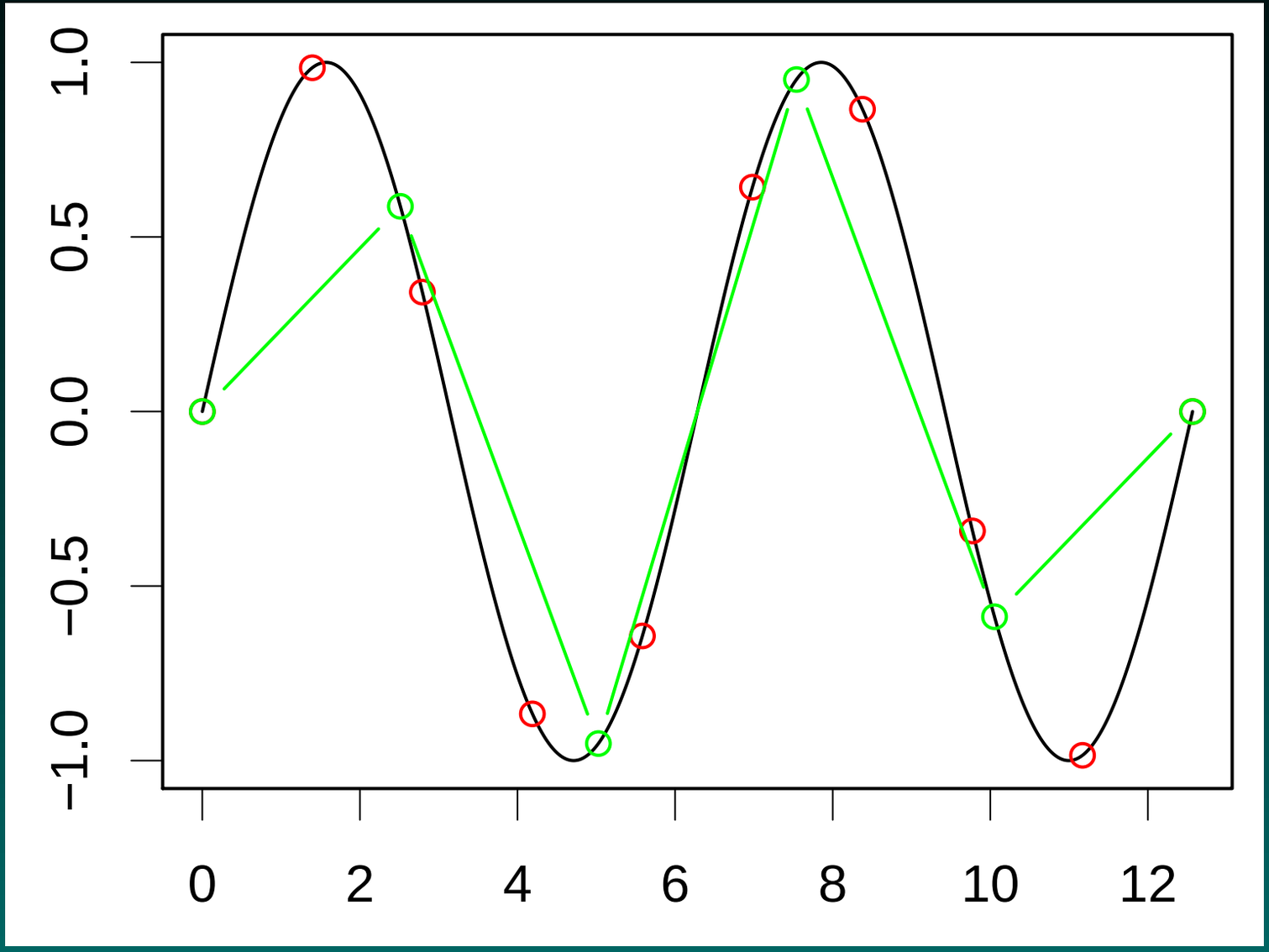
At a minimum, however, we would like to keep track of the cells through at least one full cycle, and preferably two or more. If we don't do this, we haven't given the genes time to go up and come back down, to establish a cycle.

How many?

In terms of how many times we should measure (the sampling rate), the tradeoff is one of being able to pick up fine details against that of the cost of the arrays. The first says make as many measurements as you can, and the latter says make as few.

Aside from the issue of fineness of detail, however, there is the question of how do we decide if we have found something real. In most cases this entails comparing results with a null distribution, and the behavior of the null distribution can be estimated by permutation of the values.

How many?



How many?

How many genes do we have?

How many permutations of the data are possible?

For the yeast data, there are about 6000 genes. Starting with 10 observations, $6000/10! = 0.0017$, which is the most extreme the p-values can get after adjusting for multiple testing in a conservative fashion. As a ballpark estimate, then, I'd look for at least 10, possibly with a few more to allow for the possibility of garbage measurements. We may revise this based on the actual statistic we use, but the general rule is that we want to be able to say that an effect is "significant".

How many?

There is one additional thing that I would strive for, and that would be to have measurements in two or more cycles, with measurements made at the same relative points in the cycle (to the extent possible).

This way we can superimpose points from different cycles to check for visual overlap.

How to measure?

How do we decide which genes are cyclic?

Well, if an expression pattern is cyclic, that means that it repeats over time with a specific period. Thus, if we plot the intensities from cycle 1 against the corresponding intensities from cycle 2, we should see a straight line (this type of matching and plotting is why I would like to see measurements of multiple cycles).

How to measure?

The above approach (plotting) does require that we have a moderately good idea of what the period is before we begin to collect the array samples. This is in general a good property to plan for in the design.

Note that this type of behavior (plot agreement) does not depend on the shape of the patterns. In that sense it is an omnibus test. Alternatively, we can look for cycles of a specific pattern.

Fitting patterns

The most familiar cyclic patterns are sinusoids, so we can talk about fitting the data to a sinusoidal pattern for each gene.

Fitting sinusoids is the approach taken by Spellman et al., as well as by others (eg, Jim Booth et al., from U Fl). The implicit model is

$$\mu(t) = b_0 + b_1 \cos(2\pi t/T + W) + b_2 \sin(2\pi t/T + W),$$

though the means were subtracted off in the Spellman et al analysis.

Model thinking

In looking at the model, we note that there appear to be 5 unknowns: b_0, b_1, b_2, T, and W. However, only the b_i values are gene-specific. T and W are assumed to be common for all of the genes in a given experiment, though not across experiments. Thus, T and W would have to be estimated anew for the α-factor arrest, cdc15-arrest, cdc28-arrest, and elutriation experiments.

A phase difficulty

Now, as it happens, we can't estimate the value of W from a single experiment (or gene) alone. The reason is that sin and cos trade off too well. for any two values W_1 and W_2,

we can write

$$\mu(t) = b_0 + b_1 \cos(2\pi t/T + W_1) + b_2 \sin(2\pi t/T + W_1),$$

or equivalently

$$\mu(t) = b_0 + b_1^* \cos(2\pi t/T + W_2) + b_2^* \sin(2\pi t/T + W_2).$$

A phase difficulty

The change in W has migrated to the b_1 and b_2 coefficient values, but the predicted values are *exactly the same*. Thus, within a given experiment, we assume that $W = 0$ to make the equations look simpler.

A phase difficulty

In general, differences in W values between experiments indicate that the experiments start at different points in the cell cycle, and these differences need to be corrected for in order to "line things up" before we can check for redundancy.

Within an experiment, we can look at differences in phase to give us some idea of where to position the genes relative to each other in the cell cycle.

Fitting the rest

So, for a given experiment and gene, how do we estimate the parameters?

Start with the simplified form of the model, with actual observed values plugged in:

$$y_i = b_0 + b_1 \cos(2\pi t_i/T) + b_2 \sin(2\pi t_i/T) + \epsilon_i$$

If T is known, then this is a simple linear regression equation which can be solved by least squares.

What is T? Brute force

A good value of T is one that reduces the residual sums of squares across all genes.

So, start with a guess as to the value of T derived from observing the cultures (when do new buds form), and try values of T in a grid about that value, refitting all of the genes each time.

The data

- α-factor: $n = 18$; 7 minute intervals; 2 periods
- cdc15: $n = 24$; 10 or 20 minute intervals, 2.5 periods
- cdc28: $n = 17$; 10 minute intervals (Affy chips) 2 periods
- elutriation: $n = 14$, 30 minute intervals

http://genome-www.stanford.edu/cellcycle/data/rawdata/

I grabbed the data from the CAMDA 2000 web site.

The data

The same cDNA arrays were used for the three experiments done at Stanford. The reference was a sample of the yeast cell culture before synchronization (ideally a steady state profile).

Log ratios are reported.

The data, slightly fit

Expt	n_{genes}	Known	T (min)	W
α-factor	6019	103	62	0
cdc15	5413	90	115	0.081π
cdc28	6068	103	85	0.170π
elutriation	6074	103	603	0.357π

There are a few things to note here.

Data oddities 1

The number of genes that are "usable" varies from one experiment to the next due to the loss of some observations due to crud. Booth et al (the source of the above table) retained genes only if 80%+ of the data was available.

The "known" genes refers to a subset of 104 genes that had been previously linked to cell cycle behavior by other means (Spellman et al).

Data oddities 2

The cycling times are different for the different experiments. This is due in part to the different conditions required for the arrests, but also to the fact that the media in which they were grown afterwards was not quite the same (the level of nutrients varied, for one thing).

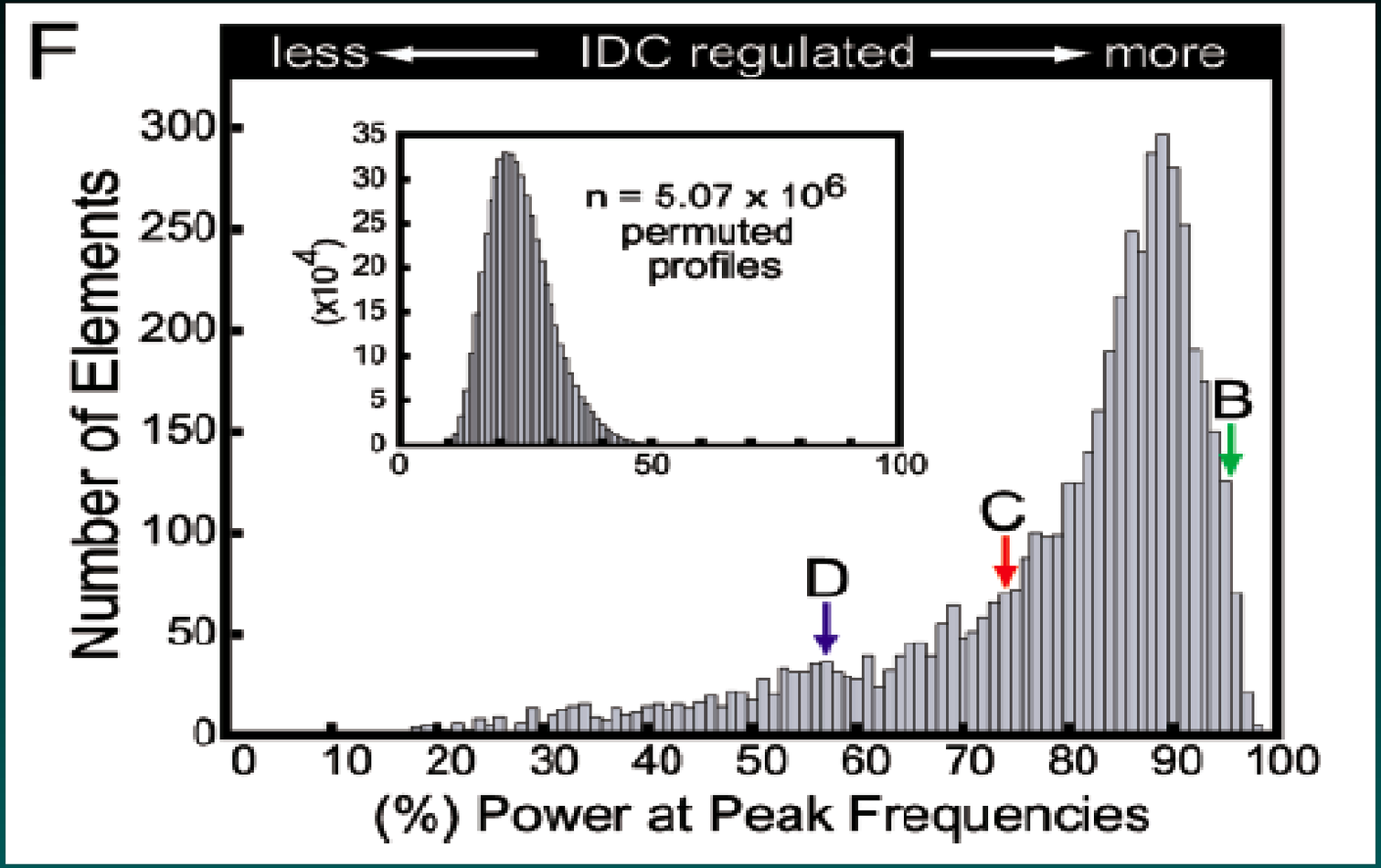
Data oddities 3

The period for the elutriation approach is much longer than for the other methods, and the number of cycles completed is much shorter (point estimates of these are 2, 2.5, 2, and 0.7, in the order above). This difference made it much harder to compare the elutriation data with the others, so this dataset was excluded by both Spellman et al and Booth et al. It should also be noted that the estimate for the elutriation period length given above is likely too high. Booth et al note that it is hard to estimate the period well from a single cycle or less.

Which genes are interesting?

In order to test significance, we are interested in whether the periodic component is large. We can define this by performing an F-test as to the significance of the model fit for each gene. Given a test statistic value, the empirical p-value can be assessed through permutation.

Which genes are interesting?



Bozdech et al, Fig 1F

Which genes are interesting? (Spellman)

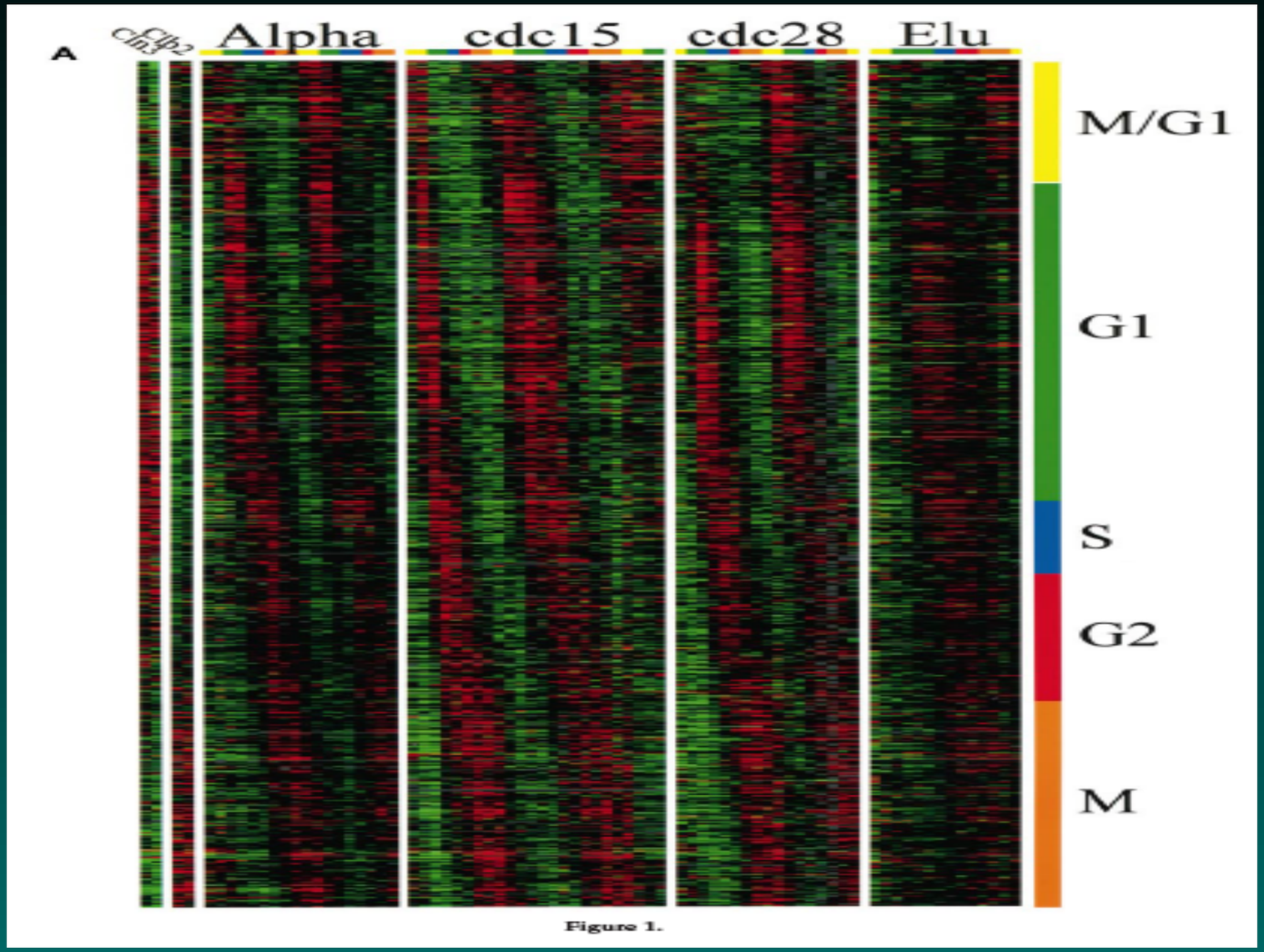


Figure 1.

Which genes are interesting? (Bozdech)

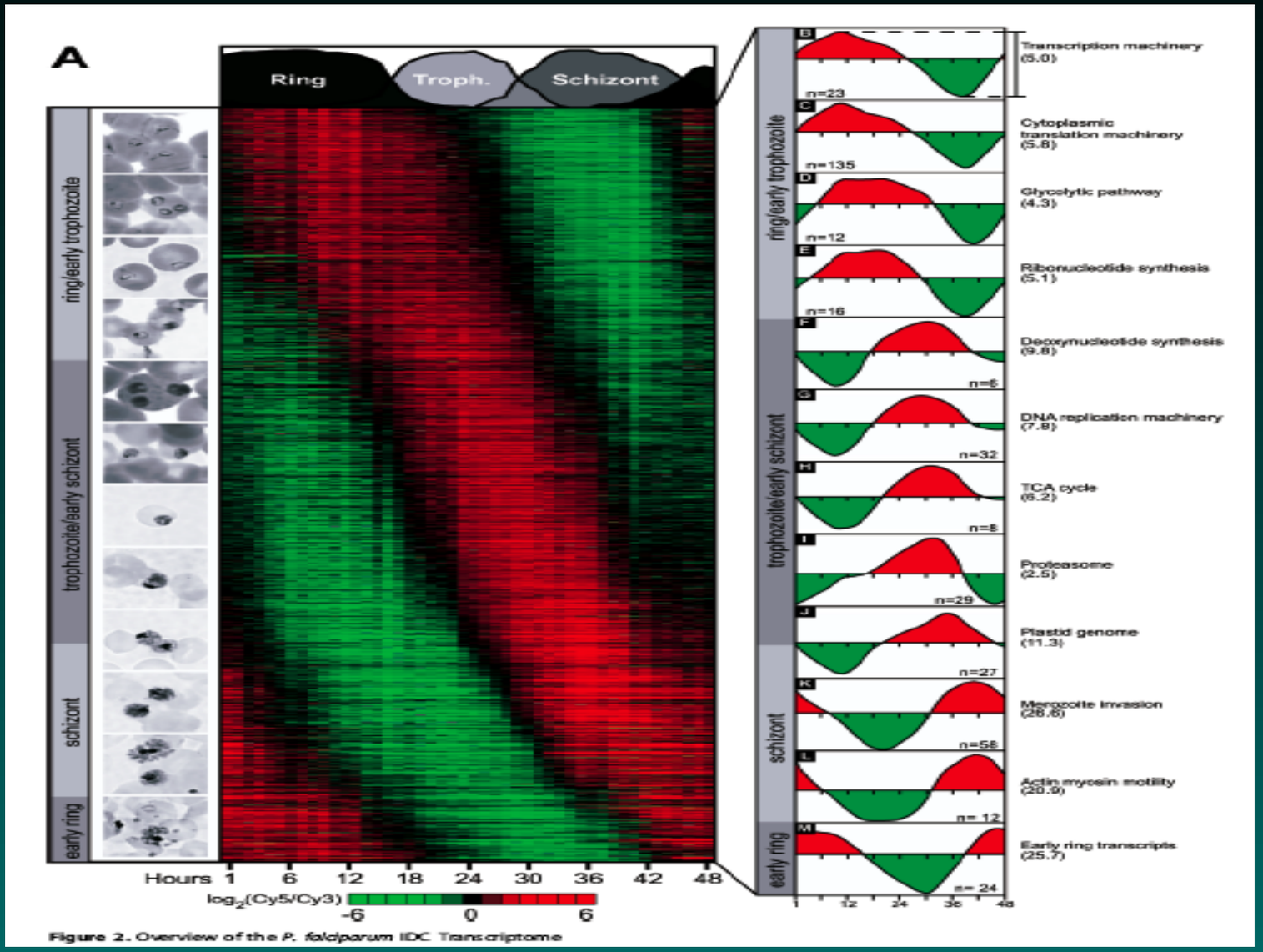


Figure 2. Overview of the *P. falciparum* IDC Transcriptome

Meta-analysis

In order to combine results from the different synchronization methods, Booth et al rely on methods for combining p-values, noting that if P_k denotes the p-value from the k^{th} experiment, then

$$C^2 = -2 \sum_{k=1}^{K} \ln P_k$$

has a χ^2_{2K} distribution, and the analytic results can be checked here.

The results

Spellman et al found about 800 genes to be cyclically regulated, whereas Booth et al found about 1170. Part of the increase is likely due to the more precise tests that Booth et al employed.

There are still some questions. Shedden and Cooper (NAR, 2002) suggest that many of the differences observed are likely due to the shock of starting from arrest, and suggest that the size-filtration method is most different because it does not include this artifact. With this in mind, they focus on the elutriation experiment and find a different list.

Are we happy?

Questions:

how could we use different periodic shapes?

is the method of combining p-values used by Booth et al really that efficient?

coupling microarray measurements with other biology – are the cyclic genes spatially grouped within a few chromosomes?

Other time course data

In other experiments, we have looked at the response of samples or cell lines (most commonly the latter) to specific stimuli or growth conditions.

In one case, we looked at the reaction to the addition of a chemotherapy agent.

In another, we tried to measure the growth and development of healthy and diseased cells.

The good and the bad

These latter examples do have one *advantage* relative to the cycling experiments – we are interested in the response or the steady state ensemble, and thus we do not need to perform a separate synchronization step (though that could be an extremely interesting separate factor to control).

However, these "response" experiments have a different and severe *disadvantage* relative to the cycling experiments, in that it is rarely possible to specify the shape of the desired response profile.

The good and the bad

Is it interesting if the expression pattern increases monotonically over time?

The good and the bad

Is it interesting if the expression pattern increases monotonically over time?

decreases monotonically?

The good and the bad

Is it interesting if the expression pattern increases monotonically over time?

decreases monotonically?

goes up sharply at the beginning, and then drops off?

The good and the bad

Is it interesting if the expression pattern increases monotonically over time?

decreases monotonically?

goes up sharply at the beginning, and then drops off?

When we have asked these questions, the standard response has been "they're all important". Unfortunately, omnibus tests (looking for anything) are (a) harder to devise and (b) less powerful than more precisely crafted tools.

Our current punt

In this case, we have tended to say "let's work with a class that's fairly simple to explain", which we have defined to be "changing monotonically, possibly after an initial delay". Building this delay into the model takes us to the technique of isotonic regression.

Do we understand the context?

When we are fitting a regression, there are further issues:

How should time be used?

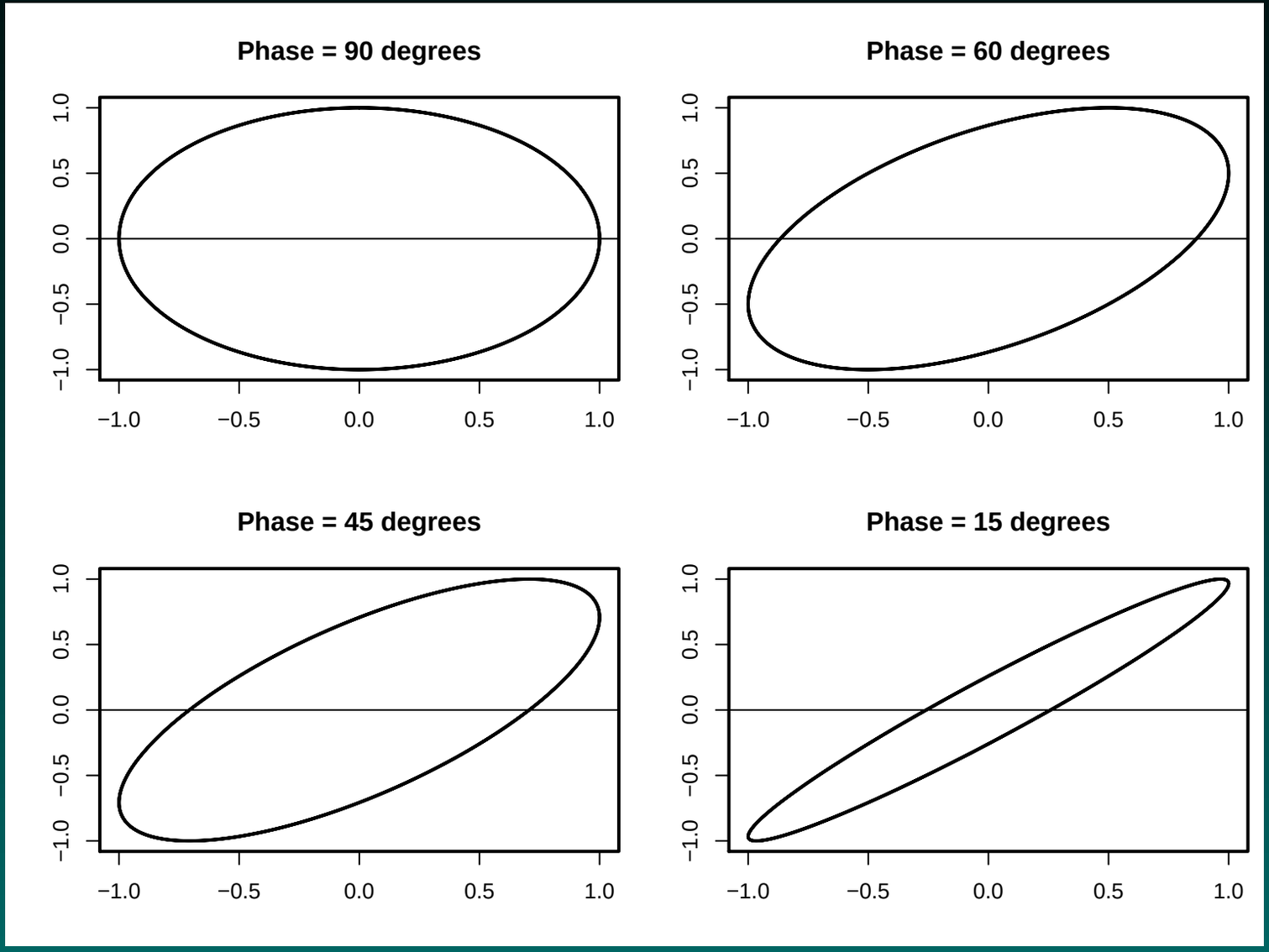
As a continuous covariate? ordinal?

Eg, there can be issues with treatment of cell lines – qualitative shifts between progress until confluence, and behavior afterwards

Choosing the times

Finally, the choice of which time points to use here is far from clear. This is the most challenging area I see, as it requires knowledge of both the likely biology and the ease of mathematical modelling.

More on Phase Shifts



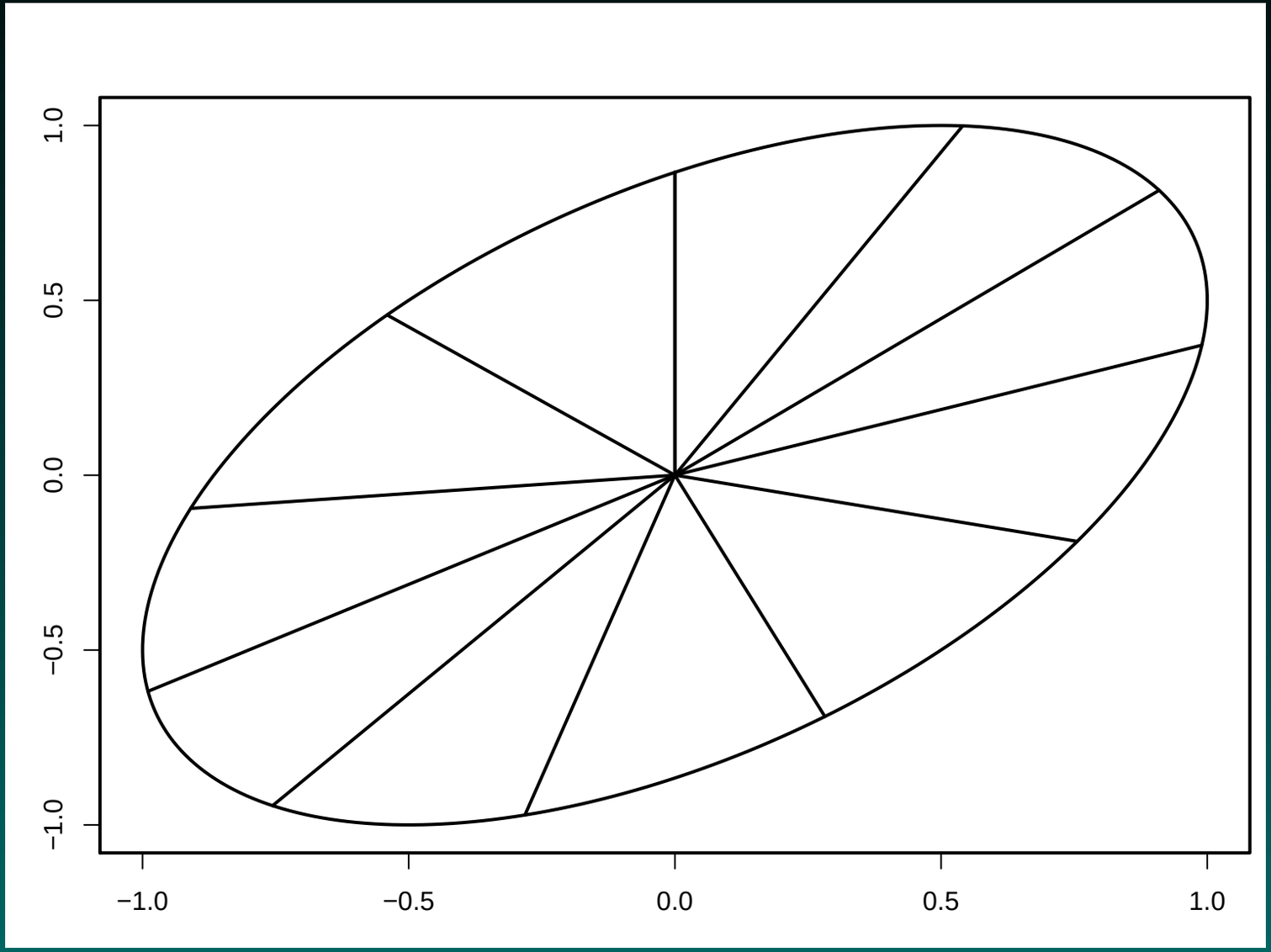
Speculation

If two genes are both periodic, and in phase, we can detect that using the usual correlation approaches, since plotting one against the other will produce a straight line.

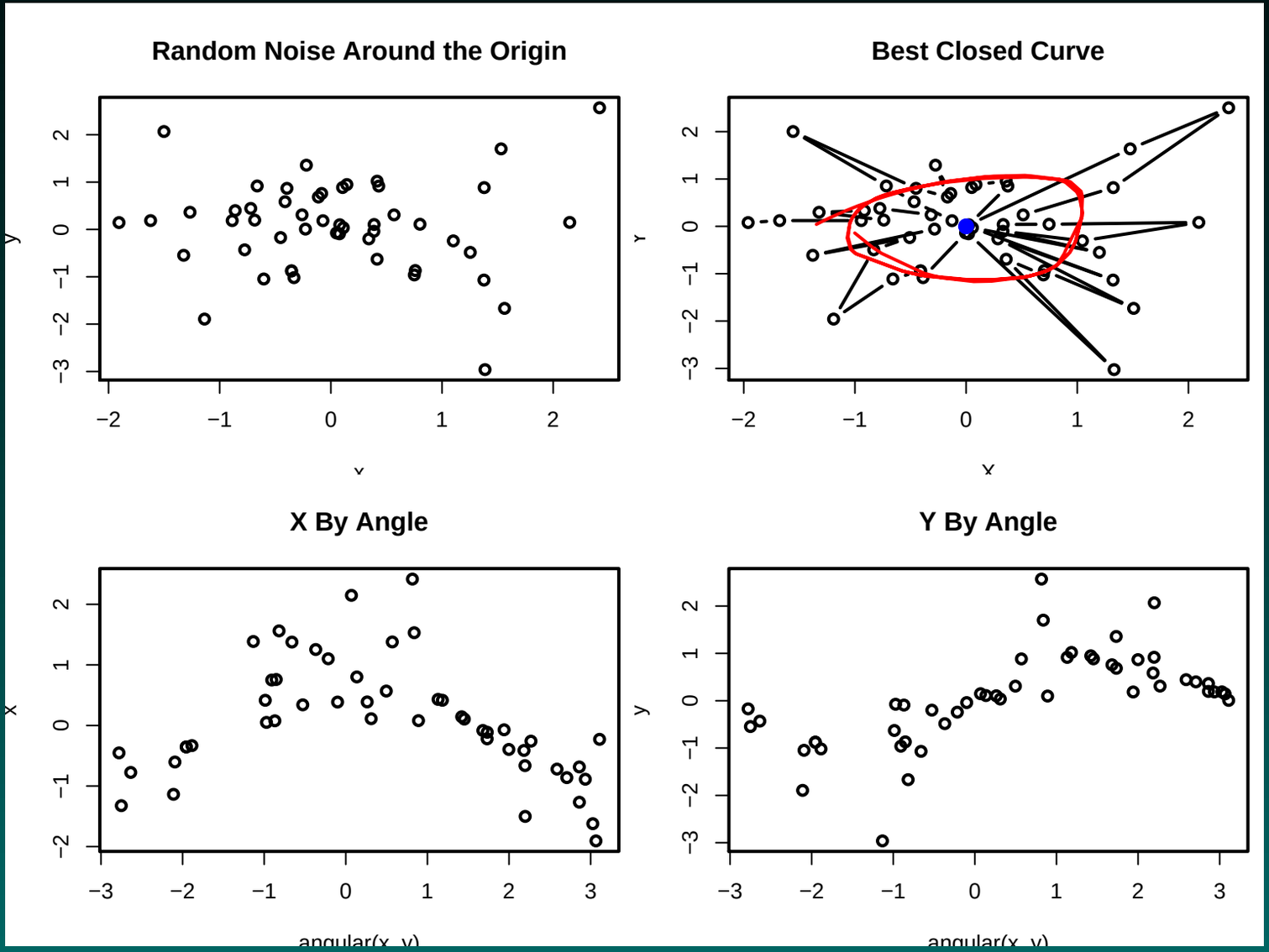
If they are periodic but out of phase, correlation will not work.

Can we detect "elliptical" patterns in plots of pairs of genes, even without the accompanying time information? (The real question is: can we find periodic phenomena in array studeis tghat are not time courses, so we do not have any way to order things temporally.)

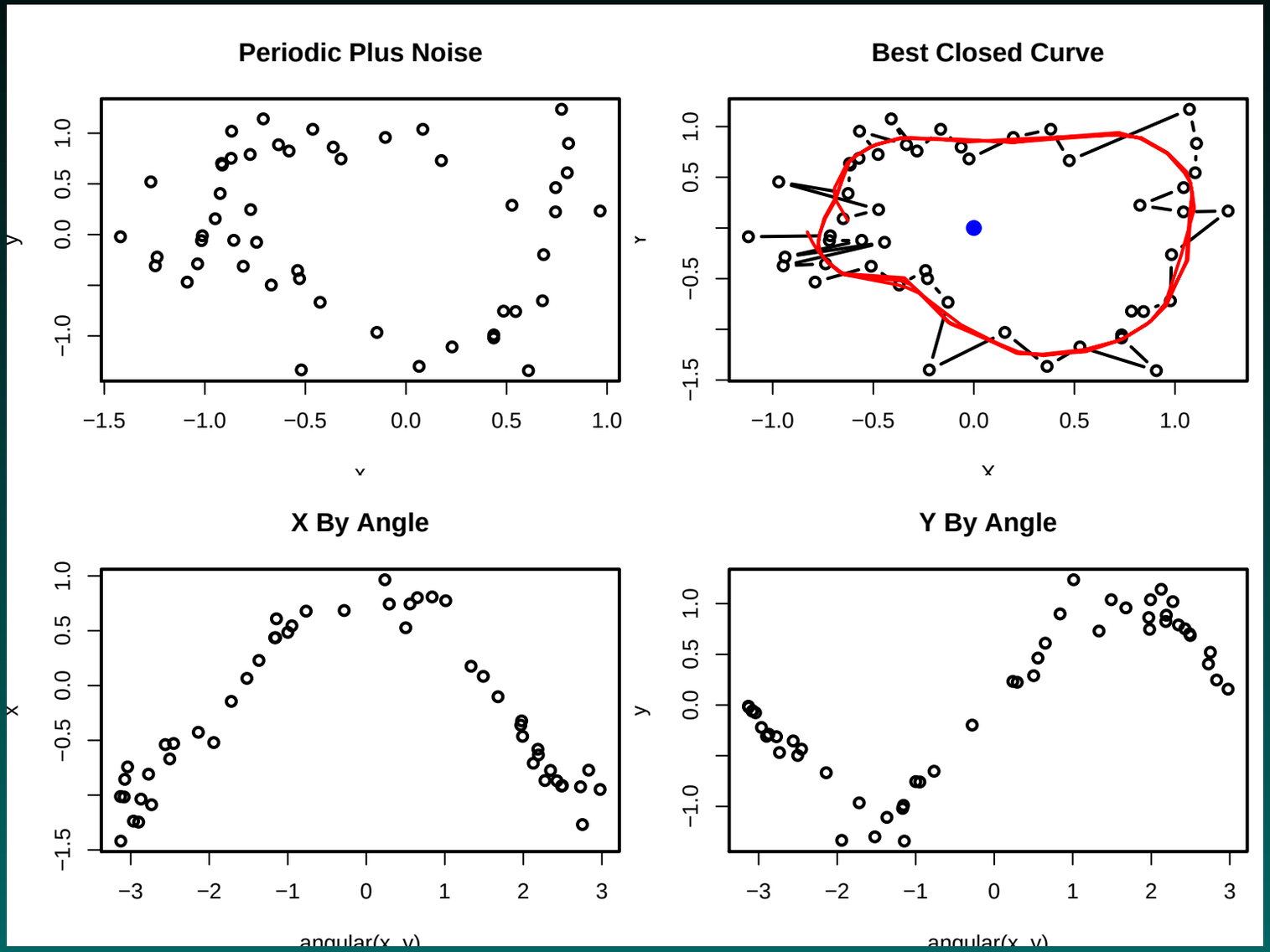
Radial Movement



Trying to Fit Closed Curves



Trying to Fit Closed Curves



Tentative Statistic

Idea: Given vectors x and y representing paired expression levels of two genes, first mean center and standardize both vectors.

Next, order the points by the angle they make with the origin.

Compute the average distance A of the points from the origin.

Compute the average distance D between consecutive points.

Use the ratio D/A to decide if there is something that looks roughly circular.

Results in the Examples

With random uncorrelated noise:

```
> summary(w)
```

```
[1] 0.8237164
```

With underlying periodic behavior:

```
> summary(w2)
```

```
[1] 0.3322947
```


Open Questions

- Tissue Arrays
- CGH Arrays
- SNP chips
- protein arrays
- data integration
- etc

General Themes

How is the data measured?

Processed?

Normalized?

Stored? (Have we kept track of things in an organized fashion?)

General Themes

Multiple Testing

Cross-Validation

Clustering

Experimental Design

Tissue Arrays

Not really high-throughput in the same sense as the others; instead of making hundreds of measurements on a single sample, we make a single measurement on hundreds of samples.

How do we define this measurement?

Integrated Optical Density

Is this the right measurement to make?

CGH Arrays

DNA, not RNA

Still worried about normalization

Scale of changes often much smaller on a log scale than with expression arrays

New feature: sequence/position information

How do we check for contiguous blocks?

What type of distance should we use?

SNP Chips

Again, DNA

Affy here

How do we define a probeset?

How do we combine probes?

Splice Variant Chips

Affy again

How do we define a gene?

Is this the natural measure a biological unit? This wording is vague. How should we sharpen the definition to make clear what we're looking for?

Protein Arrays

Smaller number of proteins being measured

Can we do absolute quantification?

Multiple spottings and dilution series – how can this be exploited?

Mass Spectrometry

Hundreds of proteins, not thousands of genes

How do we calibrate the data?

How do we quantify the data?

How do we design the studies?

Data Integration

How can we combine expression data with DNA data?

How can we combine expression data with protein data?

Lung cancer and FUS1

Some Sample Sizes

How many samples do you need?

What type of an effect are you looking for?

How big do you expect it to be?