

(Meta) Elements of transcriptome profiling / functional genomics

October 04, 2005

Alvin T. Kho
Children's Hospital Boston
Dana-Farber Cancer Institute

Functional Genomics

– the overall enterprise of de-constructing the genome to assign biological functions to, or uncover interactions between gene subsets.

Transcriptomics

– as above but focusing on subset of the genome that is “expressed”

System specific

Lecture Outline

- Review: Measuring the transcriptome. Microarrays
- Assumptions and questions in transcriptomics
 - Granularity of questions
 - Paradigm shift relative to traditional biology
 - Prototypical experiment designs
- Workflow in microarray-driven experimentation
- Analysis and modeling of transcriptome data,
 - Mathematical formulation of problem
 - “Correcting” noise and measurement variation / bias
 - Uncovering coherent geometries and dominant variance structures intrinsic to data
 - Likelihood of coherent structures / math results arising from chance
 - Squaring math results with *a priori* biological knowledge. Figure of merit
- User-friendly References

Review: Measuring the transcriptome

- Assay platform = microarray
 - Measures $\sim 10^{3-4}$ mRNA species levels at once
 - Principle: Nucleic acid complementarity – binding
 - Technical assumption I: Uniform RNA degradation and hybridization rate – independent of RNA species
 - Technical assumption II: Fluorescence intensity is proportional to expression level – independent of RNA species

Figures removed due to copyright reasons.

Please see:

Pevsner, Jonathan. *Bioinformatics and Functional Genomics*. Hoboken, NJ: Wiley-Liss, Inc., 2003.
ISBN: 0471210048.

Transcriptomics: Assumptions

- Central dogma holds
- Phenomenological (phenotypic, cellular) events of interest necessarily engage transcriptomic mechanisms, or are reflected in the transcriptome

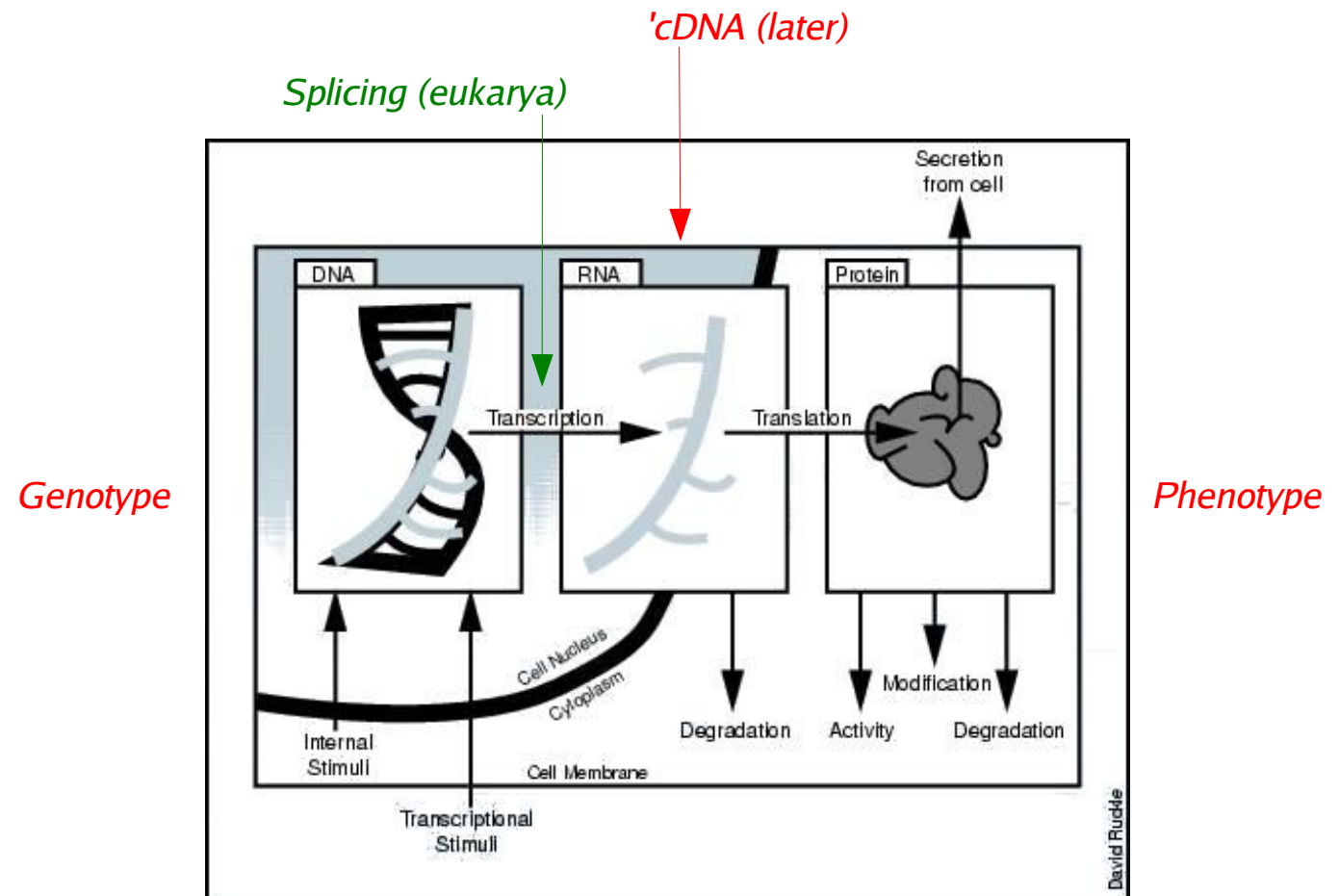


Figure 1.7 in Isaac S. Kohane, Alvin T. Kho, and Atul J. Butte. *From Microarrays for an Integrative Genomics*. Cambridge, MA: MIT Press, 2003, p. 28. ISBN: 026211271X. Courtesy of MIT Press. Copyright 2003. Used with permission.

Transcriptomics: Questions

- Granularity of questions – *3 molecular scales or levels*
 - **Single:** Identify individual molecules associated with (possibly underwriting) a biological phenomenon
 - **Network:** Identify molecular networks associated with a biological phenomenon
 - **System:** Transcriptomic / global state or characterization of a biological system
- Paradigm shift relative traditional biology
 - **Traditional:** Whole = Sum of its parts
 - **Functional genomics:** Whole \geq Sum of its parts
- Prototypical experiment designs
 - **2-group comparisons**
 - **Sequential profiling** – parametrized by a continuously-varying scalar variable
 - Hybrid of 2-group and sequential profiling

Transcriptomics: Granularity of questions

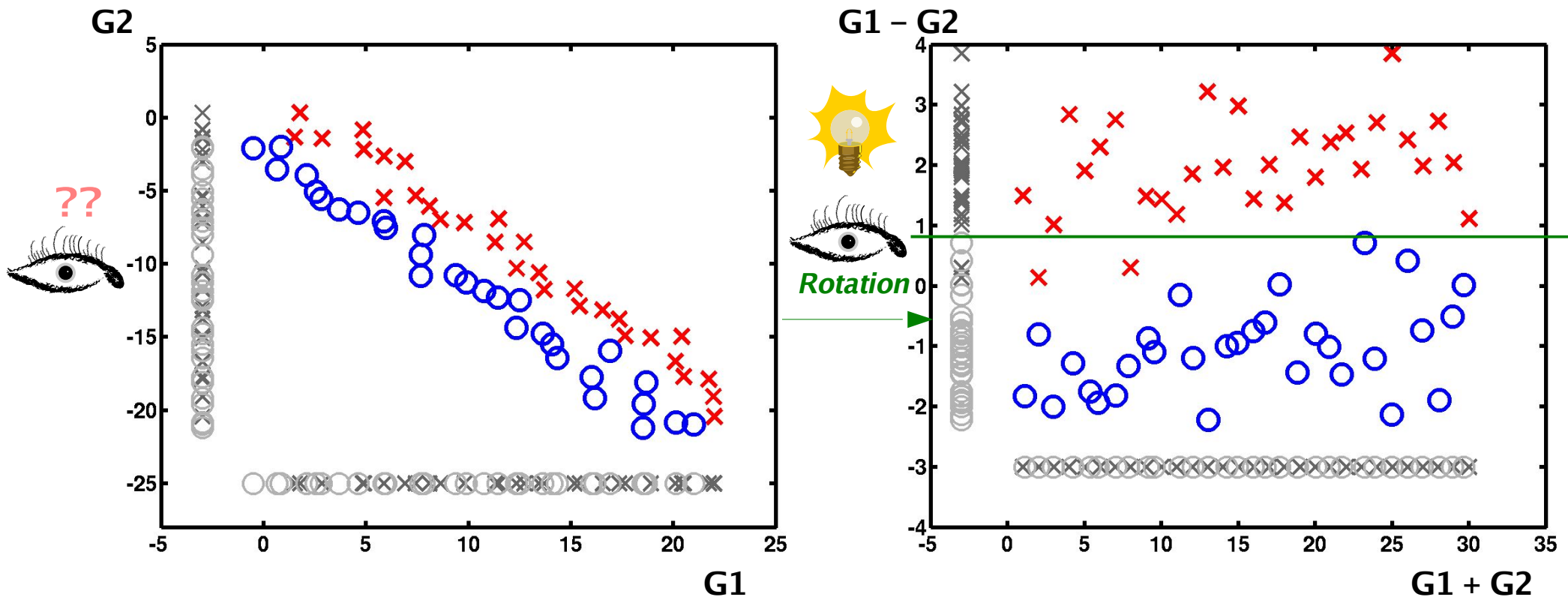
- Prototypical questions – *3 molecular scales or levels*
 - **Single:** Identify individual molecules associated with (possibly underwriting) a biological phenomenon
 - **Network:** Identify molecular networks associated with a biological phenomenon
 - **System:** Transcriptomic / global state or characterization of a biological system
- Example questions that can be practically asked given transcriptome profiling technology
 - Given known clinically-distinct disease conditions, what is the minimal gene set (their expression profiles) that distinguishes between these conditions with a reasonably high specificity and sensitivity?
 - Is there a transcriptomic (or its subset) signature that correlates with survival outcome of stage I lung adenocarcinoma patients?
 - Are the set of genes upregulated by cells C under morphogen M significantly enriched for specific functional or ontologic categories?

Transcriptomics: Paradigm shift

- How transcriptome profiling technologies change the way we think about and model biological systems, problems
- Traditional biology / genetics: Whole = Sum of its parts
 - Microarrays as a highly efficient large-scale application of northern blots, PCR
- Functional Genomics: Whole \geq Sum of its parts
 - Practical to think about combinatorial effect, leverage on multi-factorial effects.

Transcriptomics: Paradigm shift example

- Combinatorial effects example. Say we measure 2 genes G1, G2 in 30 disease X, and 30 control O subjects. Neither G1, G2, by themselves discriminate X from O. But (sign of) $G1 - G2$ does. $G1 - G2$ is the disease discriminant here.



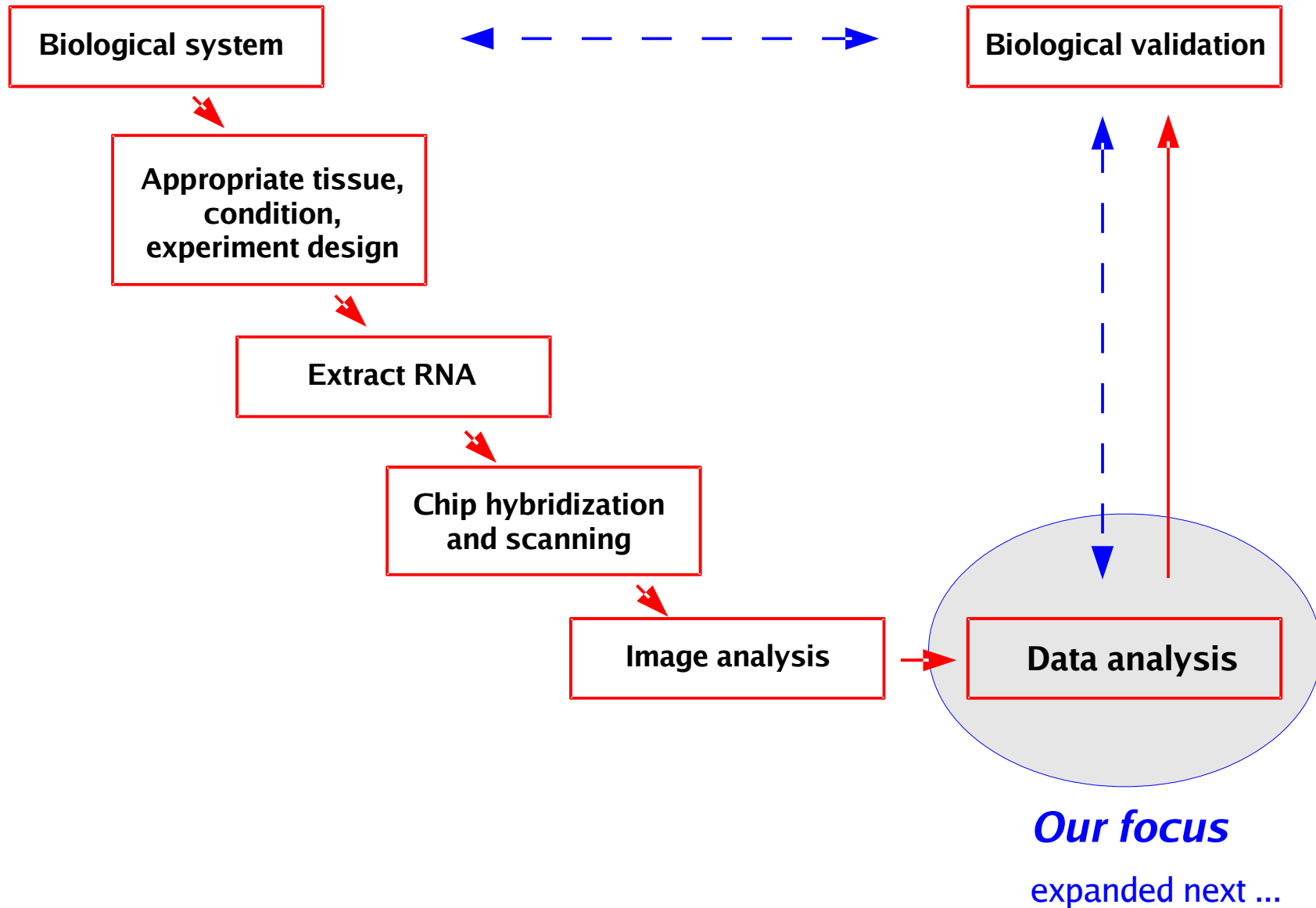
Transcriptomics: Prototypical experiment designs

- Prototypical experiment designs
 - Are not conceptually different from commonplace scientific experiment designs.
 - **2-group comparisons:** disease vs. control, treated vs. non-treated
 - **Sequential profiling** – parametrized by a continuously-varying scalar variable: time course, dosage-varying study
 - Hybrid of 2-group and sequential profiling

Typical steps in microarray-driven experimentation

- Experimental design involving biological system under investigation.
Replicates – biological and measurement / technical
- RNA target / probe preparation: Extract mRNA. Convert (to single strand cDNA typically). Label with fluorescence.
- Probe hybridization. Fluorescence scan.
- (Fluorescence) Image analysis
- (Post image) Data analysis and modeling to generate more focused hypotheses.
- Biological validation

Workflow in microarray-driven experimentation



Data analysis: Math applied to transcriptome domain

- Mathematical formulation of the biological / physical problem
 - Data representation. Modeling. Mapping physical problem into a metric space.
- “Correcting” noise and systematic measurement variation / bias
 - “Pre-processing”. Normalization. Replicate measurements.
- Uncovering coherent geometries and dominant variance structures intrinsic to data
 - “Supervised” and “unsupervised” math techniques. E.g., clustering, machine learning
- Likelihood of coherent structures / math results arising by chance
 - Statistics
- Squaring math results with *a priori* biological knowledge. Figure of merit
 - Statistics
- “Reverse engineering”. Correlation vs. causality
 - Graph and network theory.

Data analysis: The beginning

- Almost always microarray data analysis / modeling starts off with a spreadsheet (data matrix) post image analysis ...

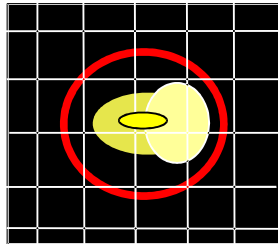


Image processing (black box?)

EntrezID	Symbol	Day1-1	Day3-1	Day5-1	Day7-1	Day10-1	Day15-1
13008	Csrp2	-2.4	74.6	25.5	-30.7	14.6	-50.1
17121	Mxd3	126.6	180.5	417.4	339.2	227.2	-76.2
17859	Mxi1	2697.2	1535	2195.6	3681.3	3407.1	1648.3
67255	Zfp422	458.5	353.3	581.5	520	348	106.3
18109	Nmyc1	4130.3	2984.2	3145.5	3895	2134.3	597.1
13555	E2f1	1244	1761.5	1503.6	1434.9	487.7	98.3
11921	Atoh1	94.9	181.9	268.6	184.5	198	-246.2
97165	Hmgb2	9737.9	12542.9	14502.8	12797.7	8950.6	458.9
18504	Pax2	379.3	584.9	554	438.8	473.9	565.2
21418	Tcfap2a	109.8	152.9	349.9	223.2	169.1	-115.3
21419	Tcfap2b	4544.6	5299.6	2418.1	3429.5	1579.4	2862.4

This is "Data"

Data analysis: Math formulation

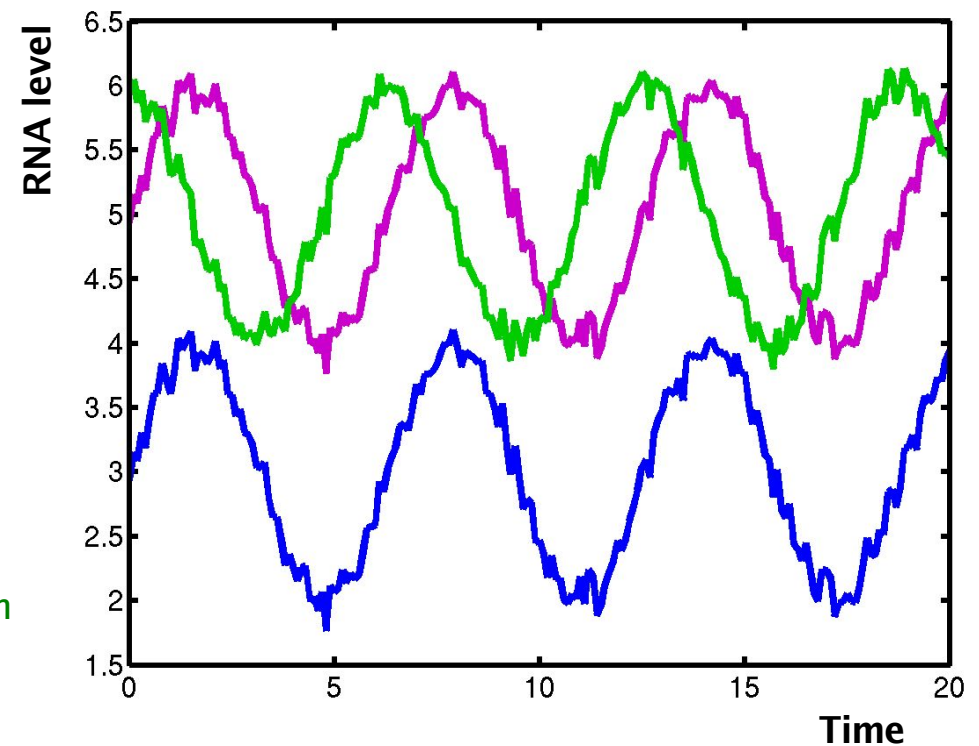
- To leverage on classical applied maths, first map data into a metric space – more generally a normed linear space
 - Our definition of similarity is embodied in the selected metric (more generally, a similarity measure)
 - Example: 2 different similarity measures are the Euclidean distance (intuitive geometric distance, this is a true metric), and Pearson linear correlation (this is not a true metric). Physically, Euclidean distance = difference in displacement, Correlation = difference in velocity

Correlation space

Magenta & Blue are *more* similar than Magenta & Green

Euclidean space

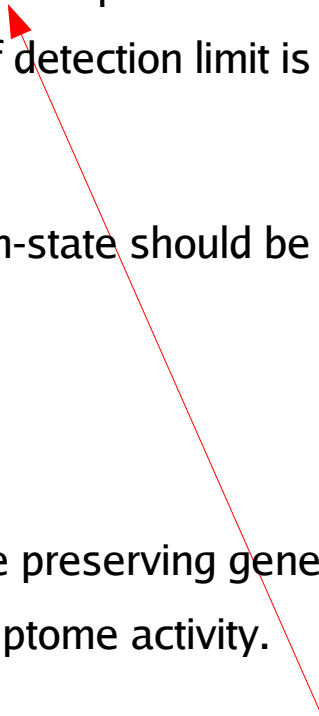
Magenta & Blue are *less* similar than Magenta & Green



Data analysis: Math formulation

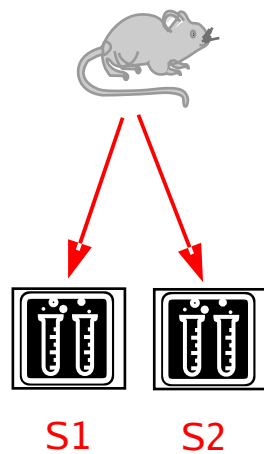
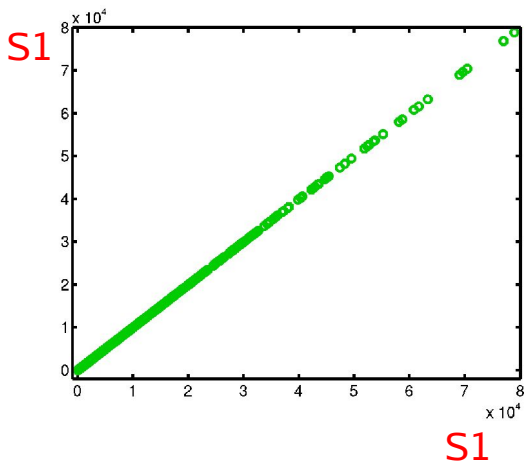
- Any gene \times sample data matrix can be viewed as,
 - *Genes in Sample space*
 - *Samples in Gene space*
 - Typically for transcriptome data, # Genes \gg # Samples
 - These spaces may have different similarity measures

Data analysis: Noise and measurement variations

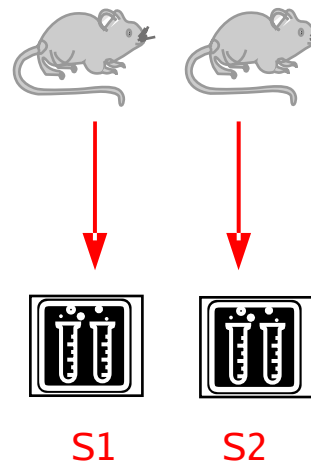
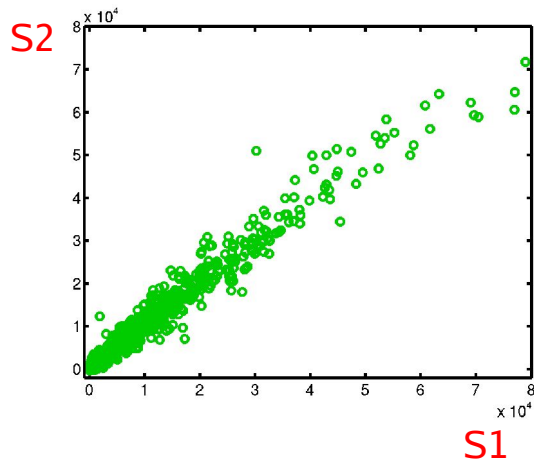
- How to detect existence of “noise” or systematic measurement biases / variations?
 - What is **Noise**? Deviations from a logical or scientific axiom / assumption. This deviation *may be* expressed / reflected in the (numerical) data. Clearly, if detection limit is gross the expression of noise is minimized.
 - Example of logical axiom: Replicate measurements of a system-state should be similar in given metric space.
 - How to correct for noise? Normalization
 - Normalization is a math transformation to minimize noise, while preserving gene expression variation resulting from biologically relevant transcriptome activity.
 - Which transformation? Depends upon reference logical / scientific axiom violated
 - Normalization example: Equalize the mean transcriptome levels across samples.
 - Replicates are critical to characterize noise, measurement variation
- 

Data analysis: Noise and measurement variations

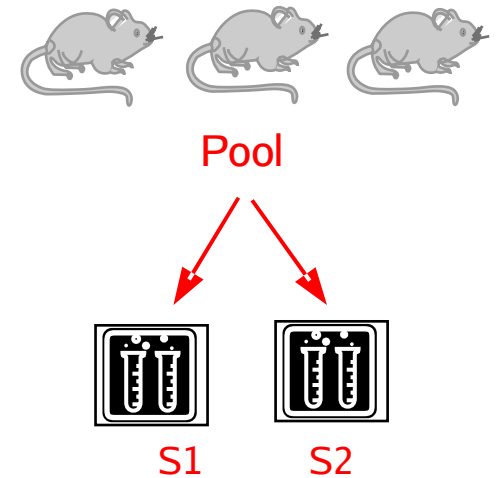
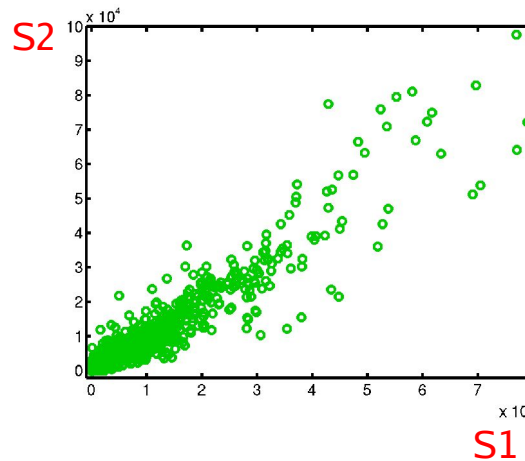
- Different concepts of a Replicate
 - Scatter plots of reported transcriptome levels between replicates



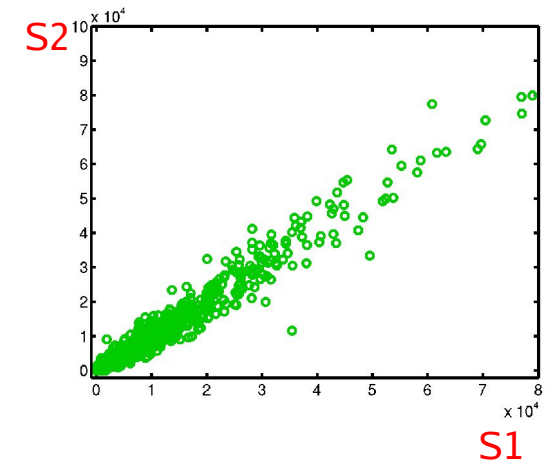
Measurement variation



Biological variation + Measurement variation



Measurement variation

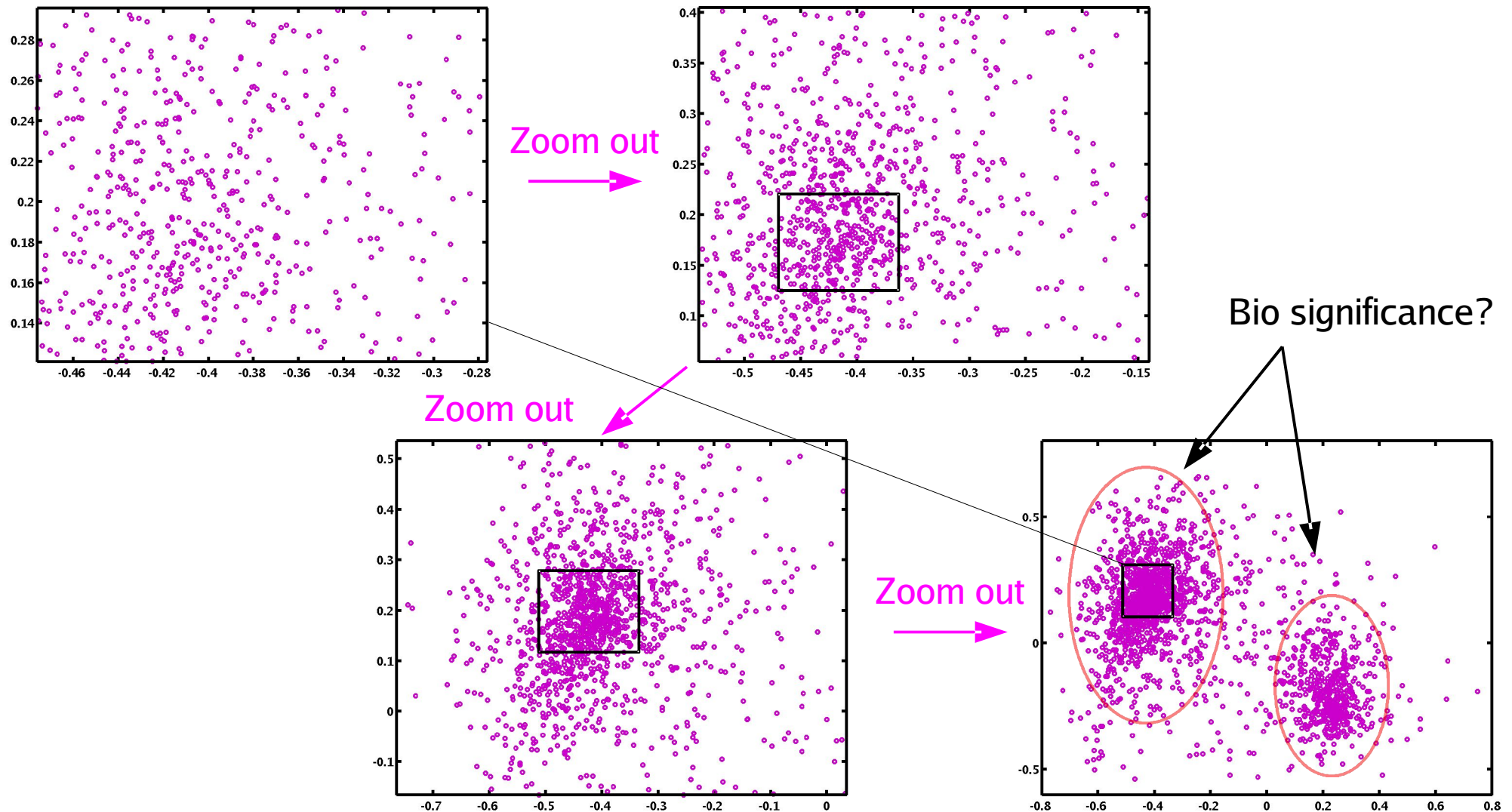


Data analysis: Intrinsic coherent structures

- Given a normalized data set. Recall that we can view the data as
 - Genes in Sample space
 - Samples in Gene space
- Question: Might these be coherent geometries and dominant variance structures intrinsic to data
 - Aim to create variationally meaningful data subsets (structure) from the mix of all features
 - Do coherent structures exist?
 - “Supervised” and “unsupervised” math techniques. E.g., clustering, machine learning
- Unsupervised = sample labels are not used by method. Supervised = sample labels are necessary input into method.
- Many math methods exist, most ported from physical and engineering science. Which is “*best*”? 2 rules of thumb
 - Scientific question should guide choice of method. Not other way around
 - Upon deciding on a method, do simulation exercise. Figure of merit

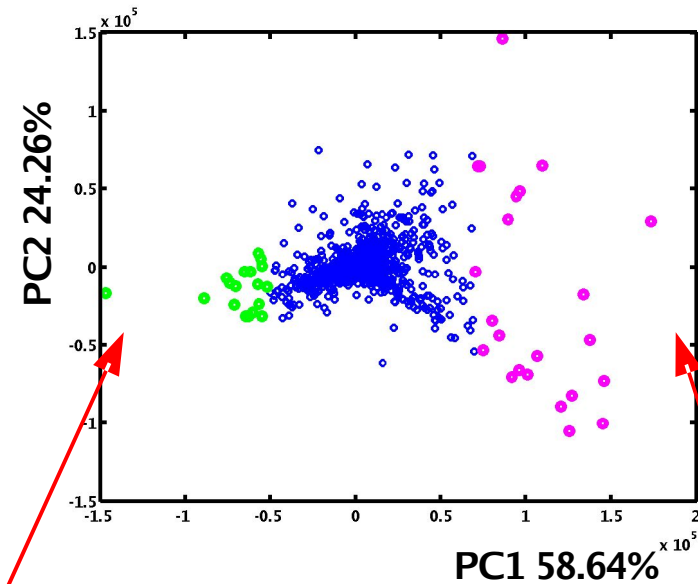
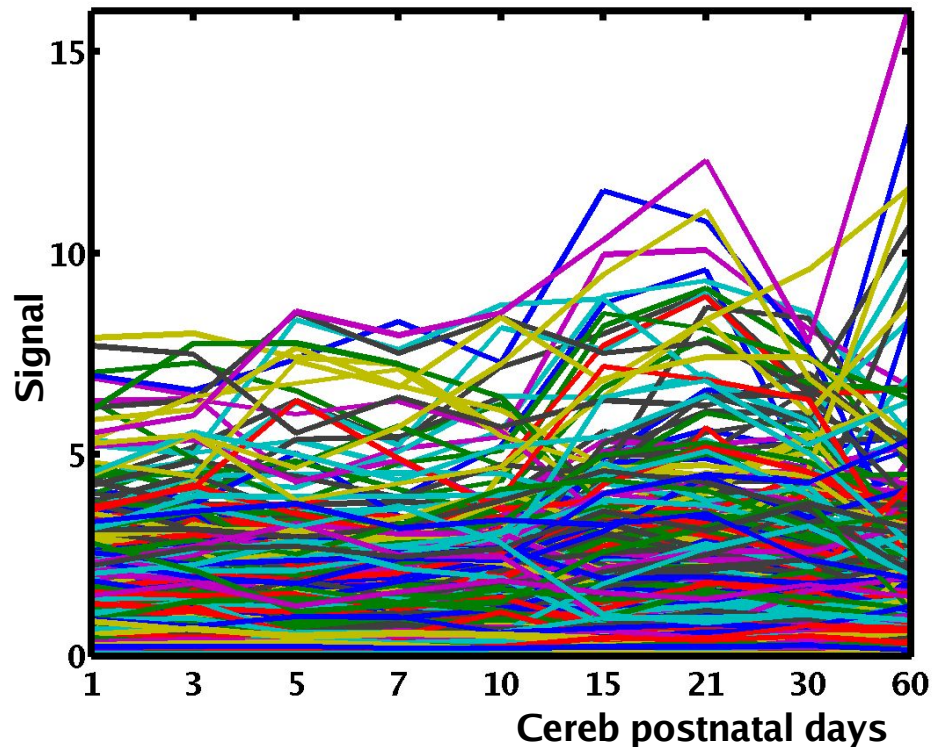
Data analysis: Intrinsic coherent structures example

- Graphical examples of internal geometries / regularity in genomic data at multiple scales



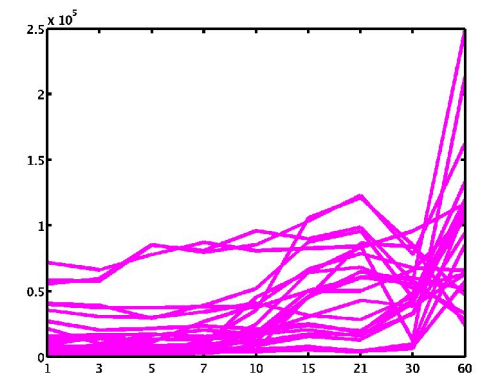
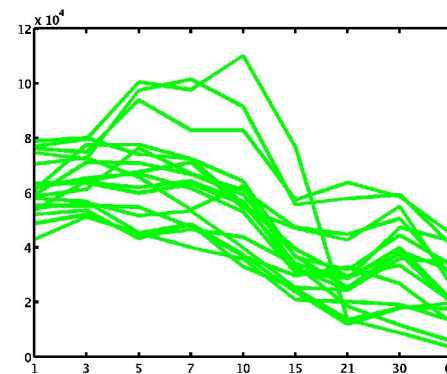
Data analysis: Coherent structures example

- Example: Mouse cerebellar development 6K genes at 9 time stages (duplicate).
 - Genes in Sample space I. Euclidean space.



PCA representation

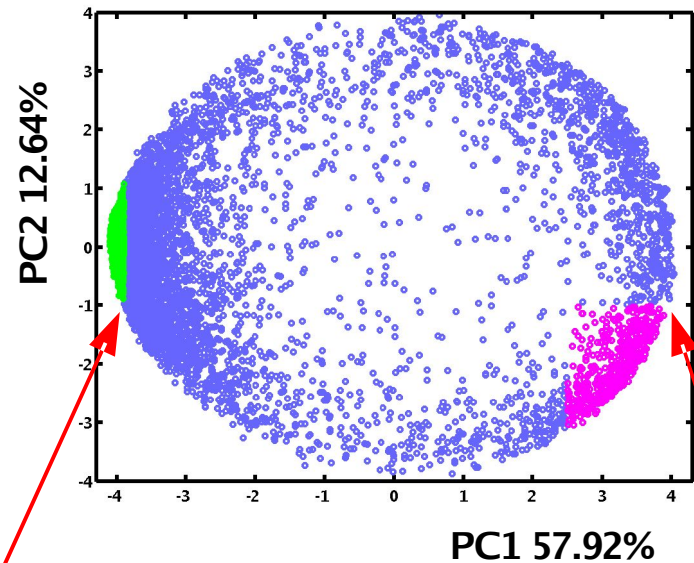
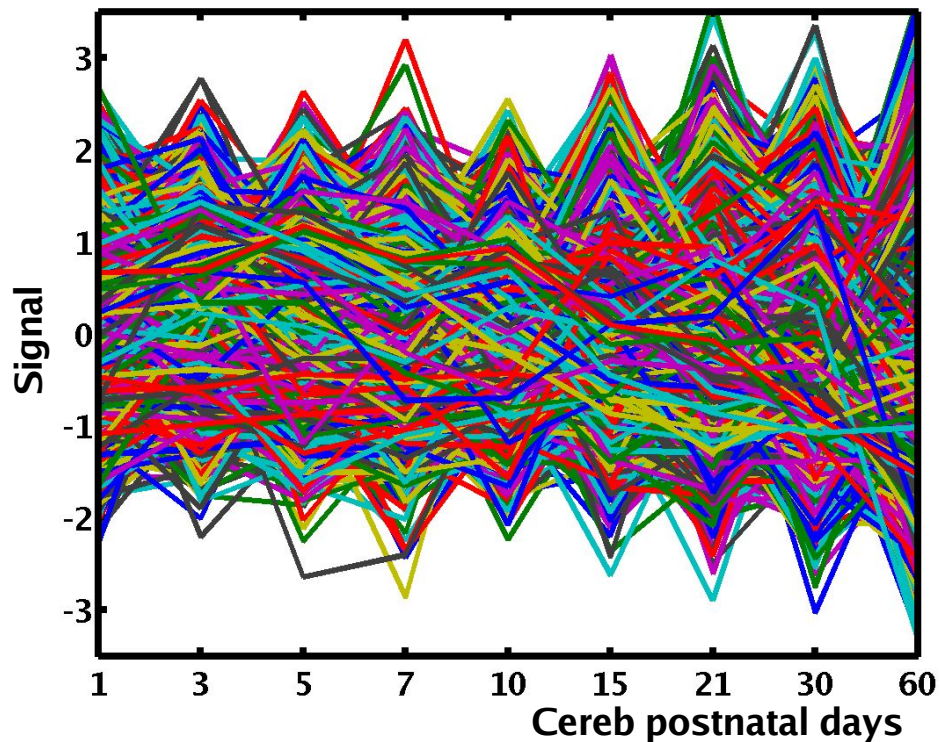
Coherent structures / regularity here?



Data analysis: Coherent structures example

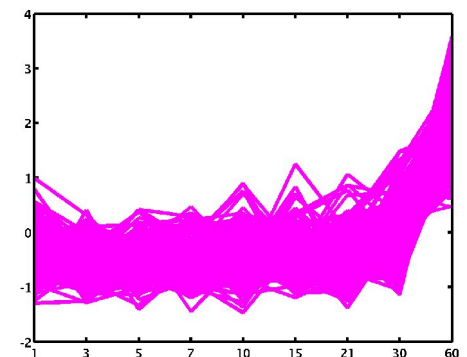
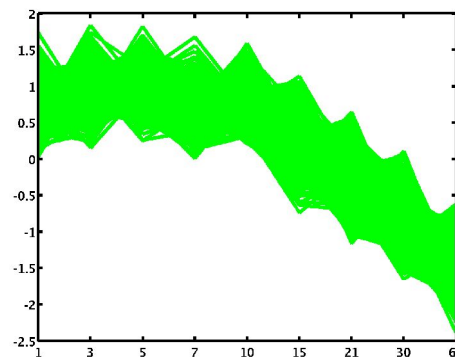
- Example: Mouse cerebellar development 6K genes at 9 time stages (duplicate).

- Genes in Sample space II. Correlation space.



PCA representation

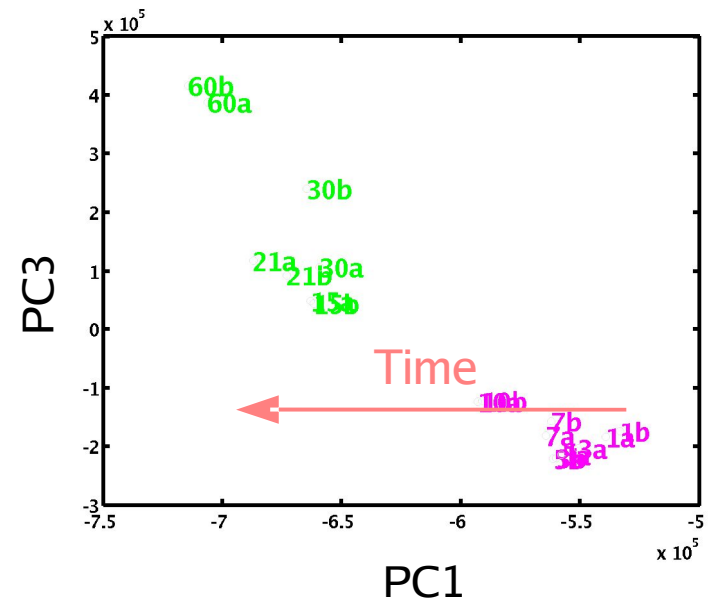
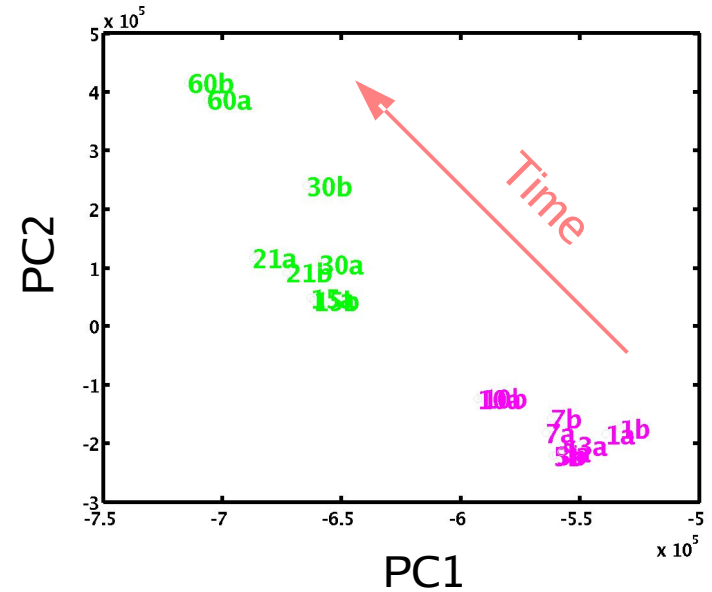
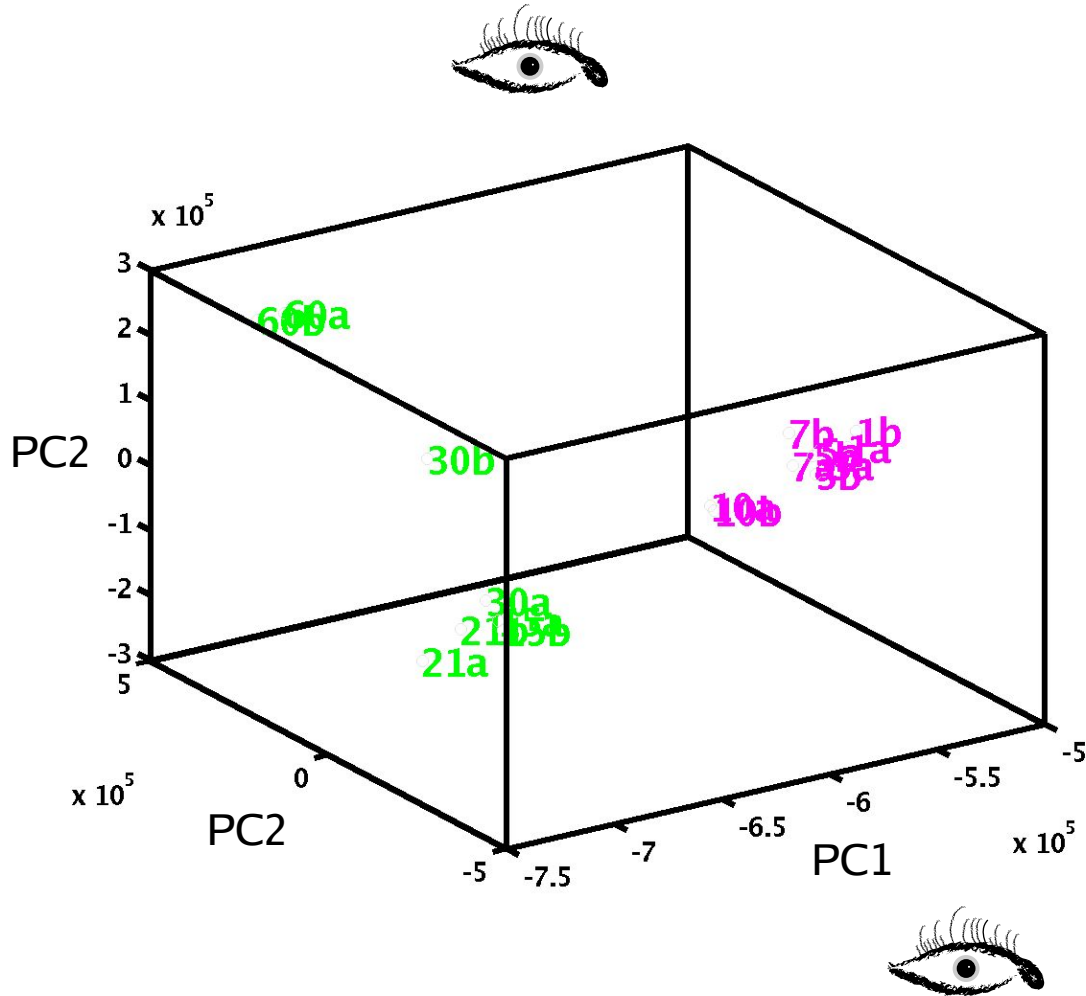
Coherent structures / regularity here?



Data analysis: Coherent structures example

- Example: Mouse cerebellar development 6K genes at 9 time stages (duplicate).

- Samples in Gene space I. Euclidean space

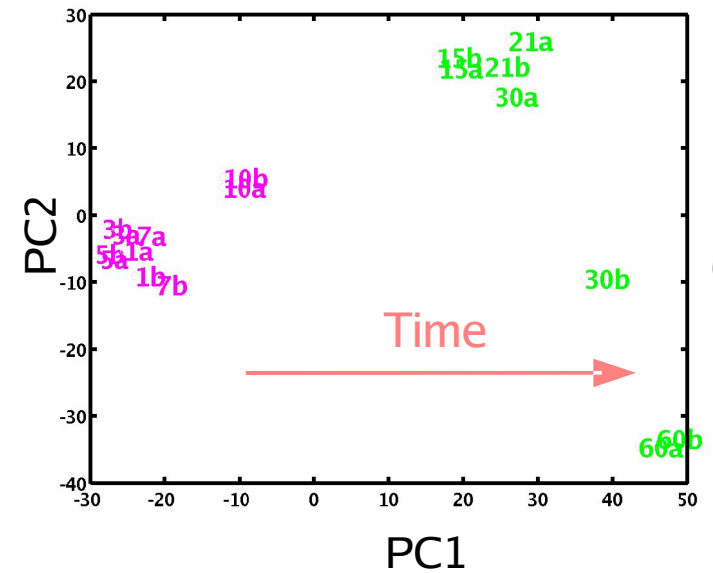
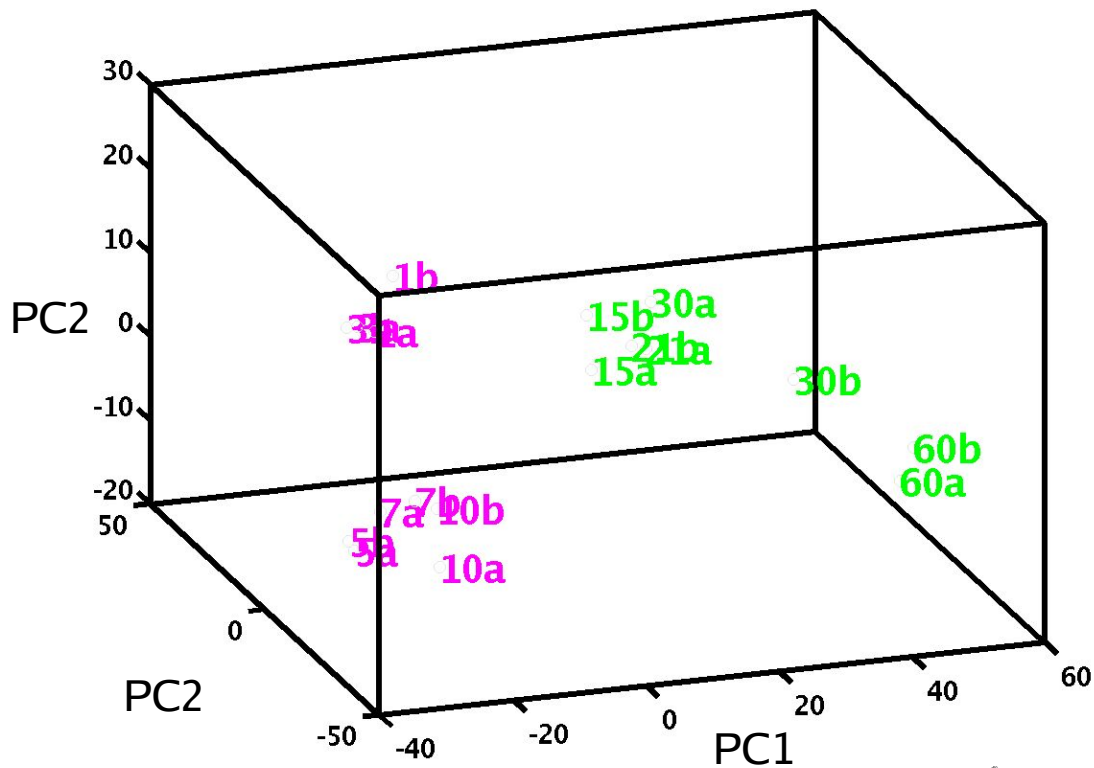


Do configurations say anything bio meaningful?

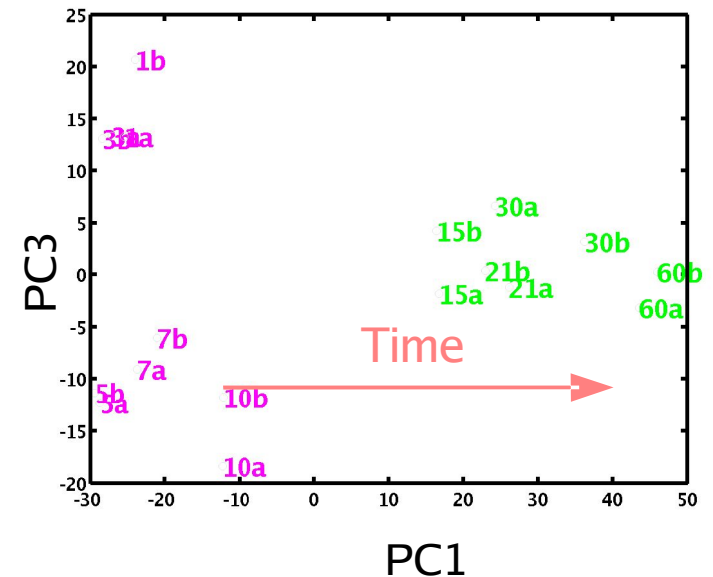
Data analysis: Coherent structures example

- Example: Mouse cerebellar development 6K genes at 9 time stages (duplicate).

- Samples in Gene space II. Correlation space



Do configurations say anything bio meaningful?



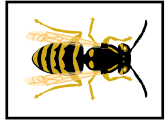
Data analysis: How likely are coherent structures due to chance?

- Squaring math results with chance
 - Statistics
- Assumptions about null hypothetical distribution
- Permutation testing: Permute data. Run similar analyses to extract coherent structures. Examine result relative to original unperturbed case

Data analysis: How well does model mirror physical system?

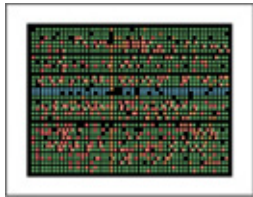
- Squaring math results with *a priori* biological knowledge. Figure of merit
 - Statistics
- Coherent / dominant mathematical structures that are uncovered via math from data should ideally have a physical or extra-math correlate.
- There are many analytic methods and attendant models that can be applied onto 1 dataset. Which best mirrors physical system?
- 1 physical system --> 1 data set --> **>1** possible models --> 1 physical system?
 - How to pick?
 - Well-definedness
 - Reality checks. How likely is this data and methods

Data analysis: Math applied to transcriptome domain



Biological system

Transcriptome



Gene	P1-1	P3-1	P5-1	P7-1	P10-1
Csrp2	-2.4	74.6	25.5	-30.7	14.6
Mxd3	126.6	180.5	417.4	339.2	227.2
Mxi1	2697.2	1535	2195.6	3681.3	3407.1
Zfp422	458.5	353.3	581.5	520	348
Nmyc1	4130.3	2984.2	3145.5	3895	2134.3
E2f1	1244	1761.5	1503.6	1434.9	487.7
Atoh1	94.9	181.9	268.6	184.5	198
Hmgb2	9737.9	12542.9	14502.8	12797.7	8950.6
Pax2	379.3	584.9	554	438.8	473.9
Tcfap2a	109.8	152.9	349.9	223.2	169.1
Tcfap2b	4544.6	5299.6	2418.1	3429.5	1579.4

Math formulation
Data representation

Map data into metric/measure space, model appropriate to biological question

Normalization
Replicates

Correct for noise, variation arising not from bio-relevant transcriptome program

Un/supervised math techniques. E.g., clustering, networks, graphs, myriad computational techniques guided by overriding scientific question !

Uncover coherent / dominant variance structures in data

Chance modeled by null hypothesis
Statistics
Permutation analyses

Likelihood of coherent structures arising from chance alone

Prediction. Inferential statistic.
Hamilton's principle – minimizing an energy functional
Correlation vs causality

Do coherent structures mirror biological parameters, system?

The Big Picture

User-friendly references

- These readings don't refer to microarrays per se, but capture the ethos of applied math in science, biology
- E. Wigner. *The unreasonable effectiveness of mathematics in the natural sciences*. Comm. In Pure and Applied Math. 13 (1), 1963
<http://www.dartmouth.edu/~matc/MathDrama/reading/Wigner.html>
- A. Turing. *The chemical basis of morphogenesis*. Phil Trans Royal Soc London (Series B), 641 (237), 1952 <http://www.turingarchive.org/browse.php/B/22> (not easy read but enlightening, DNA structure was resolved in 1954. Computer science founder ...)
- Good luck!
- *The discoveries that one can make with the microscope amount to very little, for one sees with the mind's eye and without the microscope the real existence of all these little beings*. Georg-Louis Leclerc de Buffon (of the Buffon needle problem, and 44-volume Histoire Naturelle)