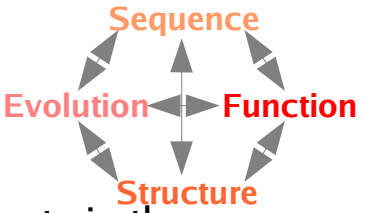
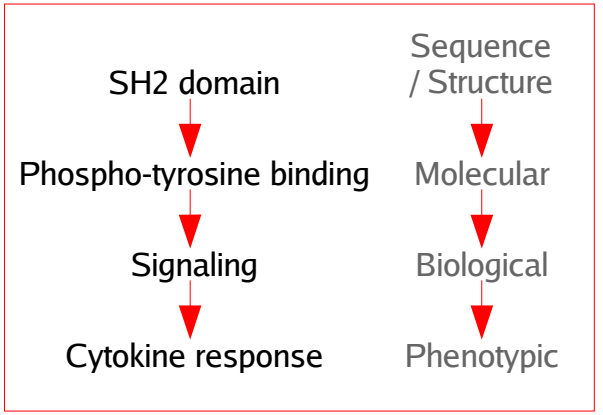
Module 4 (Functional Genomics/FG) Lecture 1 recall



- FG definition, place in tetralogy
- Review heritable primary functional elements in the genome and "epigenome". Central dogma – 1-way transfer of transcriptome info.
- Survey 2 primary scalable transcriptome profiling principles (technologies): sequencing (SAGE), nucleotide complementation (microarrays). Main idea: A small subset of nucleotides uniquely represents each RNA species / transcript
- Transcriptome profiling study assumptions, caveats
  - Biological system / state of interest engages transcriptome machinery.
  - Averaging RNA levels across heterogeneous cell populations
  - Technical issues for microarrays
    - diff hybridization rate for each RNA species
    - Cross-hybridization.
    - Fluorescent intensity $\propto$ RNA abundance
  - (Today) **Noise**: measurement / technical / biological variations. Choice of reference system.

Module 4 (Functional Genomics/FG) Lecture 1 recall

- 2 archetypal FG questions
  - What *function* does a given molecule X have in a specific biological system / state?
  - Which molecules (interactions) characterize / modulate a given biological system / state?
- Different levels of *function* for a bio molecule
  - Chemical / mechanistic – microscopic scale event
  - Biological – effects a phenotype (a macroscopic scale event)



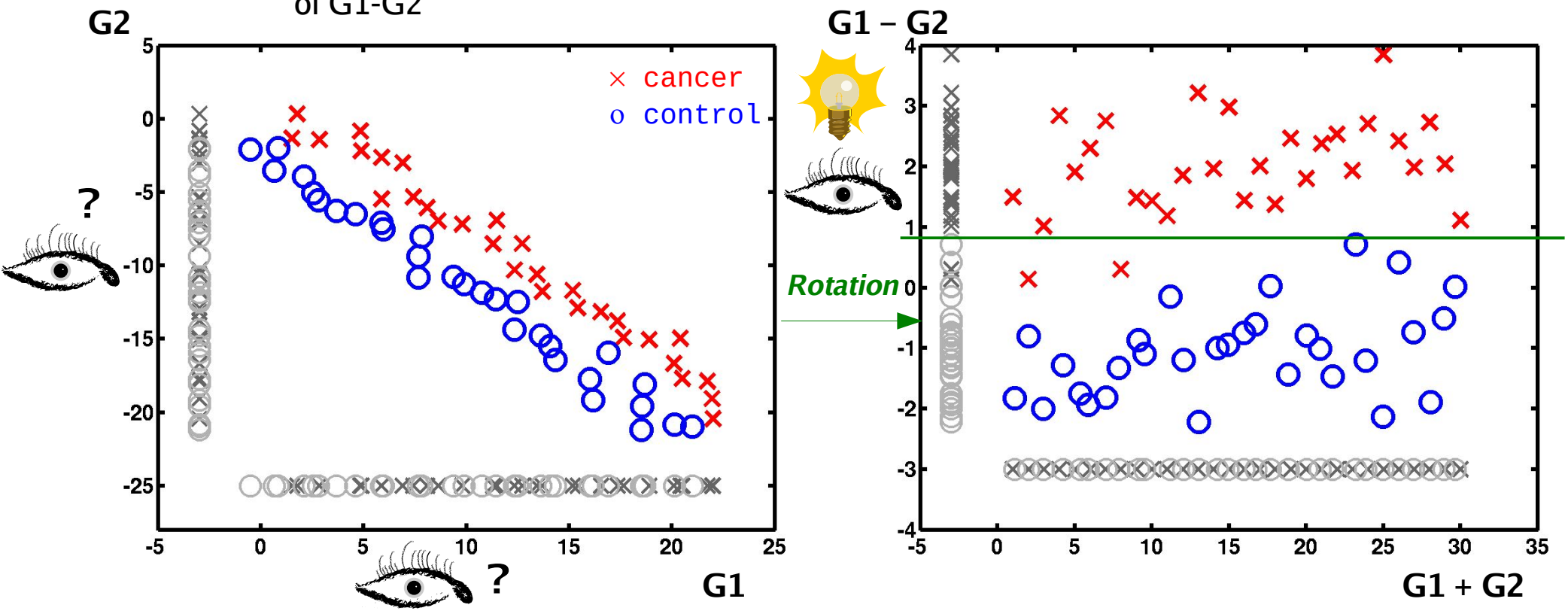
- How does the ability to measure many (~$10^{3\text{-}4}$) *different* bio-molecules at a time resolve the 2 questions above?
  - Without appropriate experiment design, and data analysis / interpretation, it does not help

FG: Pre-analysis pointers

- Example questions that may be asked given profiling technology
  - For a clinically-distinct set of disease states, what is the minimal transcriptome subset distinguishing between these conditions?
  - Is there a (sub)transcriptome 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 particular functional / ontologic categories?
  - Can gene interaction pathways be recovered from the transcriptome profiles of a time resolved biological process?

Module 4 (Functional Genomics/FG) Lecture 1 recall

- 2 diff qualitative views with parallel high throughput transcriptome profiling technologies
  - View 1*: Whole = Sum of individual parts*. Only an efficient way to screen many molecular quantities at a time.
  - View 2*: Whole > Sum of individual parts*. As above, plus unraveling intrinsic regularities (eg. correlations) between measured molecula quantities.
  - Illustration: Measure Not many **2** quantities G1, G2 in 2 disease populations. Discriminant is sign of G1-G2



Module 4 FG Lecture 2 outline: Large-scale transcriptome data analysis

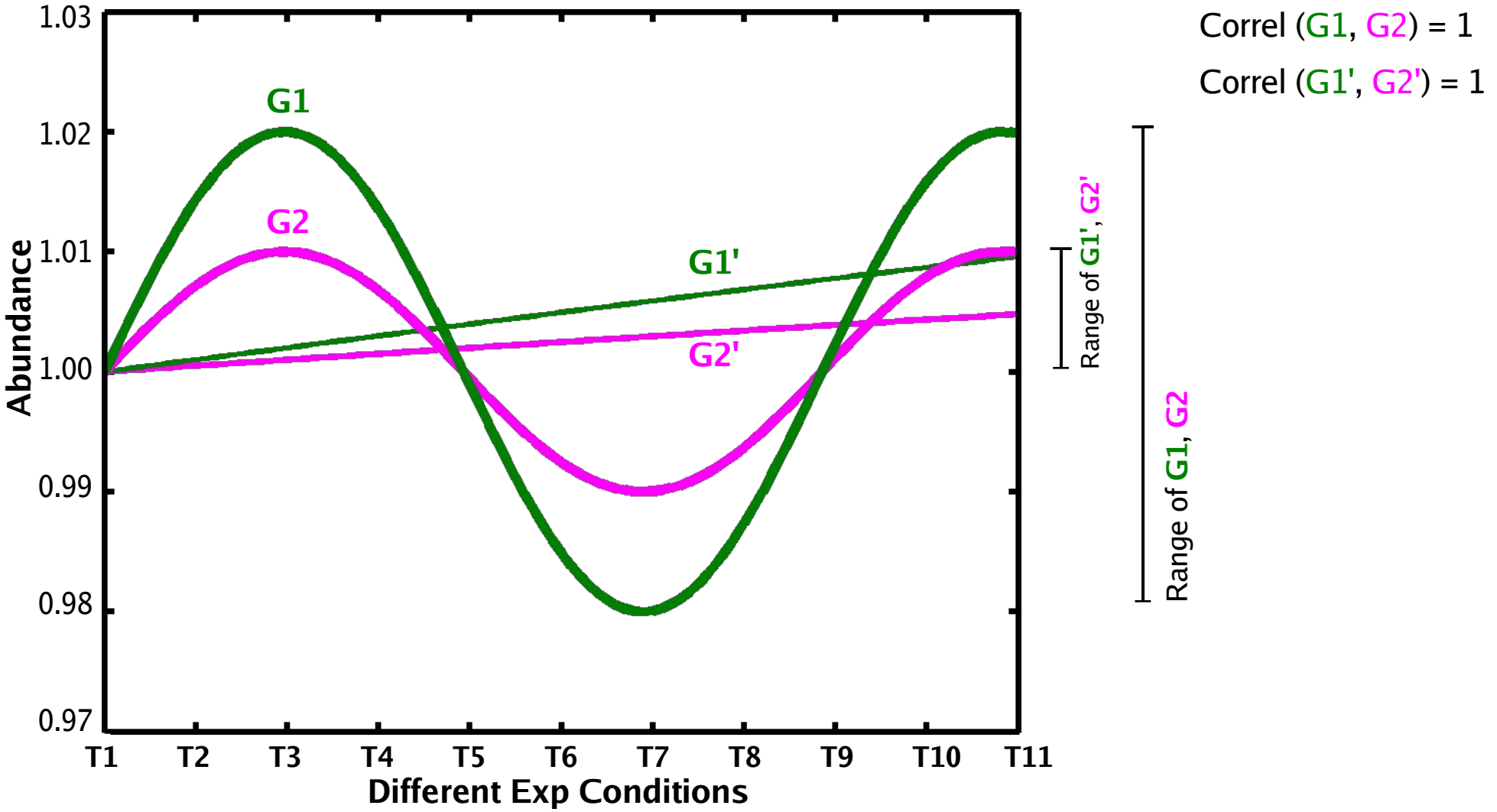
- Pre-analysis
  - Prototypical experiment designs. Exercising the transcriptome machinery
- Generic workflow in transcriptome profiling-driven studies
- Data representation, modeling, intrinsic regularities
  - Dual aspects of a data matrix
  - Defining a measure / metric space
  - Distributions (native, null) of large scale expression measurements
  - Noise and definitions of a replicate experiment. Pre-processing, normalization
  - What are *regularities*?
  - Unsupervised and supervised analyses. What is a cluster? Clustering dogma.
  - Statistical significance of observed / computed regularities
- Figure/s of merit for data modeling
  - Correspondence between regularities and biologic (non-math) parameters
  - Model predictions

FG: Pre-analysis, experiment design

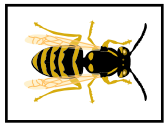
- Prototypical experiment designs
  - 2-group comparisons **(A)**
  - Series of experiments parametrized by a well-ordered set (eg. time, dosage) **(B)**
  - Hybrid of above
- Categorizing example questions posed earlier
  - For a clinically-distinct set of disease states, what is the minimal transcriptome subset distinguishing between these conditions? **(A)**
  - Is there a (sub)transcriptome signature that correlates with survival outcome of stage I lung adenocarcinoma patients? **(B+A)**
  - Are the set of genes upregulated by cells *C* under morphogen *M* significantly enriched for particular functional / ontologic categories? **(A)**
  - Can gene interaction pathways be recovered from the transcriptome profiles of a time resolved biological process? **(B+A)**
- Exercising the transcriptome machinery to its maximal physiological range

FG: Pre-analysis, experiment design

- Exercising the transcriptome machinery to its maximal physiological range. How? Subject system to extremal conditions.Define extremal condition? Important when looking for relationship between measured quantities. Eg. molecules G1, G2, G1', G2' under conditions T1-11.



FG: Generic workflow in transcriptome profiling-driven studies



Biological system and a question

Biological validation

Appropriate tissue, condition, experiment design

Extract RNA, chip hybridization, scanning

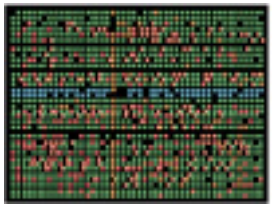


Image analysis

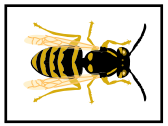
Data analysis modeling

Transcriptome Profiling Big Picture

Focus today

expanded next ...

FG: Generic (expanded) workflow in transcriptome data analysis



Biological system / state

Transcriptome

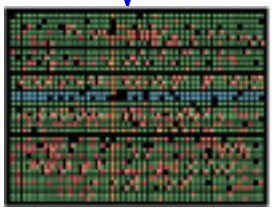


Image Analysis

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 overiding scientific question !

Uncover regularities / dominant variance structures in data

Chance modeled by null hypothesis
Statistics
Permutation analyses

Likelihood of regularities arising from chance alone

Prediction. Inferential statistic.
Minimizing an energy functional
Correlation vs causality
Figure of Merit

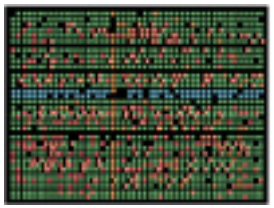
Do regularities reflect biological system – state?

Analysis / Modeling

Big Picture

FG: Data representation, modeling, intrinsic regularities ... starting point

- Almost always microarray data analysis / modeling starts with a spreadsheet (data matrix) post image analysis
  - Data representation = mapping a physical problem and measurements onto a math framework. Why? Leverage of classical math results.



Chip for Day 1-1

Zoom into grid coordinate corresponding to probe sequence for Mxd3

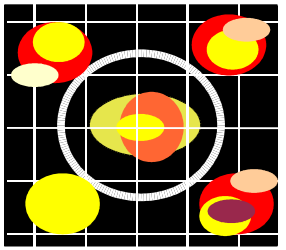


Image processing (black box?)

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

Here is “data” - a matrix, typically #Rows (Genes) >> #Columns (Samples) ... linear algebra will be very useful here

FG: Dual aspects of a data matrix

- Typical trancriptome profiling studies have data matrix D of N genes $\times$ M samples / conditions where N >> M.
- Entries of D are typically real numbers that's either dimensioned/dimensionless. Eg. fold is dimensionless. This fact affects expectation about the characteristics of the distribution of gene intensities (Later).
- 2 dualistic views of D (emphasizing different aspects of the system / state)
  - Genes in sample space: Each gene has an M-sample profile / distribution of intensities
  - Samples in gene space: Each sample has an N-gene “transcriptome” profile
- In either case, data's high dimensionality makes it non trivial to visualize this data
  - Recalling the 2 gene G1, G2 example (cancer vs. control) from last time / earlier – coherent relationships / regularities (if indeed they exist!!!) between genes and sample conditions are far less obvious now

FG: Defining a measure space to work in

- Define a measure / metric space S where data lives:
  - To quantify dis/similarity between objects – (a fundamental requirement!)
  - Choice of dis/similarity measure, metric (should) reflect biological notion of "alikeness"
  - A metric d is a map from $S \times S \rightarrow$ non-negative reals such that for any x, y, z in S, 3 cardinal properties hold
    - M1 Symmetry: $d(x, y) = d(y, x)$
    - M2 Triangle inequality: $d(x, y) + d(y, z) \geq d(x,z)$
    - M3 Definite: $d(x, y) = 0$ if and only if $x = y$
  - Metrics d_1 and d_2 are equivalent if there exists $\lambda , \lambda' > 0$ such that $\lambda d_1 \leq d_2 \leq \lambda' d_1$ on S.
- Ideally, data is mapped into metric space to leverage on classical math theorems – even better map into inner product space ($\mathbb{R}^n$, $\langle *, * \rangle$ dot product)
- Metric examples: Euclidean, Taxicab / City Block / Manhattan, Geodesics
- Non-Metric dissimilarity measures:
  - Correlation coefficient (violates M2, M3) – angle between the projections of X and Y onto the unit hypersphere in S ($\mathbb{R}^n$, X and Y are n-vectors)
  - Mutual information (violates M3) – average reduction in uncertainty about X given knowledge of Y.

FG: Defining a measure space to work in

- Basic connections between Euclidean metric e, correlation coefficient p on same data in S = ($\mathbb{R}^n$, $\langle *, * \rangle$ dot product). Let x, y be in S; $\underline{1}$ = n-vector of 1's; $\underline{0}$ = n-vector of 0's or origin of $\mathbb{R}^n$; μ_x and μ_y be avg of x, y components. $|x| = \sqrt{\langle x, x \rangle}$ = length of x. [Simple exercise to check these calculations]

  - A1 $\underline{x} = x - \underline{1}\mu_x$, $\underline{y} = y - \underline{1}\mu_y$ are orthogonal to $\underline{1}$. Mean (μ) centering is equiv to map from $\mathbb{R}^n$ to the $\mathbb{R}^{n-1}$ hyperplane orthogonal to $\underline{1}$. Picture next slide

  - A2 Variance (Std^2) of x, $\sigma_x^2 = \langle \underline{x}, \underline{x} \rangle / (n-1) = |\underline{x}|^2/(n-1)$. So $\underline{x}/\sigma_x$ lives on the hypersphere radius $\sqrt{(n-1)}$ centered at $\underline{0}$ in $\mathbb{R}^{n-1}$. Picture next slide

  - A3 Recall correlation, $p(x,y) = \langle \underline{x}/|\underline{x}| , \underline{y}/|\underline{y}| \rangle = \langle \underline{x}, \underline{y} \rangle / (|\underline{x}| \, |\underline{y}|) = \cos(\angle(\underline{x},\underline{y}))$. Since $\langle a, b \rangle = |a| \times |b| \times \cos(\angle(a,b))$. Easy to see that $|\underline{x}| \, |\underline{y}| \, p(x, y) = \langle \underline{x}, \underline{y} \rangle$.

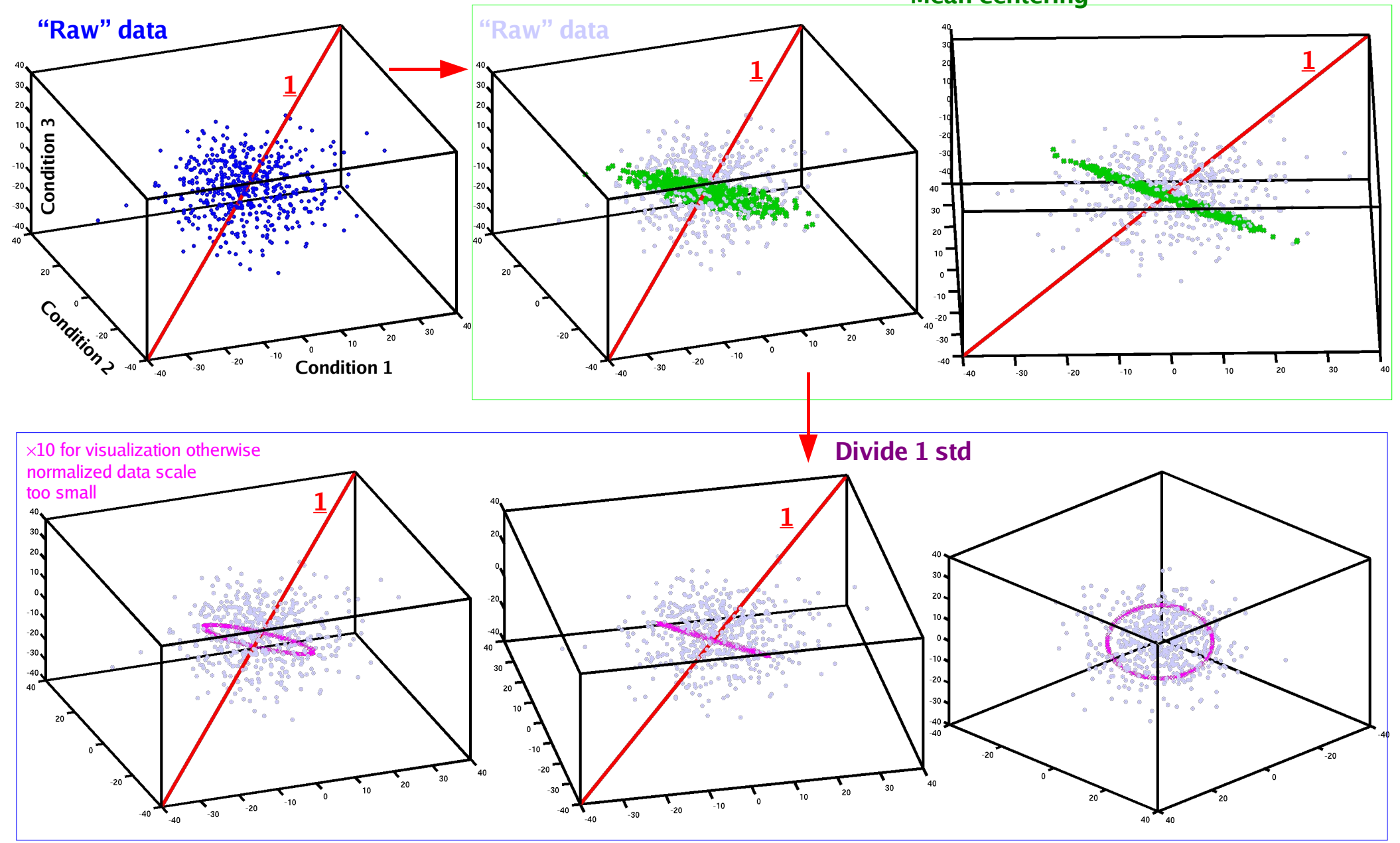
  - A4 Recall euclidean dist x to y: $e^2(x,y) = \langle x-y, x-y \rangle = \langle x, x \rangle + \langle y, y \rangle - 2 \langle x, y \rangle$.

  - A5 Say x and y are μ centered with std 1. Clearly $x = \underline{x}$, $y = \underline{y}$ and $\langle x, x \rangle = \langle y, y \rangle = |x|^2 = |y|^2 = (n-1)$. Plugging A3 into A4, $e^2(x,y) = \langle x, x \rangle + \langle y, y \rangle - 2\langle x, y \rangle = 2(n-1) - 2|x| \, |y| \, p(x,y) = 2(n-1)(1- p(x,y))$.

  - ie, $e^2(x,y) \propto -p(x, y)$ for x,y on the $\underline{0}$ centered radius $\sqrt{(n-1)}$ hypersphere of $\mathbb{R}^{n-1}$. We have encoded correlation structure in S = ($\mathbb{R}^n$, $\langle *, * \rangle$ dot product) with Euclidean metric. More about this in Preprocessing / Normalization section

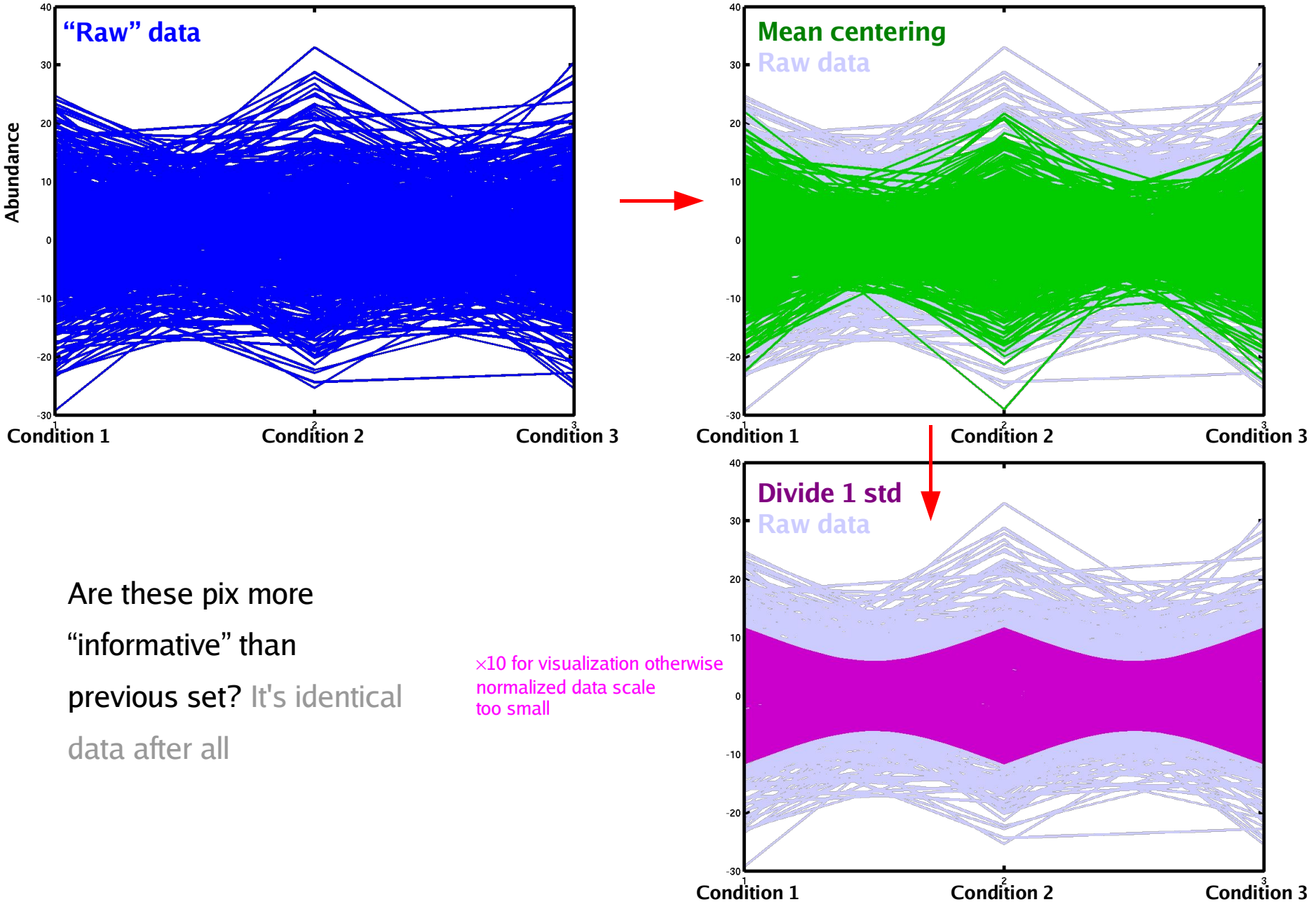
FG: Example transformations of data in $\mathbb{R}^3$ vision #1

- Geometric action of μ centering, $\div 1\ \sigma$ in $\mathbb{R}^3$. Data = 500 genes $\times$ 3 conditions. Genes in sample space view



FG: Example transformations of data in $\mathbb{R}^3$ vision #2 (flatland)

- Geometric action of μ centering, $\div 1$ σ in $\mathbb{R}^3$. Data = 500 genes $\times$ 3 conditions. Genes in sample space view



Are these pix more "informative" than previous set? It's identical data after all

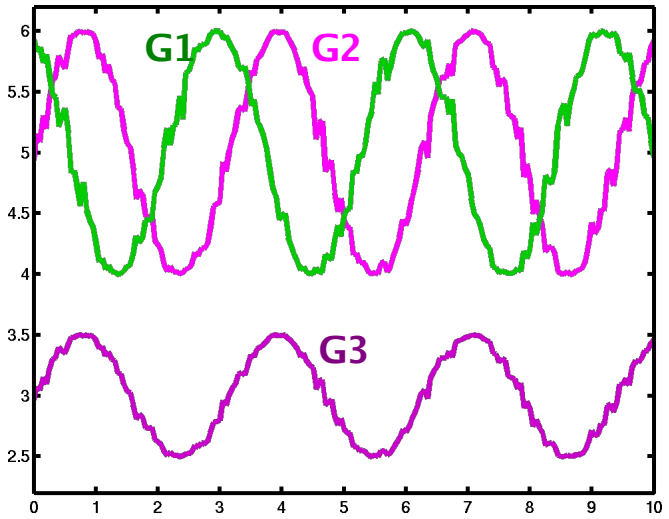
×10 for visualization otherwise normalized data scale too small

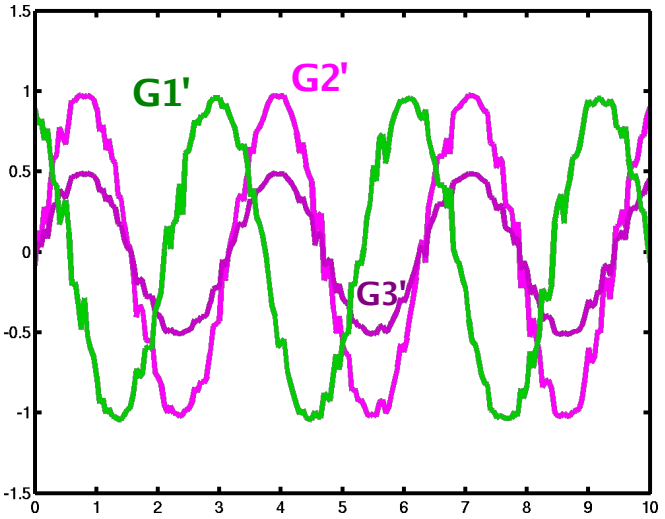
FG: Defining a measure space to work in

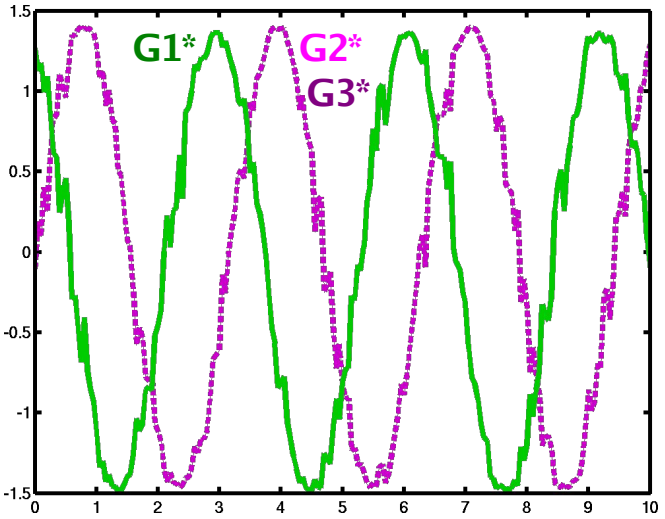
- Graphic example 1 of differences in notion of similarity embedded in Euclidean metric vs. correlation space. Time series of 3 genes: G1, G2, G3
  - Euclidean space: (G2 and G1) are more similar than (G2 and G3)
  - Correlation space: (G2 and G3) are more similar than (G2 and G1)
  - After mean centering, ÷1 std, in both Euclidean and correlation spaces (G2* and G3*) are more similar than (G2* and G1*)

Mean center

÷ 1 std



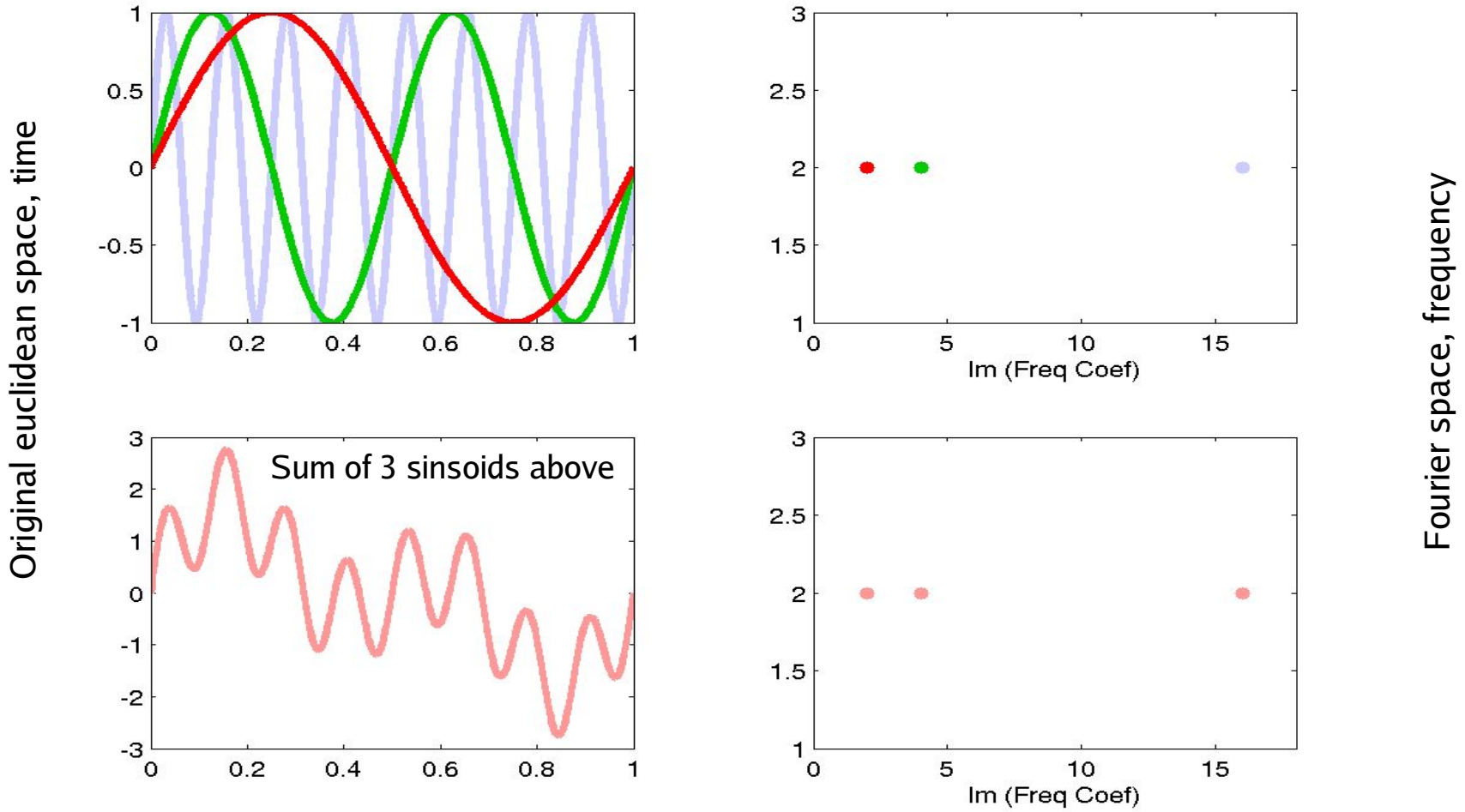




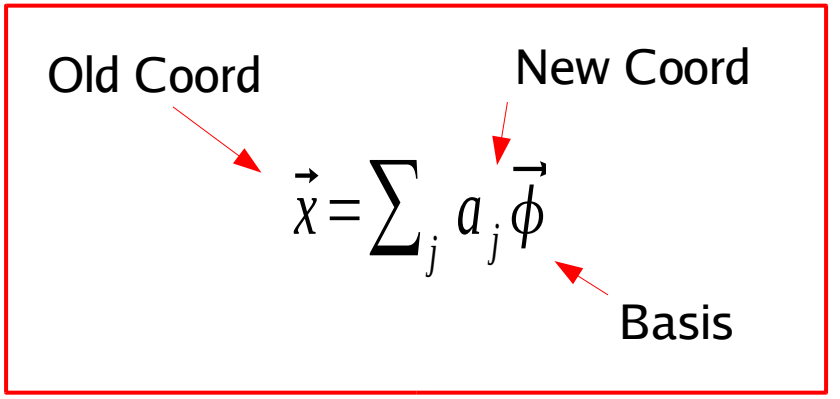
Obviously I rigged G2 to be scalar multiple of G3. So G2*=G3*

FG: Defining a measure space to work in

- Graphic example: When periodicity is a property of interest in data. Fourier representation
  - Three genes G1, G2, G3 – their time series



FG: Defining a measure space to work in



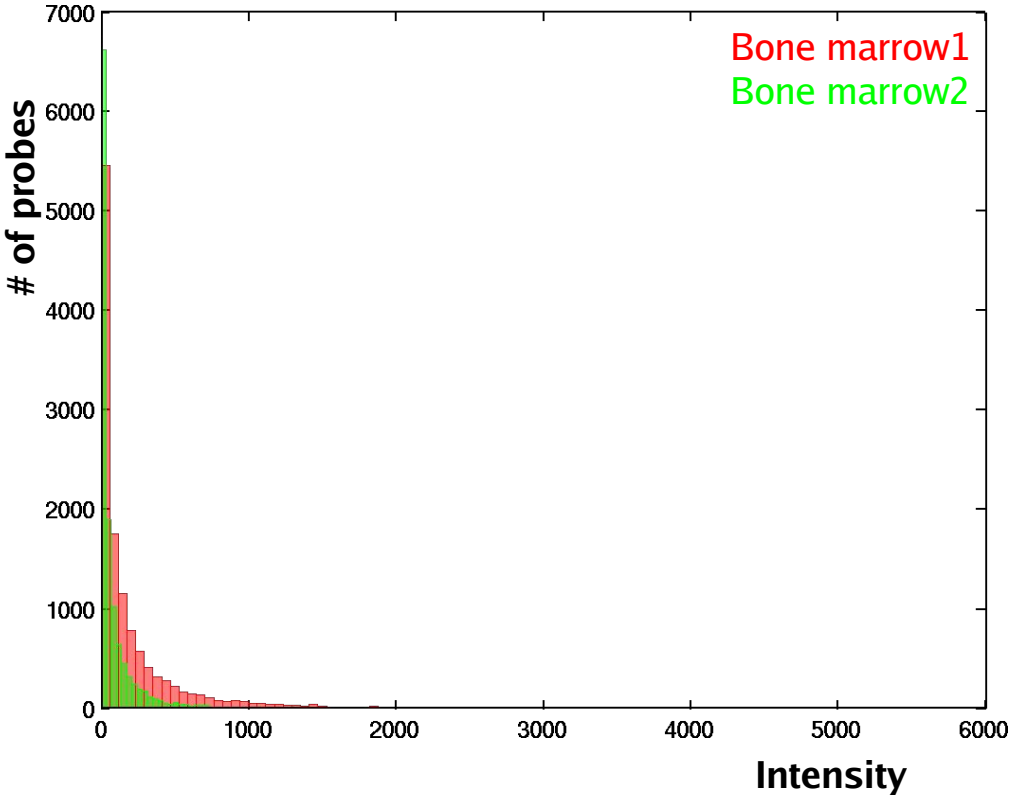
- Some modes of data representation:
  - PCA (Euclidean) - finite bases. Rotation/translation.
  - Fourier (Euclidean) - infinite bases (localize “freq” domain). Signal decomposed into sinusoids. Periodic boundary conditions.
  - Wavelet (Euclidean) - infinite bases (localize “time” domain). Signal decomposed by discretized amplitudinal range.
- Different approach emphasizes different regular structures within the data. There is almost always a geometric interpretation.
- Secondary uses: Feature reduction, de-noising, etc.

FG: Native distributions of large scale expression measurements

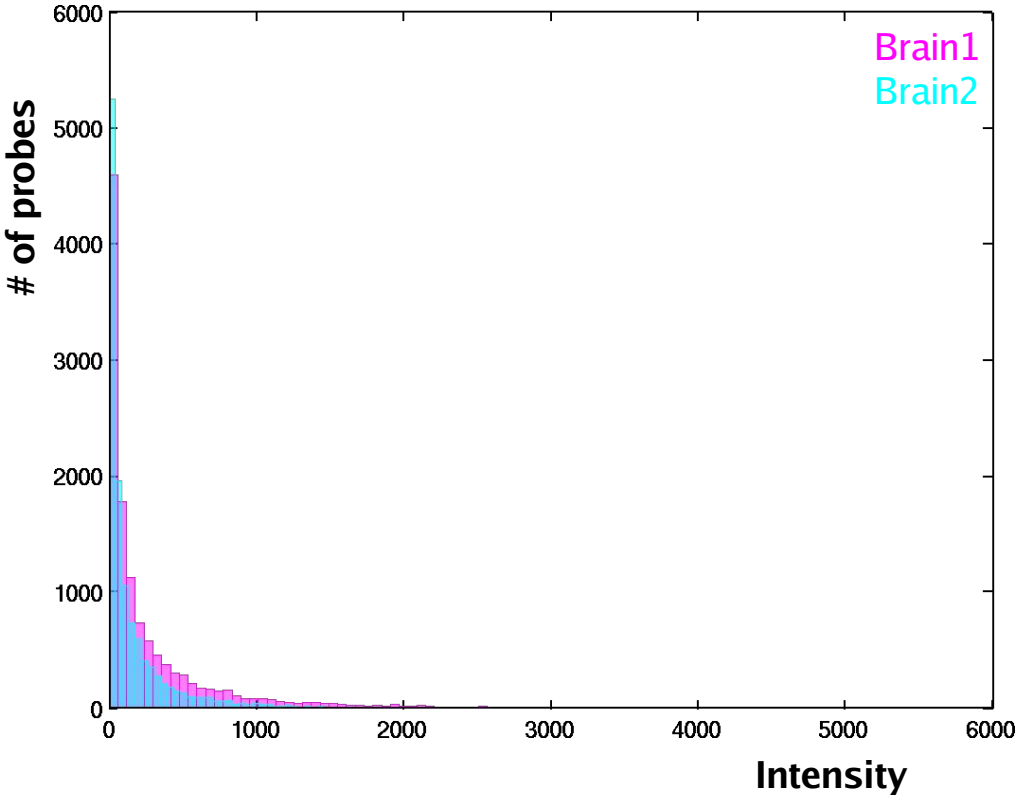
- Recall data matrix D = N genes × M samples / conditions. Do D entries have units (dimensioned), or not (dimensionless, say fold in 2-channel microarray read-outs)?
- Consider, the transcriptome profile of a sample (samples in gene space view).
  - 2 basic questions (not well-defined):
    - Is there a generic form for distribution of microarray data (independent of technology, bio-system/state)? If yes, what are it's characteristics?
    - Can these characteristics be used for quality control, or provide info about underlying biologic properties / mechanism?
  - Power law, log normal - Hoyle et al., Bioinformatics 18 (4) 2002 *Making sense of microarray data distributions*.
  - Lorentz - Brody et al. PNAS 99 (20) 2002 *Significance and statistical errors in the analysis of DNA microarray data*.
  - Gamma – Newton et al, J Comput Biol 8 2001 *Improving statistical inferences about gene expression changes in microarray data*.
  - Application: Characteristics of these distributions used for rational design of a null hypothesis for evaluating significance of intrinsic data regularities (defined later)

FG: Native distributions of large scale expression measurements

- Pictures of distributions human bone marrow (n=2 sample *replicates*), human brain (n=2 sample *replicates*). All on Affymetrix u95Av2 platform



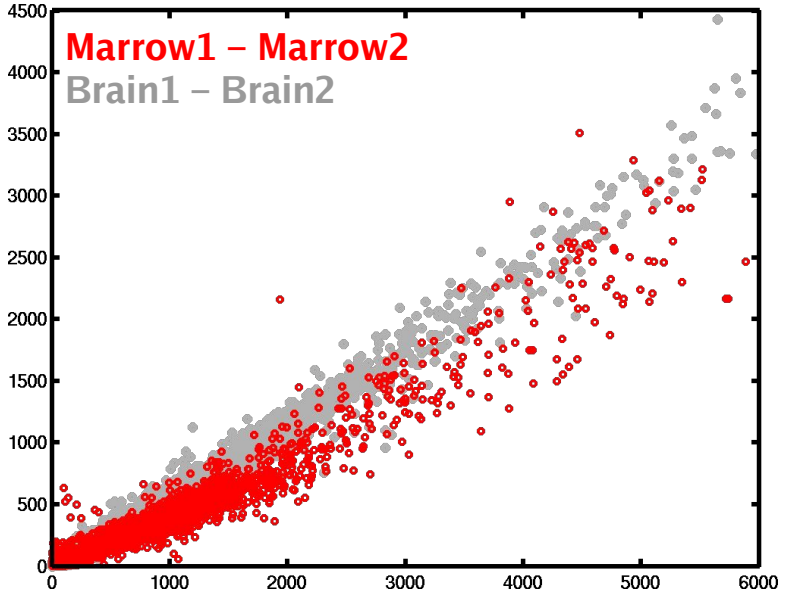
Bone marrow
2 separate histograms

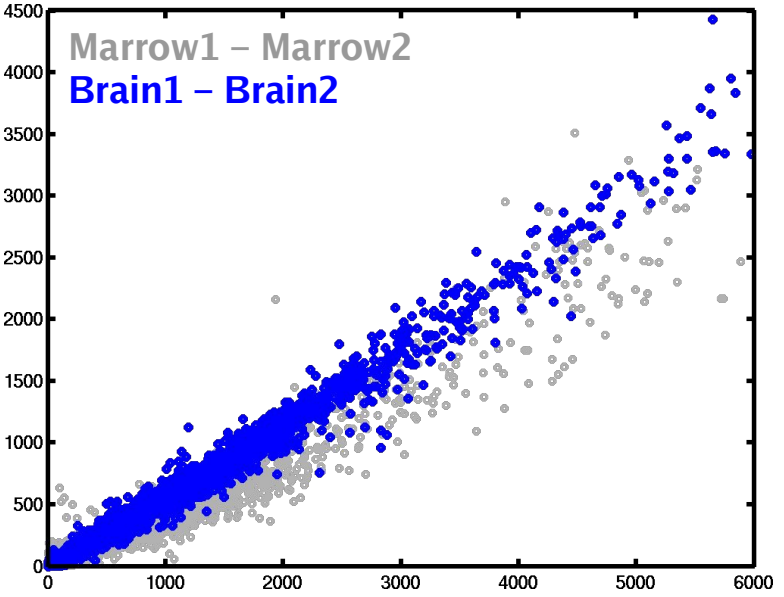


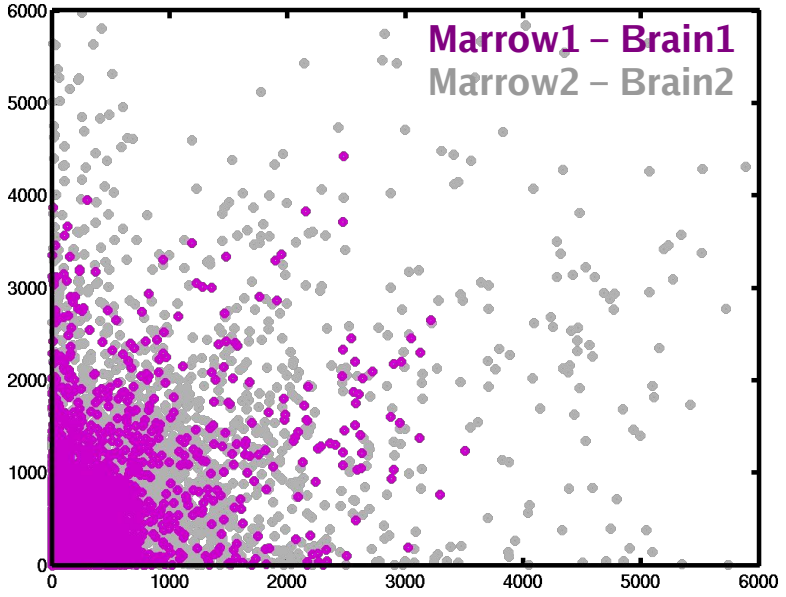
Brain
2 separate histograms

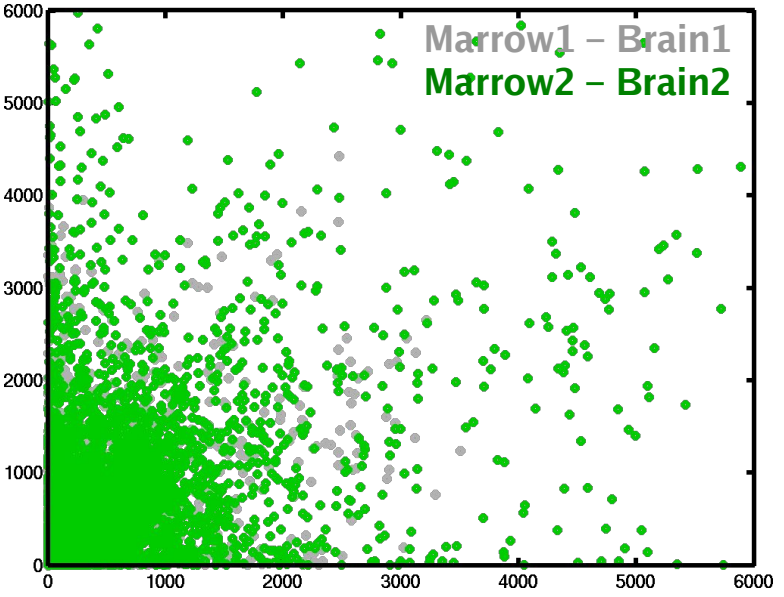
FG: Native distributions of large scale expression measurements

- Another view: Intensity-intensity scatter plots intra- and inter- tissue. Same data as previous slide







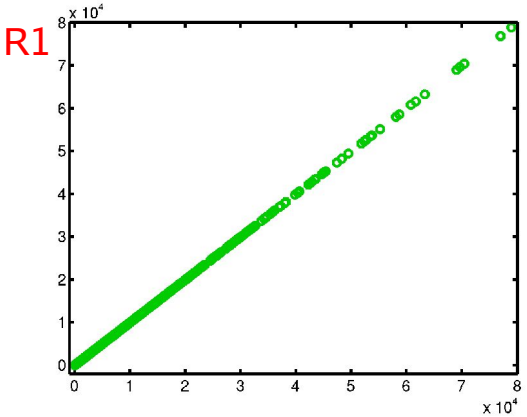


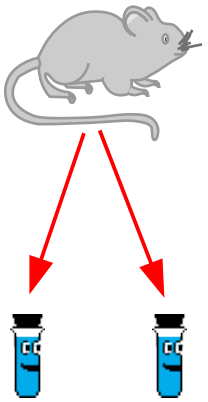
FG: Characterizing and correcting noise

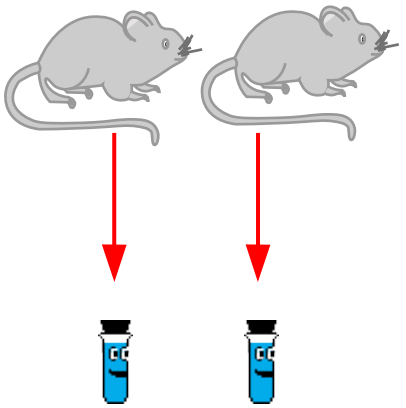
- *Noise* are variations from logical / math / scientific assumption that is expressed in the data.
  - **N1** Example of a scientific assumption: Closely timed repeat measurements of a system – state are similar (in a given metric space). *Nature does not make leaps* (Leibniz, New Essays)
  - **N2** Conceptual sources of noise: failure to account for all parameters affecting a system – state, basic assumption is ill-defined.
  - **N3** Technical sources of noise: biological heterogeneity, measurement variation, chip micro-environment / thermodynamics
- First characterize noise, then correct for it based on characteristics (generalization)
  - Characterization: Depends on N1-3, esp. N1. Practically from N3, we design replicate experiments to asssess the level of measurement / technical and biological (more complex) variation. Some chips have built-in housekeeping probes – benchmark dataset.
  - Correction: Via "pre-processing" (data from a single chip independent of other chips) or normalization (data across different chips) – essentially math transformations that aim to minimize noise while preserving expression variations that arise from biologically relevant transcriptome activity [... vague]
  - Question: Intensity distribution over all probes (2-3 slides ago) related to variations of a probe?

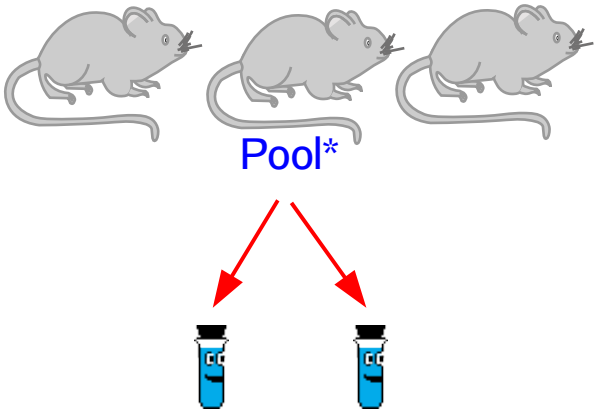
FG: Characterizing noise – replicates

- Different levels of being a "Replicate" experiment
  - Intensity-intensity scatter plots of chip read-outs between "Replicates"
  - These data used to characterize noise (eg. define a noise range / interval) – many standard ways to do this

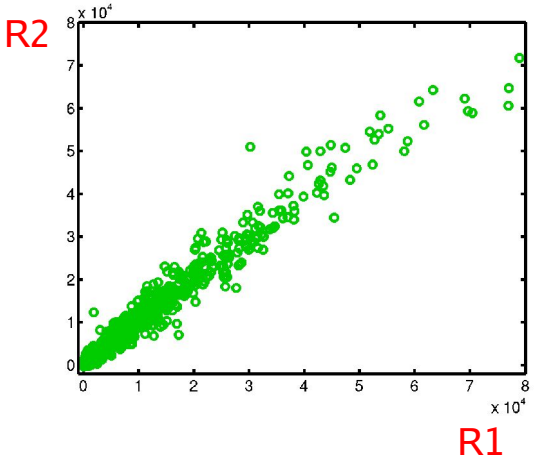




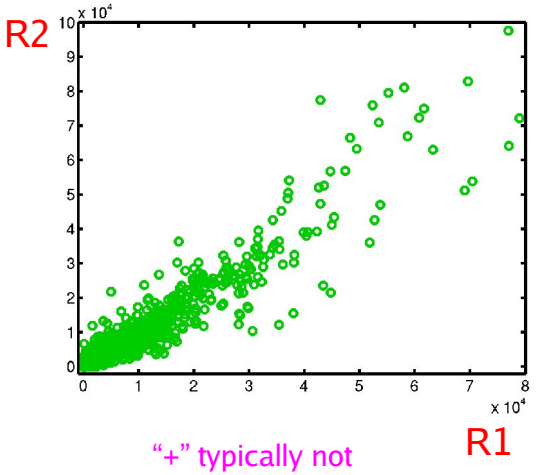




Measurement variation

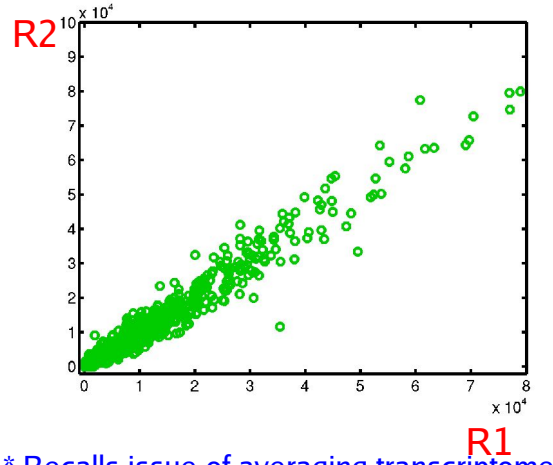


Biological variation "+"
Measurement variation



"+" typically not linear / additive

Measurement variation



* Recalls issue of averaging transcriptome of heterogeneous cell populations

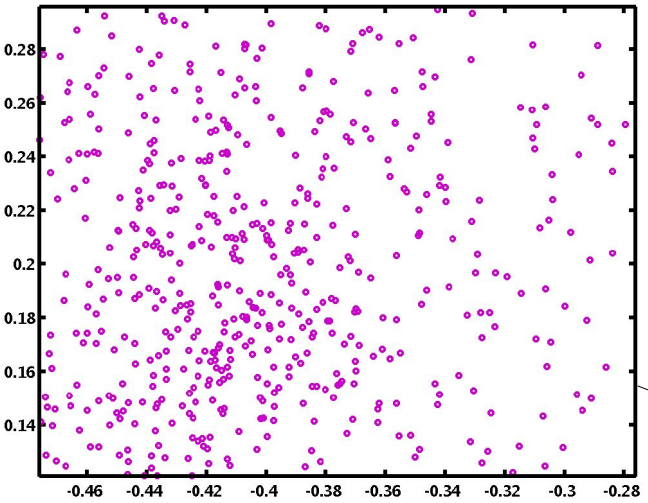
FG: Correcting noise – pre-processing, normalization

- Math transformations that aim to minimize noise while preserving expression variations that arise from biologically relevant transcriptome activity
  - "Pre-processing" (data from a single chip independent of other chips)
  - Normalization (data across different chips). We'll use this to refer to pre-processing too.
  - House-keeping probes on some chips – benchmark dataset to characterize chip technical micro environmental variations
- Frequently used transformations
  - "Standardization" / central-limit theorem scaling (CLT) – mean center , divide by 1 std
  - Linear regression relative to a Reference dataset. Corollary: All normalized samples have same mean intensity as Reference.
  - Lowess, spline-fitting, many others – all with their own distinct assumptions about characteristics of noise and array-measured intensity distribution.

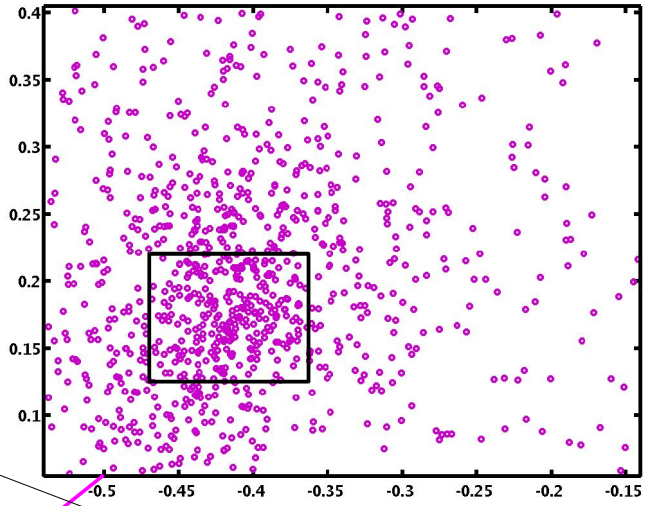
FG: Regularities in data

- A graphical example, intensity-intensity plot of a dataset at multiple scales. 3 questions:
  - Criteria for regularity?
  - Likelihood of such regularities arising by random confluence of distributions?
  - Biological relevance?

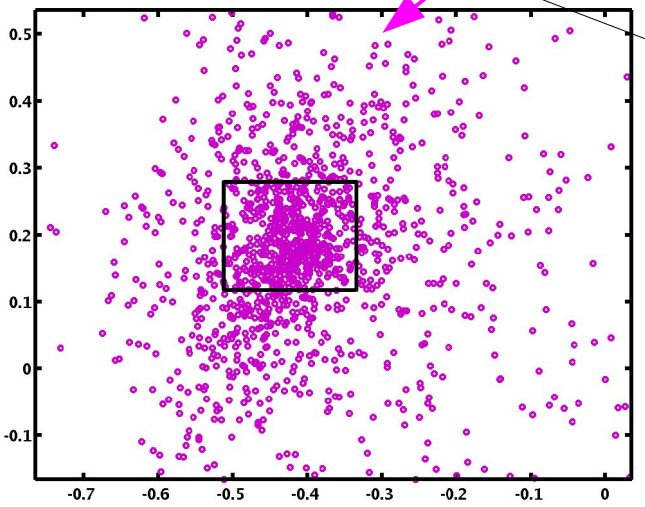
ie. How do we know that we are not hallucinating?



Zoom out

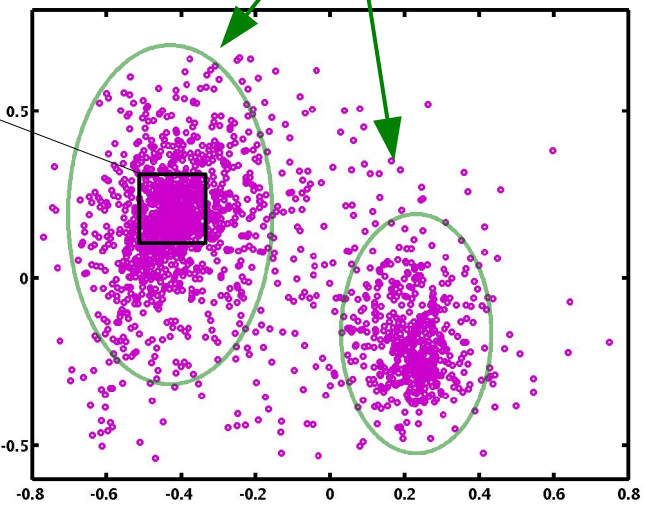


Zoom out



Zoom out

Regularities, coherent structures.



FG: Regularities in data

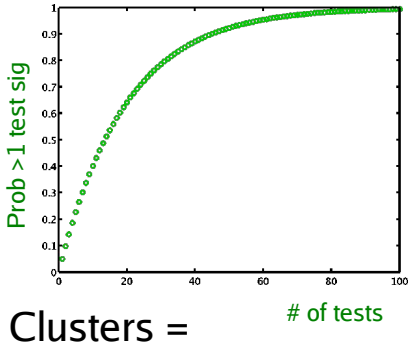
- *Regularities* refer to dominant variance structures or coherent geometric structures intrinsic to data with respect to a particular measure space. An observed pattern can sometimes be regarded a regularity if it can be correlated to *a priori* scientific knowledge.

- Recall a data matrix D = N × M may be view as genes in sample space, or samples in gene space. Questions:

  - Given any D in a specific metric space, do regularities exist? Eg. If D was considered a linear transformation $\mathbb{R}^M \rightarrow \mathbb{R}^N$, it would make sense to look for invariants of D (eigenvectors)

  - Why would anyone look for regularities?

  - How to look for regularities? ... cluster-finding algorithms? What is a "cluster"? Clusters = regularities?



Multiple tests

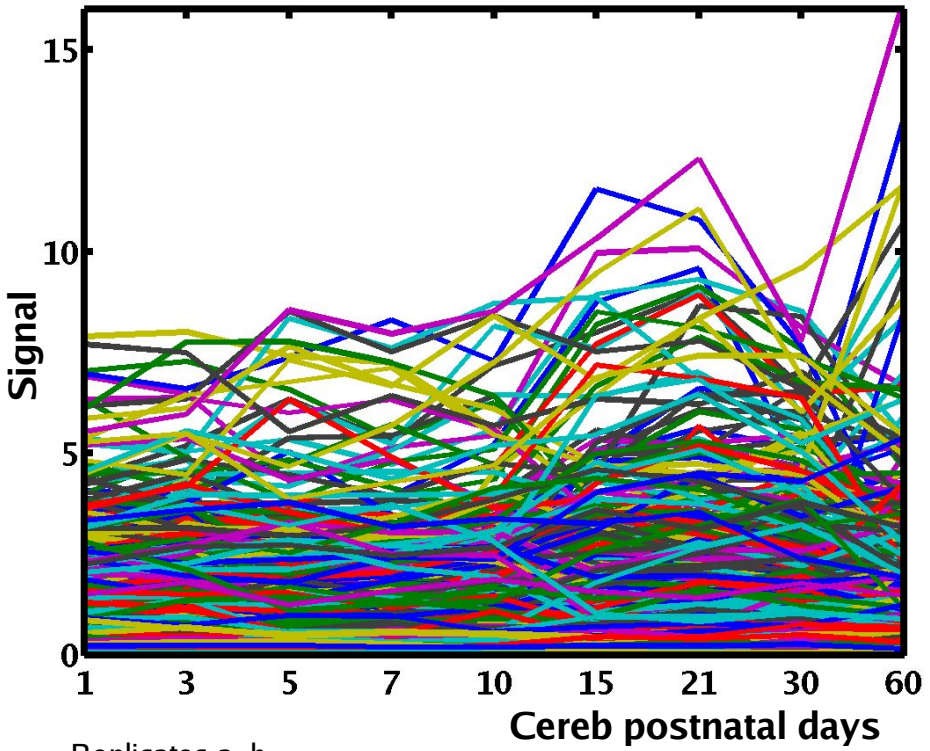
- Now we arrive at heart of analysis proper: Supervised versus unsupervised techniques,

  - **Supervised**: *a priori* scientific knowledge is explicit input parameter into computation, eg. ANOVA ranksum test, t test.

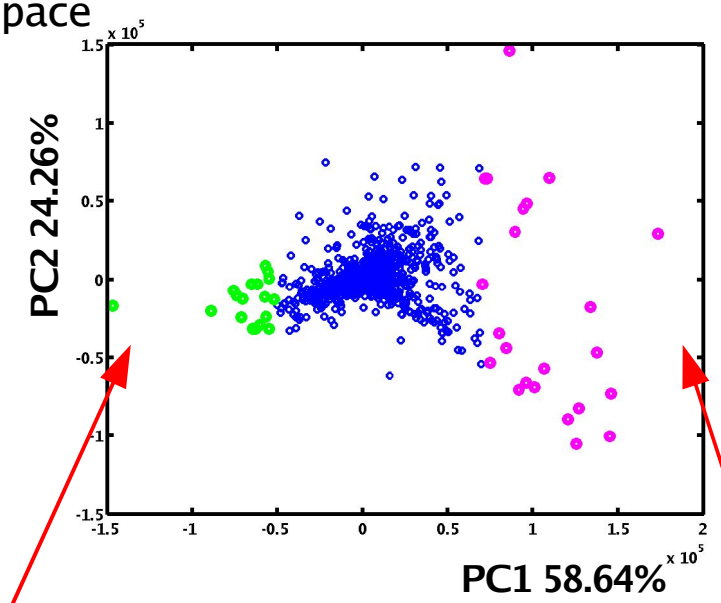
  - **Unsupervised**: Input parameters do not explicitly include *a priori* scientific knowledge, eg. clustering algorithms such as k-means, ferromagnetic (Ising-Potts)

FG: Regularities in data example

- Example: Mouse cerebellum postnatal development, data matrix D = 6,000 genes × 9 time stages (measurement duplicated at each stage, so really 6,000 × 18)
  - Genes in sample space view 1. Euclidean space

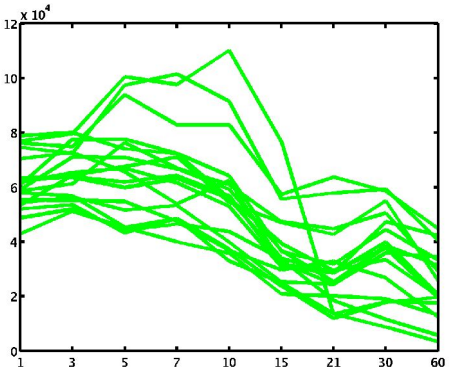


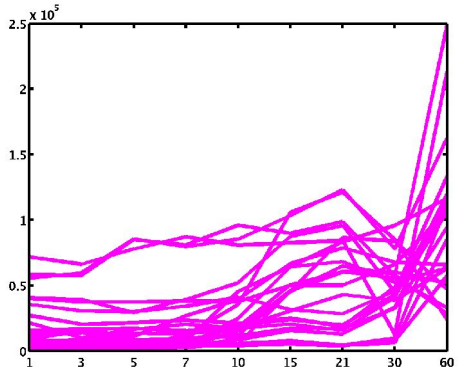
Replicates a, b @each stage



PCA* representation

Coherent structures / regularity here?

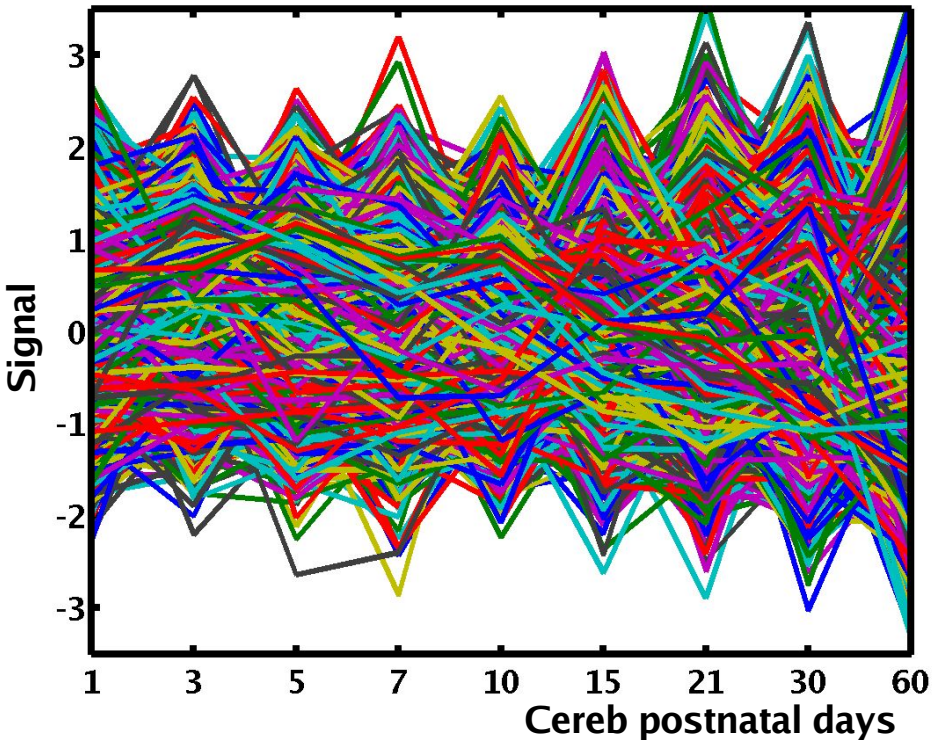




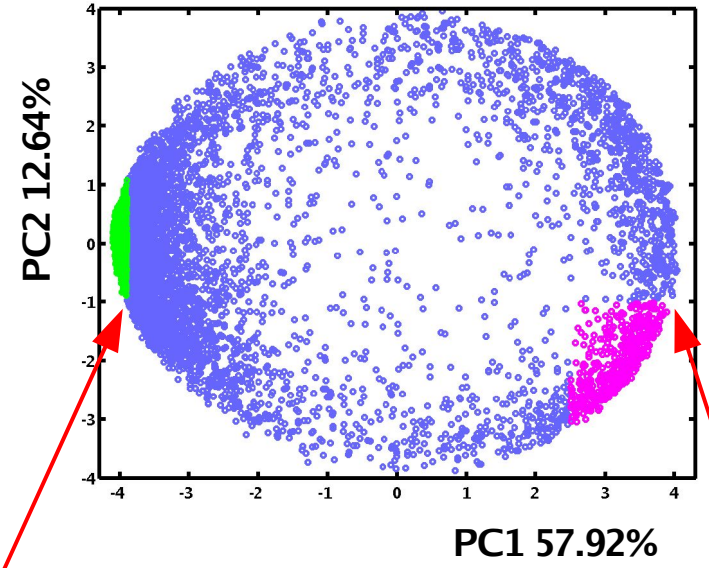
*PCA = principal component analysis, singular value decomposition

FG: Regularities in data example

- Same data as previous slide. Each gene is standardized / CLT normalized across conditions.
  - Genes in sample space view 2. Correlation space

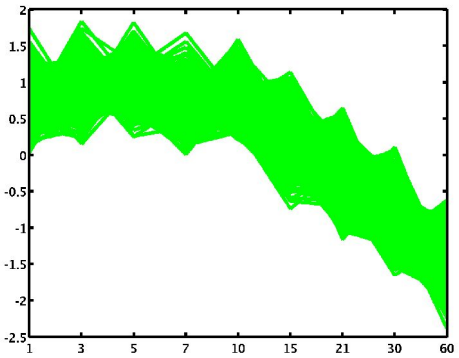


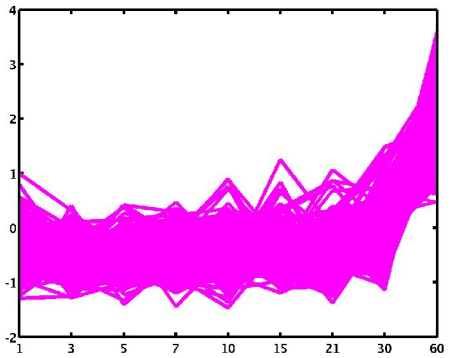
Replicates a, b
@each stage



PCA* representation

Coherent structures / regularity here?

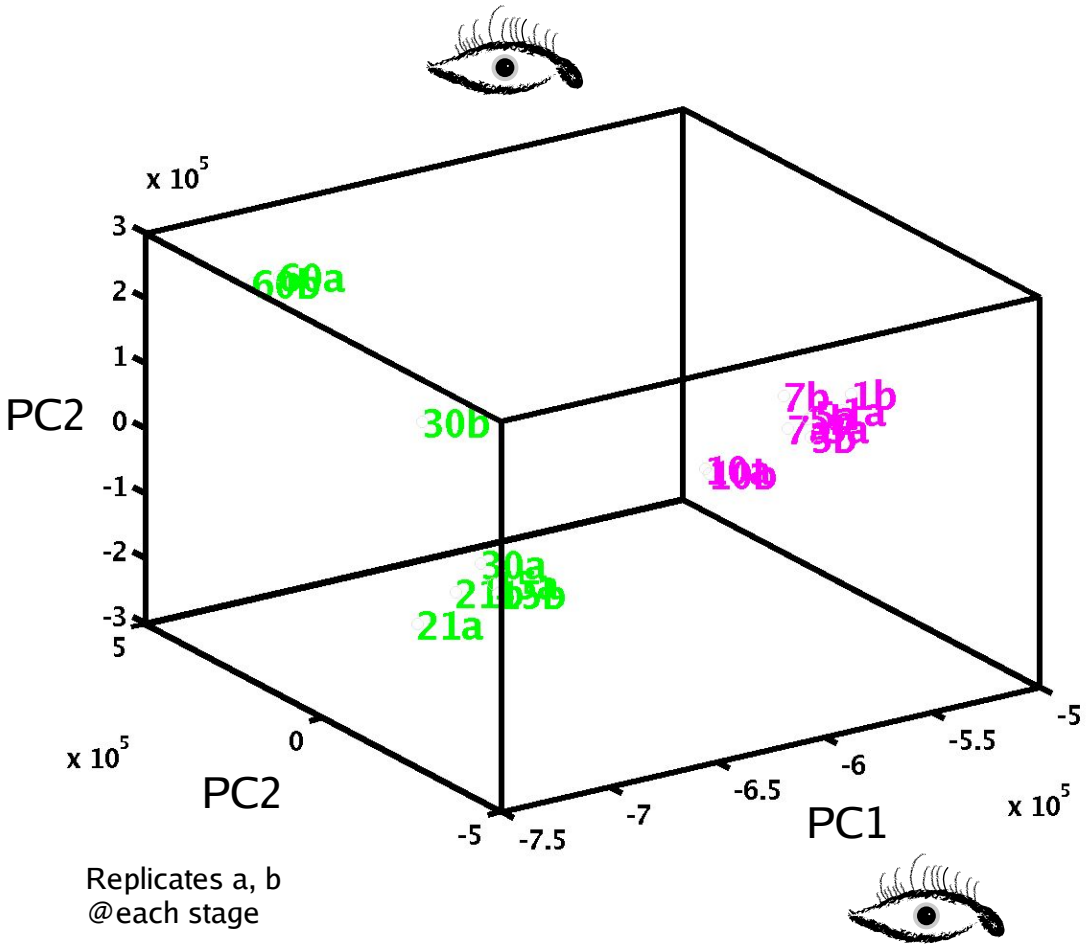


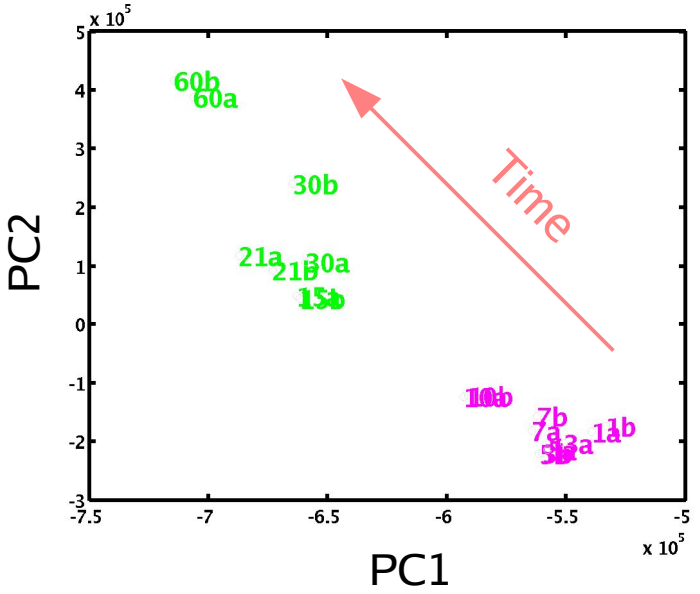


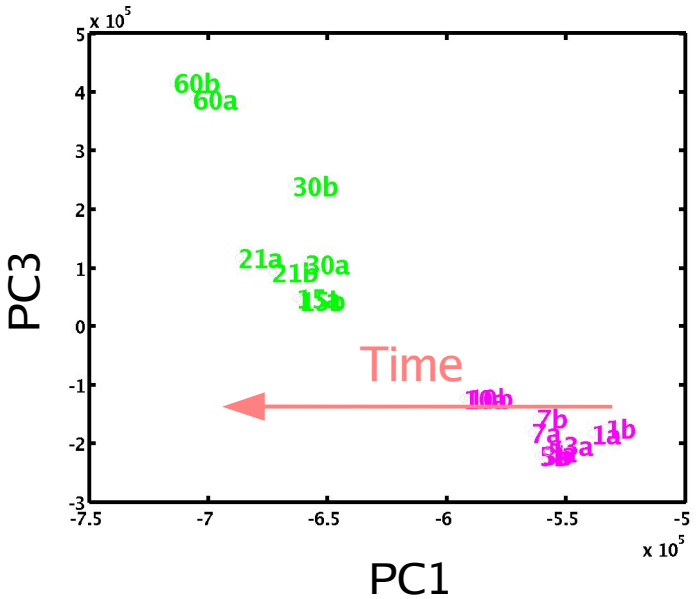
*PCA = principal component analysis, singular value decomposition

FG: Regularities in data example

- Same data as previous slide. The dual to genes in sample space view 1
  - Samples in gene space view 1. Euclidean space





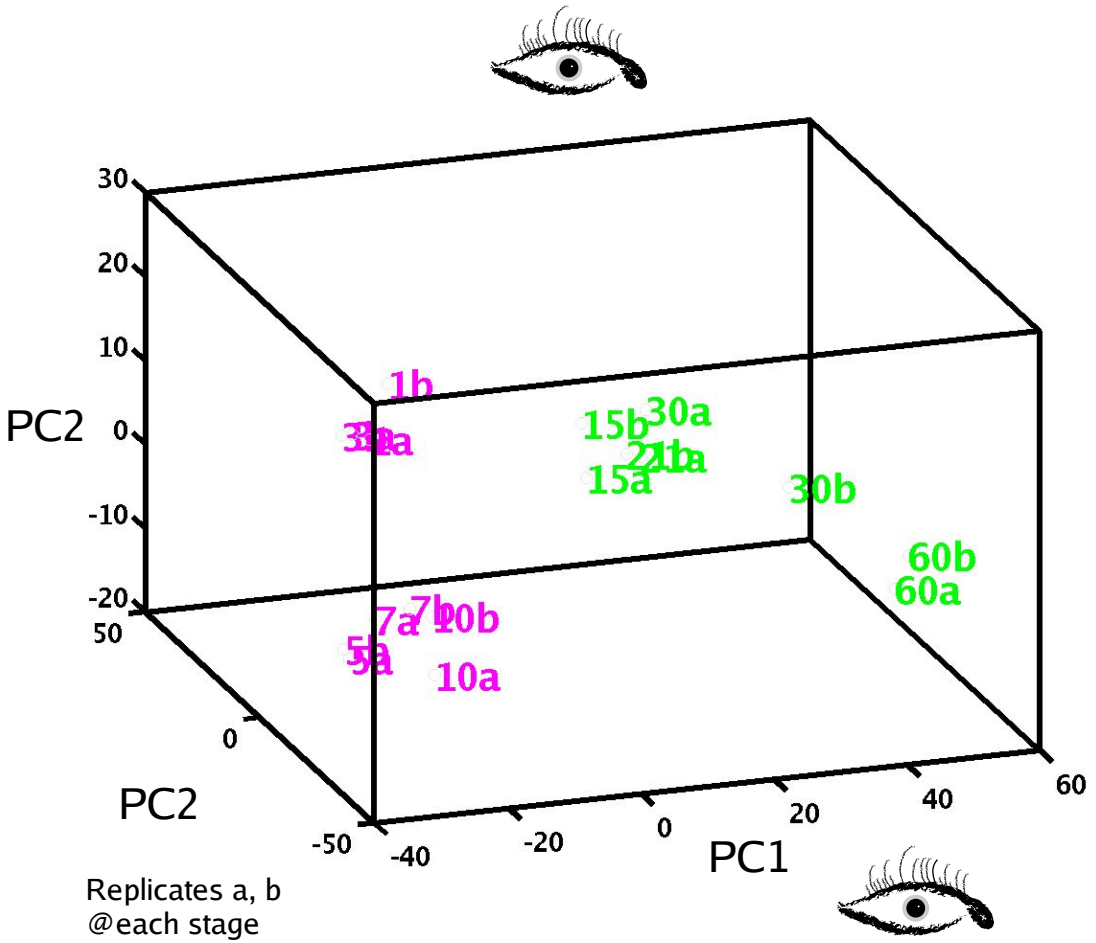


Do configurations say anything bio relevant?

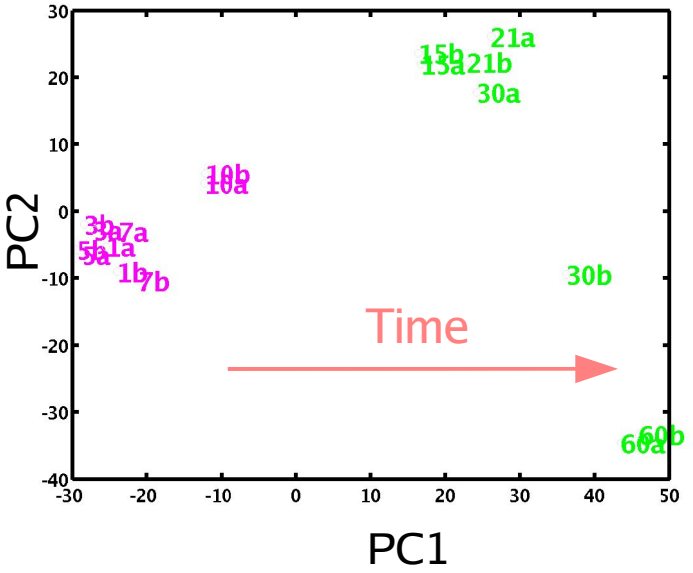
*PCA = principal component analysis, singular value decomposition

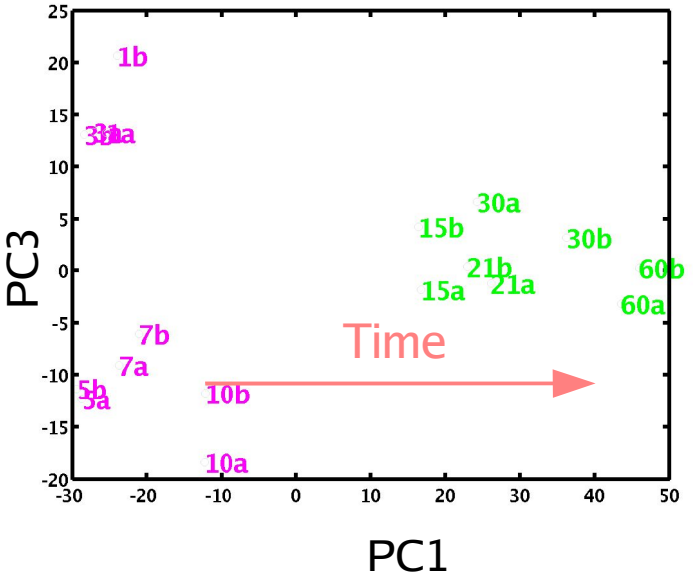
FG: Regularities in data example

- Same data as previous slide. Each sample is CLT normalized across genes.
  - Samples in gene space view 2. Correlation space



Replicates a, b
@each stage





Do configurations say anything bio relevant?

*PCA = principal component analysis, singular value decomposition

FG: Reality check and figures of merit in modeling

- Unsupervised and supervised analyses. What is a cluster? Clustering dogma. Next lecture
- Statistical significance of computed / observed regularity
  - Need null hypothetic distribution as reference system to assess significance. Standard statistics.
  - Non-parametric: Permutation tests. Mode of permutation modulated by assumptions about random action / effects (null hypothesis ... modeling chance) producing a given transcriptome profile / state. Re-analyze permuted data following same approach as for unpermuted, and watch for regularities.
- Figure/s of merit:
  - 1 biological system – state $\rightarrow$ 1 transcriptome profile dataset $\rightarrow$ >1 possible models. Which model is optimal?
  - For each model, perform sensitivity / specificity analysis (ROC curves) on *simulated data* (drawn from *putative* distribution of actual data) which have regularities explicitly embedded
  - Prediction on independent test dataset, coupled with biological knowledge.Overfitting and non generalizability

Module 4 FG Lecture 2 outline: Large-scale transcriptome data analysis

- Murray Gell-Mann, *The Quark and the Jaguar*: “Often, however, we encounter less than ideal cases. We may find regularities, predict that similar regularities will occur elsewhere, discover that the prediction is confirmed, and thus identify a robust pattern: however, it may be a pattern for which the explanation continues to elude us. In such a case we speak of an "empirical" or "phenomenological" theory, using fancy words to mean basically that we see what is going on but do not yet understand it. There are many such empirical theories that connect together facts encountered in everyday life.”

  Meaning of regularities

- Comte de Buffon, “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.”

  Point of mathematical formulation