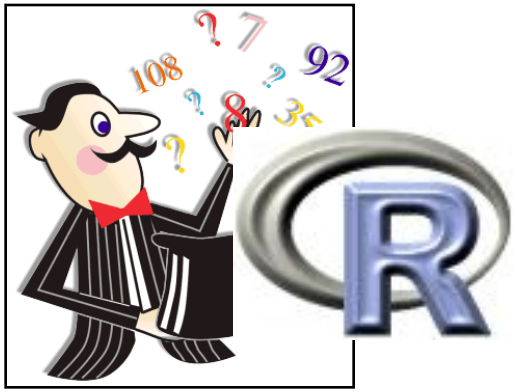


Exploratory data analysis

Aedín Culhane
aedin@jimmy.harvard.edu

Dana-Farber Cancer Institute/Harvard School of Public Health.

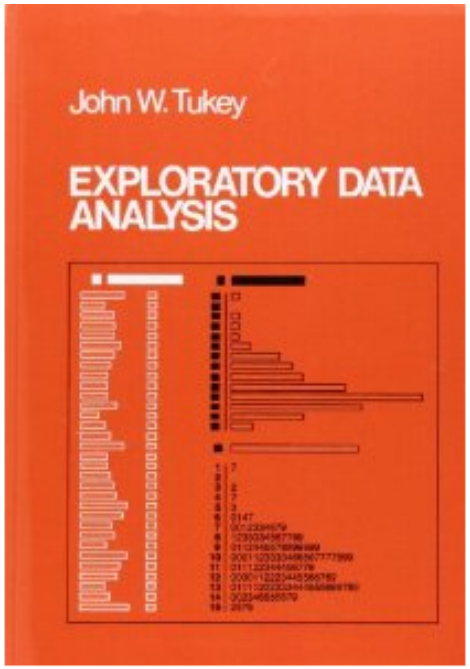
What is EDA?

- **Exploratory data analysis** (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

- using the “picture-examining eye” (Tukey)

Tukey; A father of EDA.

> The greatest value of a picture is when it forces us to notice what we never expected to see.
>
> — John W. Tukey. Exploratory Data Analysis. 1977.



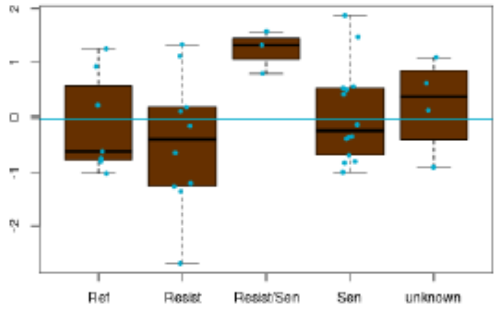
Exploratory Data Analysis [Paperback]
John W. Tukey (Author)
(8 customer reviews)

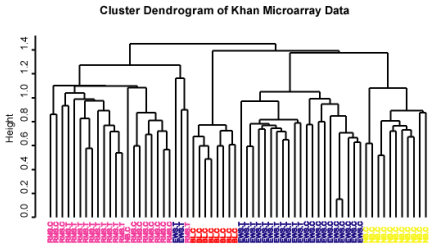
Exploration of Data is Critical

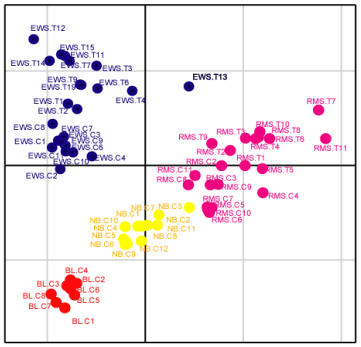
- EDA should take place before more rigorous statistical analysis. 1st part of larger process
- First line of defense against bad data, use to check assumptions about data.
- Maybe lead to new insights, new questions or feed into process of building predictive models

Exploration of Data is Critical

- Detect unpredicted patterns in data
- Decide what questions to ask
- Can also help detect cofounding covariates







What do with a new dataset? – Jeef Leek

Blog: http://simplystatistics.org/

What I do when I get a new data set as told through tweets

Posted on June 13, 2014 by Jeff Leek

Hilary Mason asked a really interesting question yesterday:

What do with a new dataset? – Jeff Leek

- **Step 0: Figure out what I'm trying to do with the data**
  - “Look, Stop, Think…”,
  - Check-in with person generating data

- **Step 1: Learn about the elephant**
  - figure out what the data set "looks" like: head(), tail, sapply(df, class)
  - look for NA,
  - check for personally identifiable information
  - colnames(), summary(), str() etc

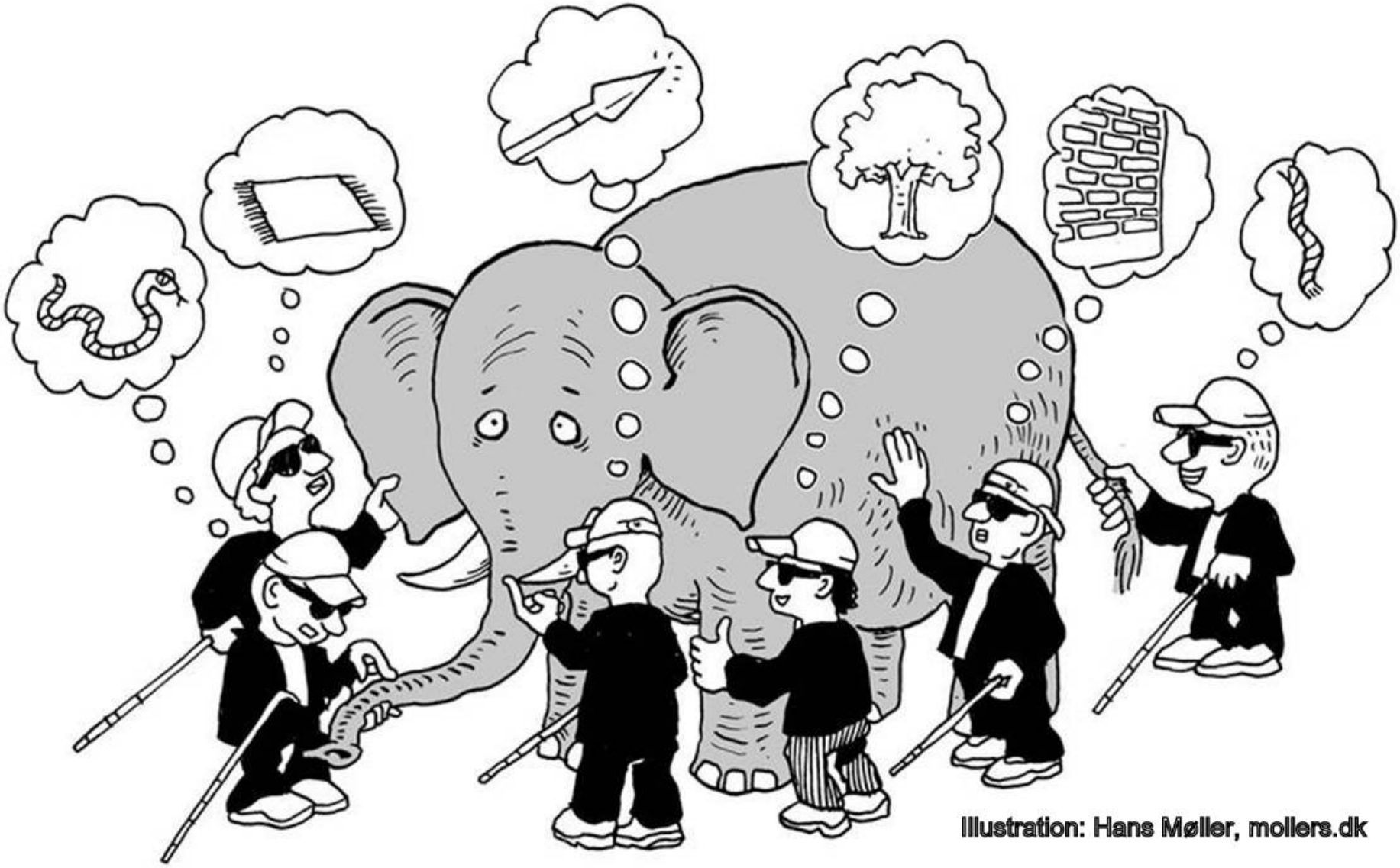


Illustration: Hans Møller, mollers.dk

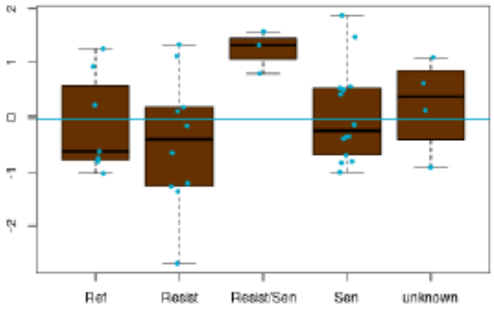
What do with a new dataset? – Jeff Leek

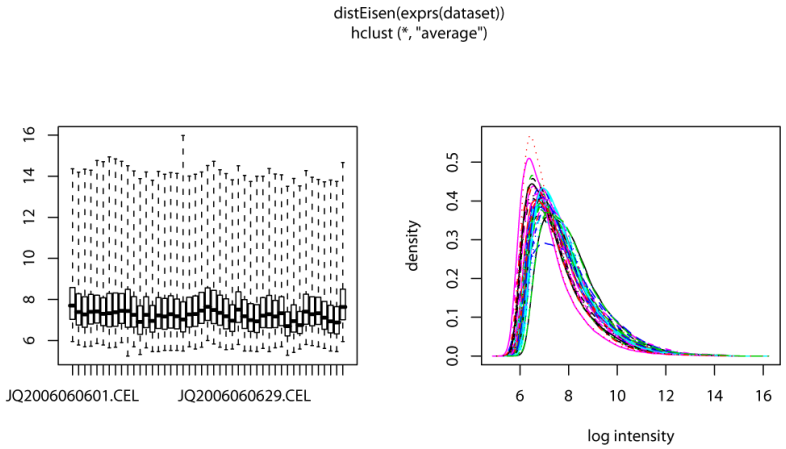
- **Step 0: Figure out what I'm trying to do with the data**
- **Step 1: Learn about the elephant**
- **Step 2: Clean/organize**
  - Fix data, NA,
  - swear a lot

- **Step 3: Plot. That. Stuff**
  - look at variables one by one like
    - Histograms,
    - scatterplots,
    - density plots,
    - jittered 1d plots
  - If data are multivariate ,get a feel for high dimensional structure
    - dimension reduction : principal components,
    - hierarchical clustering analysis

Visualize variables one by one

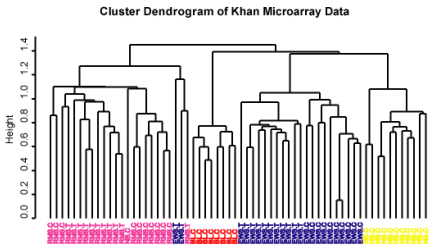
- Histograms, density plots
- Boxplots
- Scatterplots
- MA plots (variance v mean)
- Pairs
- Correlations, Corrplot



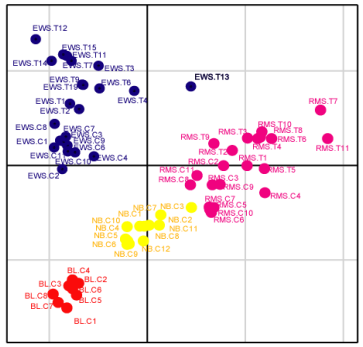


Get a feel for high dimensional structure of multivariate data

- Clustering
  - Hierarchical
  - Flat (k-means)



- Ordination (Dimension Reduction)
  - Principal Component analysis, Correspondence analysis



What do with a new dataset? – Jeff Leek

- **Step 0: Figure out what I'm trying to do with the data**
- **Step 1: Learn about the elephant**
- **Step 2: Clean/organize**
- **Step 3: Plot. That. Stuff**
- **Step 4: Get a quick answer to the question from Step 1**
  - quick and dirty answer to the question;
  - simple predictive model or a really basic regression model.
  - Check back with person generating data.

Exercise 1

- Take Duncan Dataset
  - Take all numeric variables
  - Pairs plot
  - Do Corrplot
  - Boxplot
  - Density plot

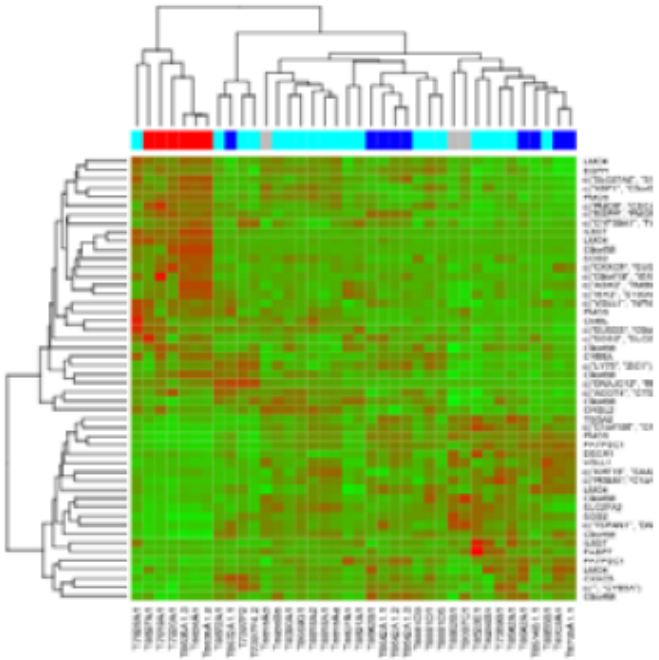
Cluster Analysis

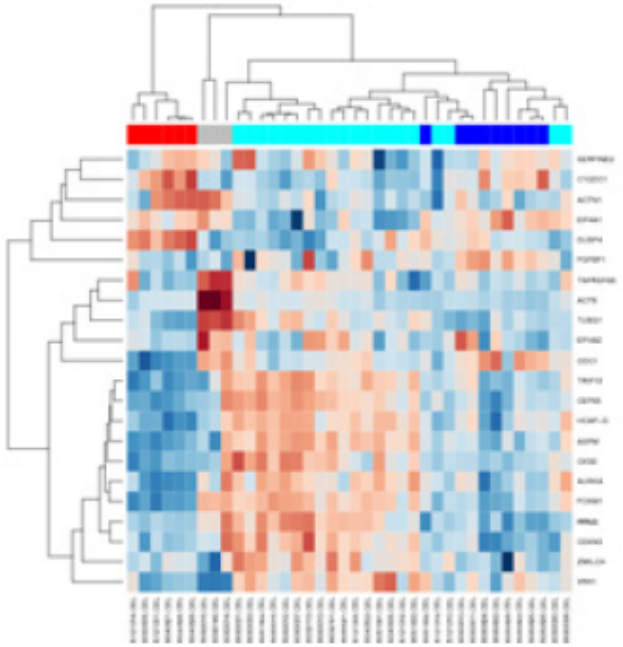
`dist()`

`hclust()`

`heatmap()`

`library(heatplus)`





Clustering Algorithms

- Clusters can be generated using agglomerative or divisive methods
  - **divisive** k-means
  - **hierarchical agglomerative clustering**

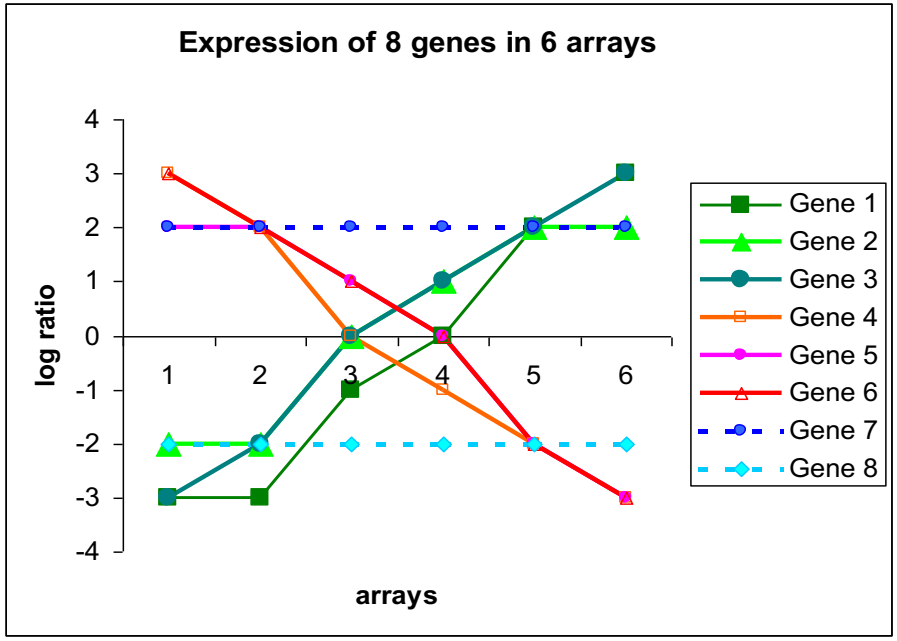
- The clusters are dependent on the **distance** measure used

A Distance Metric

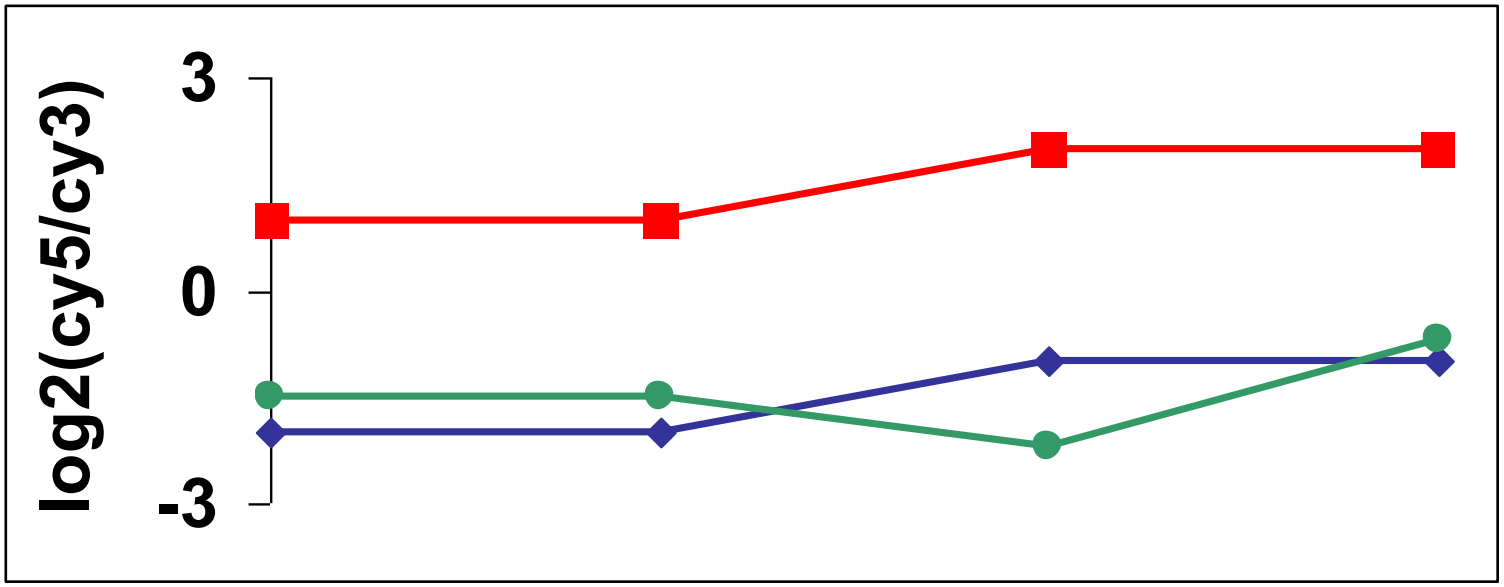
- In exploratory data analysis
  - only discover where you explore..

- The choice of metric is fundamental
  - no right or wrong distance measure
  - it depends on what relationships you wish to visualize

8 Genes: Which is “closest”?



Distance Is Defined by a Metric

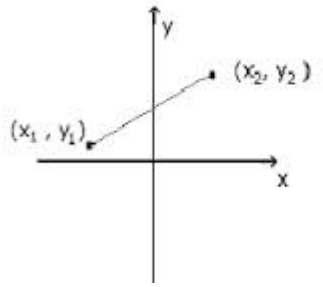


Distance Metric:	Euclidean	Pearson*
– D →	1.4	-0.05
– D →	6.0	+1.00

Clustering: Distance metrics

- **Euclidean distance**
  - A commonly used measure
  - Ruler distance

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

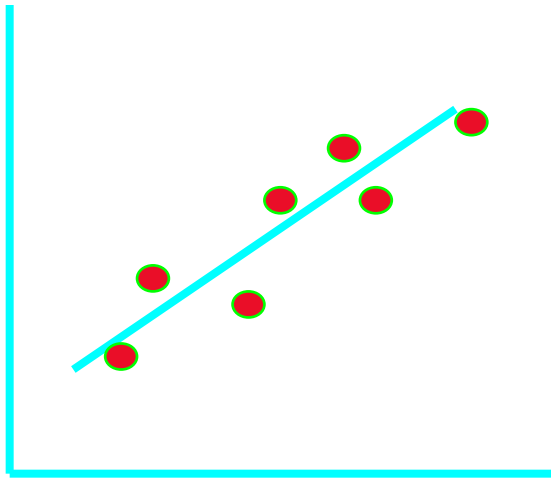


Pythagorean Theorem

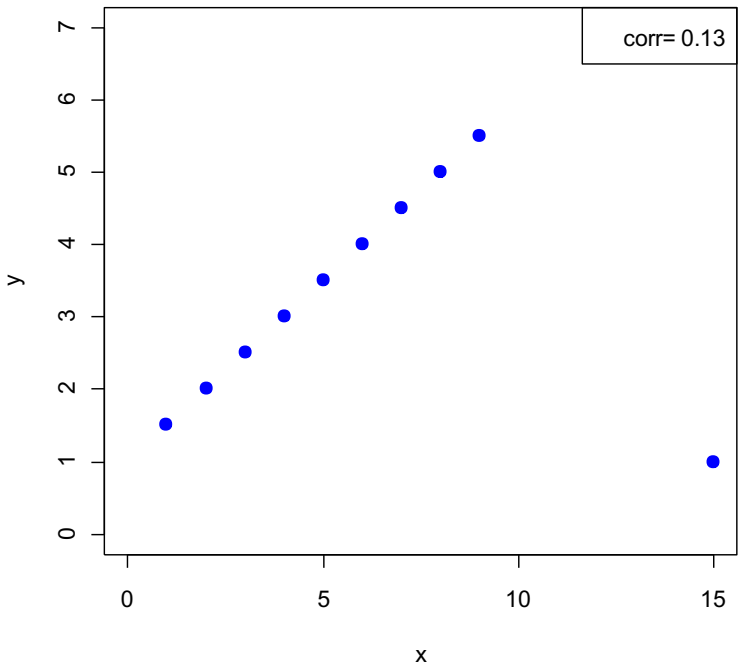
$$Euclidean\ Distance = d = \sqrt{\sum_{i=1}^{N} (Xi - Yi)^2}$$

Clustering: Distance metrics

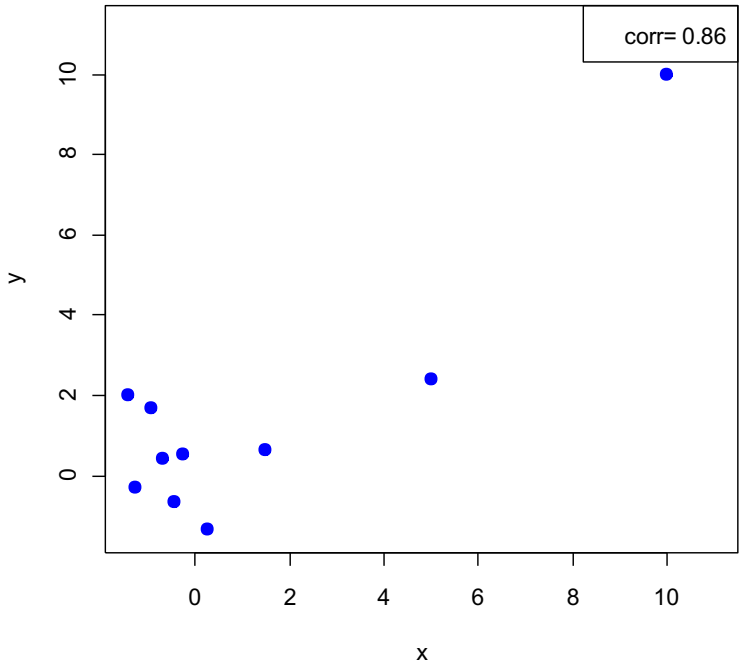
- **Pearson Correlation Coefficient**
  - A commonly used measure
  - As its a similarity measure, for distance = 1- PCC



Correlations gone wrong

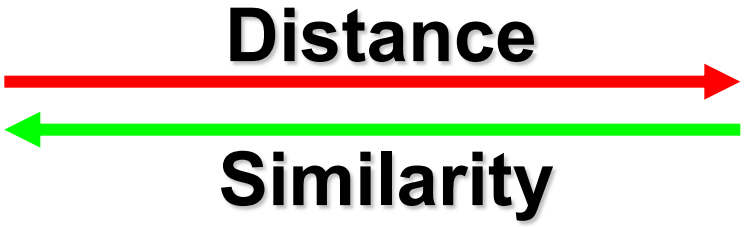


y=x/2+1



Random Noise
rnorm(10, mean=1, sd=2)

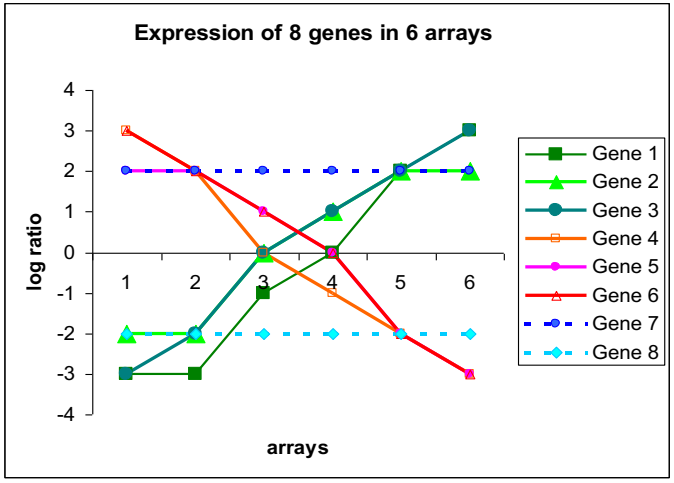
Distance Metrics



- **Euclidean distance**
- **Pearson correlation coefficient**
- **Spearman rank**
- **Manhattan distance**
- **Mutual information**
- **etc**

Each has different properties and can reveal different features of the data

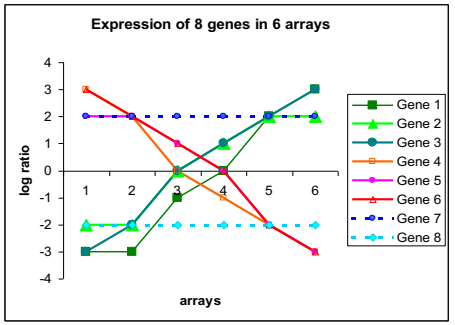
Which is “closest”?



	Expt1	Expt2	Expt3	Expt4	Expt5	Expt6
Gene 1	-3	-3	-1	0	2	3
Gene 2	-2	-2	0	1	2	2
Gene 3	-3	-2	0	1	2	3
Gene 4	3	2	0	-1	-2	-3
Gene 5	2	2	1	0	-2	-3
Gene 6	3	2	1	0	-2	-3
Gene 7	2	2	2	2	2	2
Gene 8	-2	-2	-2	-2	-2	-2

$$Euclidean\ Distance = d = \sqrt{\sum_{i=1}^{N}(Xi - Yi)^2}$$

Which is "closest"?



- Calculate $$Euclidean\ Distance = d = \sqrt{\sum_{i=1}^{N}(Xi - Yi)^2}$$

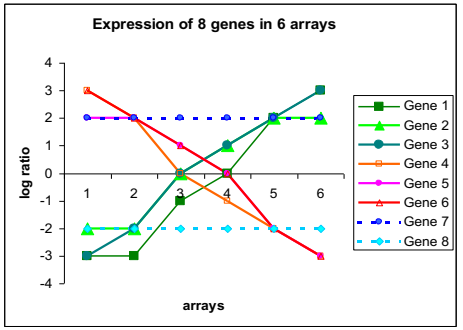
	Expt1	Expt2	Expt3	Expt4	Expt5	Expt6
Gene 1	-3	-3	-1	0	2	3
Gene 2	-2	-2	0	1	2	2

$\sqrt{5} = 2.24$

	Expt1	Expt2	Expt3	Expt4	Expt5	Expt6
Gene 1	-3	-3	-1	0	2	3
Gene 3	-3	-2	0	1	2	3

$\sqrt{3}=1.73$

Distance Matrix



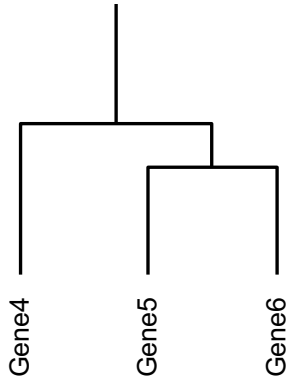
	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7	Gene 8
Gene 1	0	2.24	1.73	10.72	10.30	10.82	8.00	6.93
Gene 2		0	1.41	9.27	8.66	9.17	6.08	6.71
Gene 3			0	10.39	9.75	10.30	6.86	7.42
Gene 4				0	1.73	1.41	7.42	6.86
Gene 5					0	**1**	6.78	6.78
Gene 6						0	6.86	7.42
Gene 7							0	9.80
Gene 8								0

5,6 have the min distance between them (dist = 1)

Join 5,6

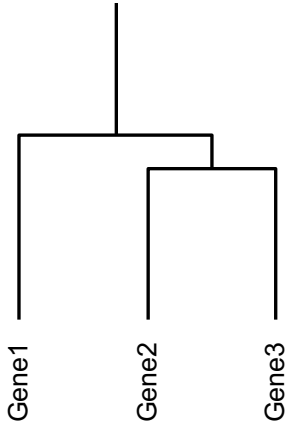
	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5,6	Gene 7	Gene 8
Gene 1	0	2.24	1.73	10.72	10.30	8.00	6.93
Gene 2		0	**1.41**	9.27	8.66	6.08	6.71
Gene 3			0	10.39	9.75	6.86	7.42
Gene 4				0	**1.41**	7.42	6.86
Gene 5,6					0	6.78	6.78
Gene 7						0	9.80
Gene 8							0

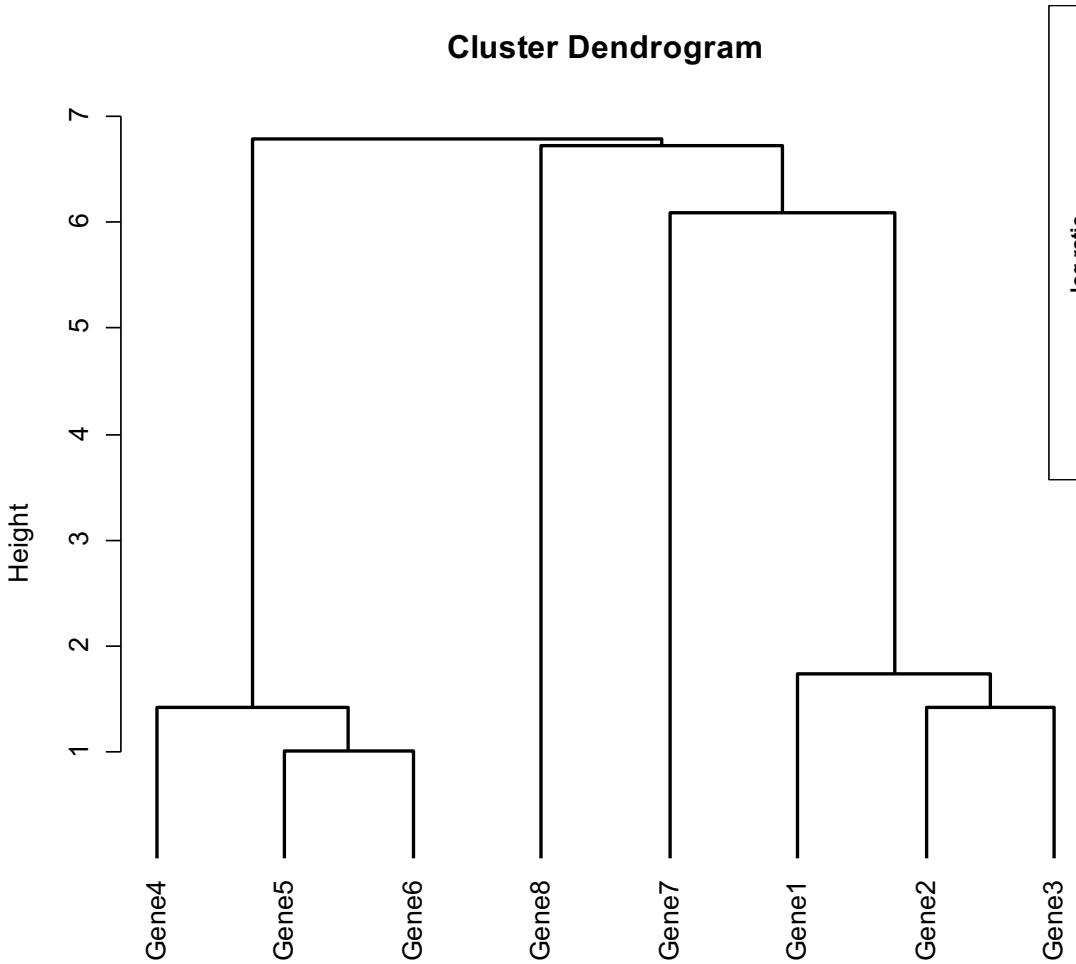
	Gene 1	Gene 2,3	Gene 4	Gene 5,6	Gene 7	Gene 8
Gene 1	0	1.73	10.72	10.30	8.00	6.93
Gene 2,3		0	9.27	8.66	6.08	6.71
Gene 4			0	**1.41**	7.42	6.86
Gene 5,6				0	6.78	6.78
Gene 7					0	9.80
Gene 8						0

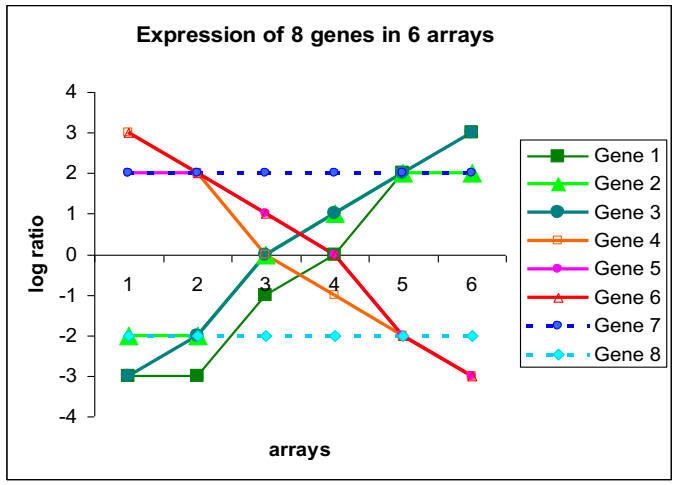


	Gene 1	Gene 2,3	Gene 4,(5,6)	Gene 7	Gene 8
Gene 1	0	1.73	10.30	8.00	6.93
Gene 2,3		0	8.66	6.08	6.71
Gene 4,(5,6)			**0.00**	6.78	6.86
Gene 7				0	9.80

… continue, join 1 to (2,3) at 1.73…. until done

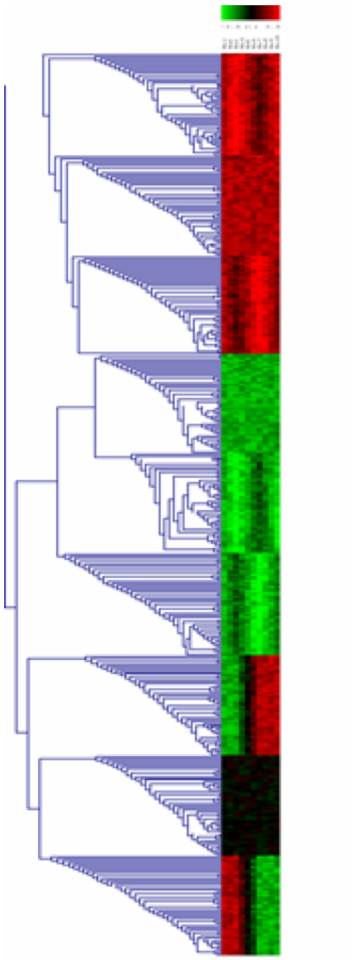


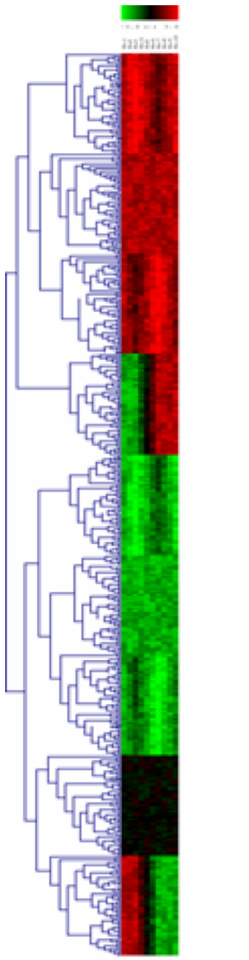


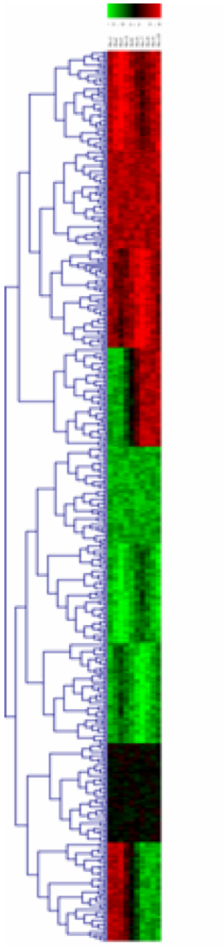


Hierarchical clustering assembles a number of items into a tree where items that are joined by short branches if they are very similar to each other and by increasingly longer branches as their similarity decreases.

Different Linkage Methods







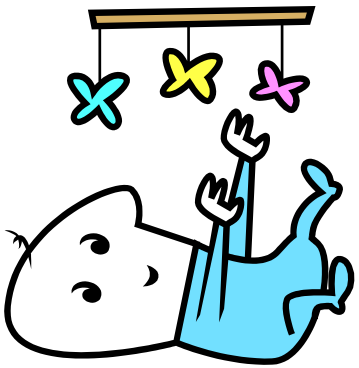
Single **Average** **Complete**

Join by min average max

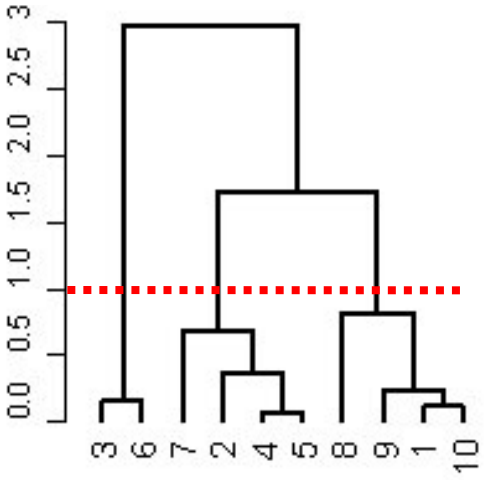
Quick Aside: Interpreting hierarchical clustering trees

Hierarchical analysis results viewed using a dendrogram (tree)

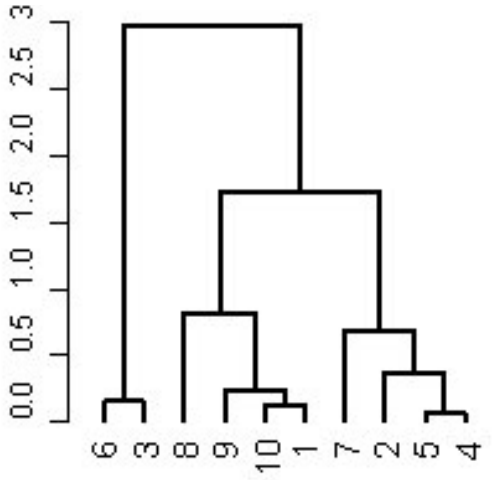
- Distance between nodes (Scale)
- Ordering of nodes not important (like baby mobile)



A



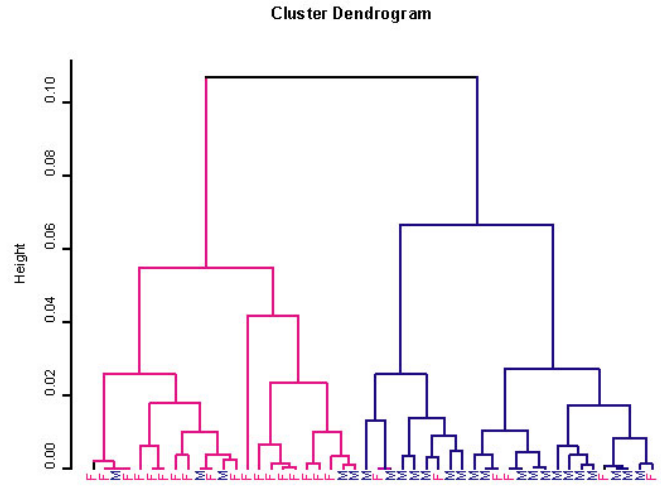
B

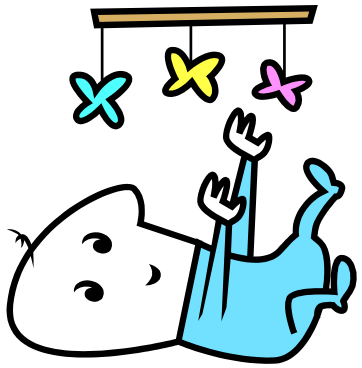


Tree A and B are equivalent

Limitations of hierarchical clustering

- Samples compared in a pair wise manner
- Hierarchy forced on data
- Sometimes difficult to visualise if large data
- Overlapping clustering or time/dose gradients ?





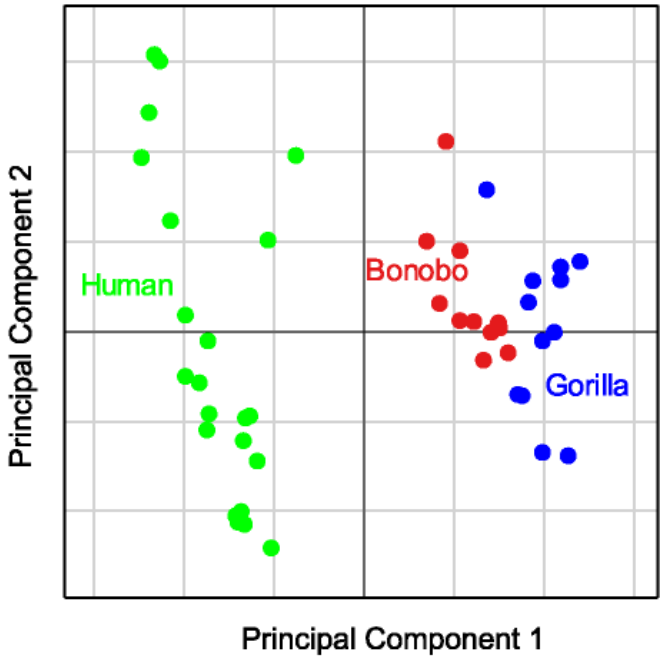
Ordination

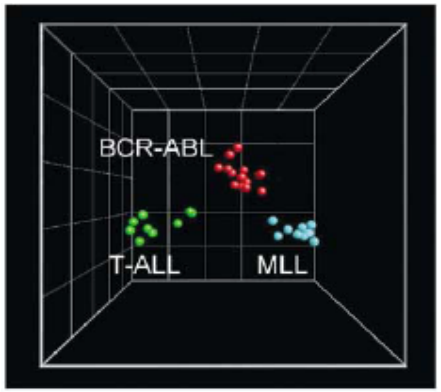
- Also refers to as
  - Latent variable analysis, Dimension reduction

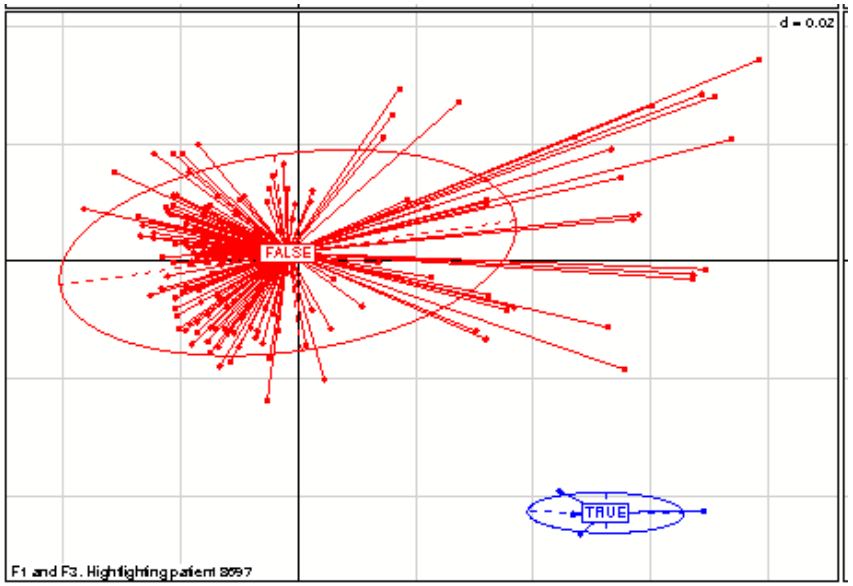
- Aim:

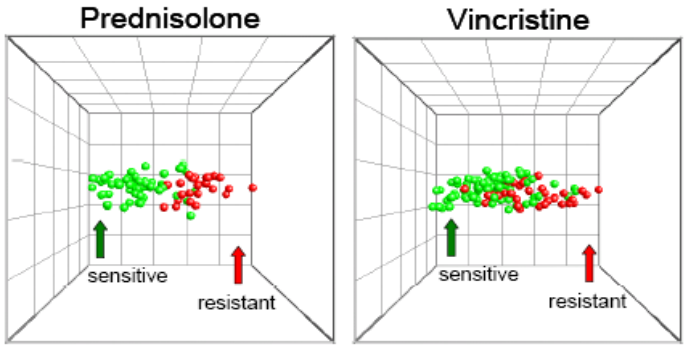
  Find axes onto which data can be project so as to explain as much of the variance in the data as possible

Ordination of Genomics Data

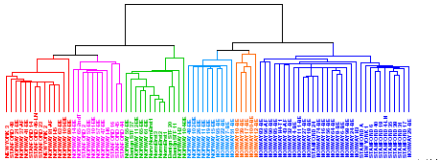




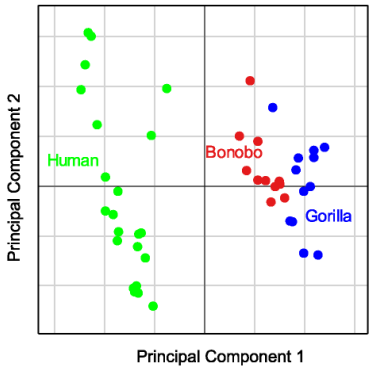




Complementary methods



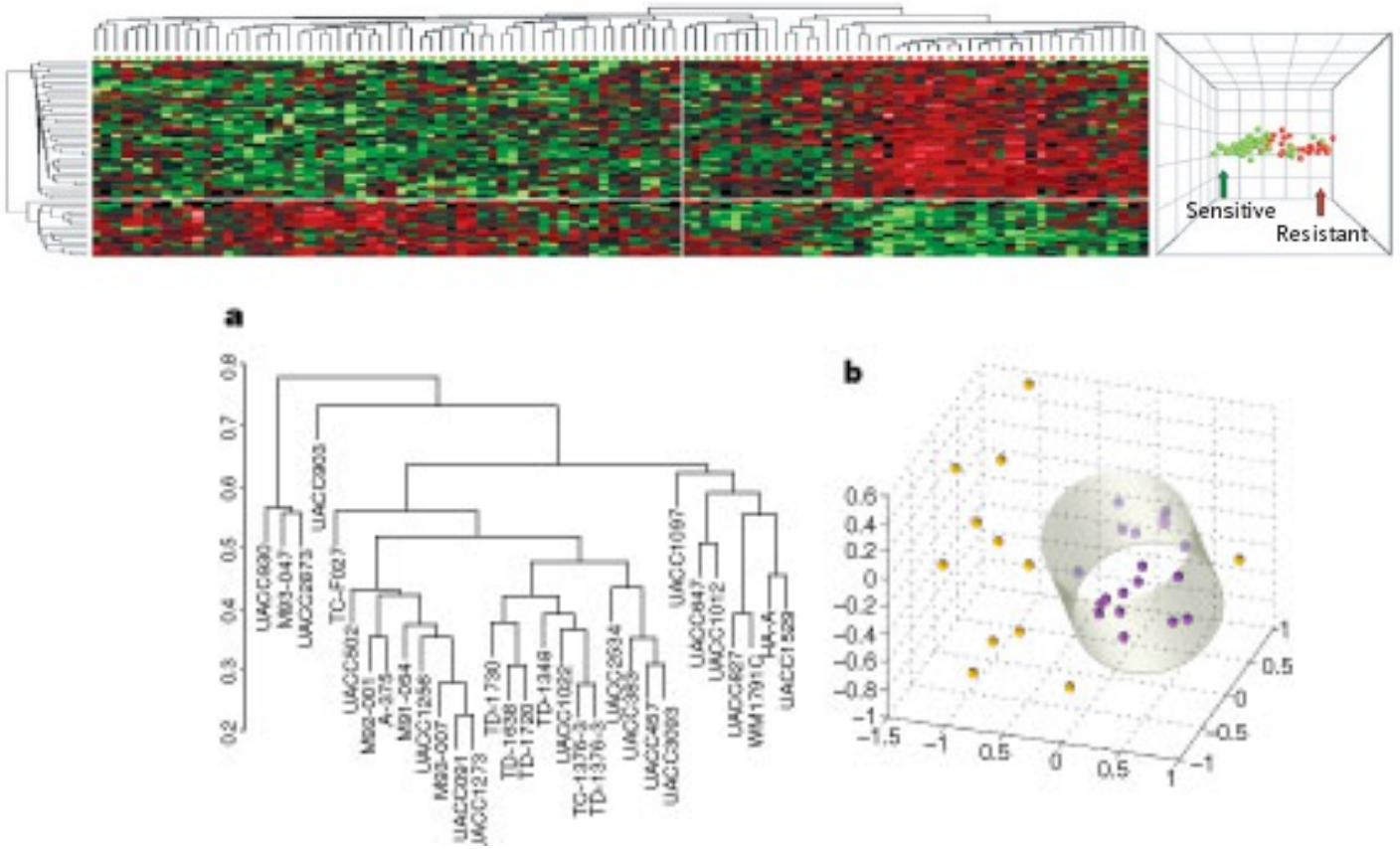
Cluster analysis generally investigates pairwise distances/similarities among objects looking for fine relationships



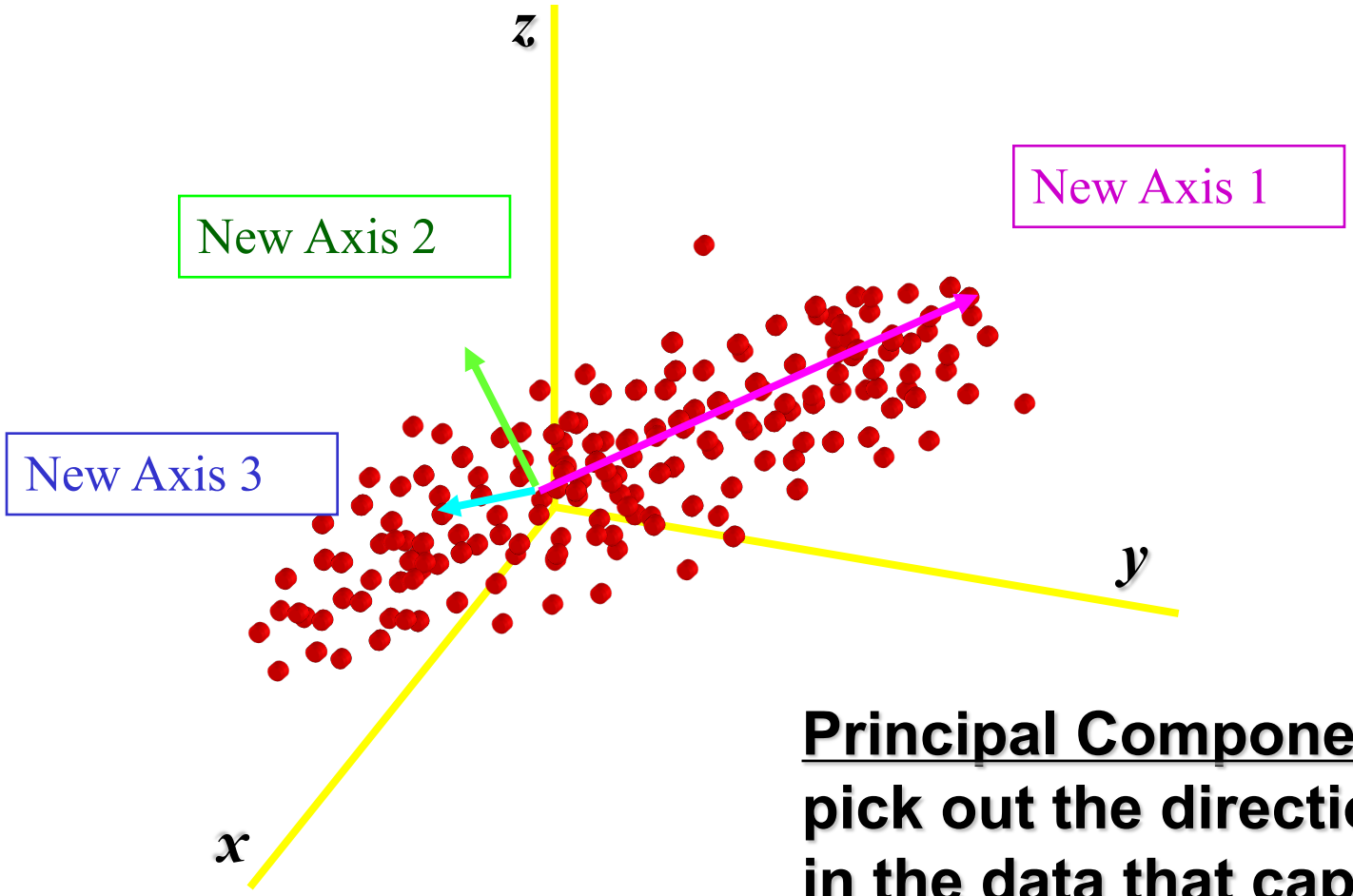
Ordination in reduced space considers the variance of the whole dataset thus highlighting general gradients/patterns

(Legendre and Legendre, 1998)

Many publications present both

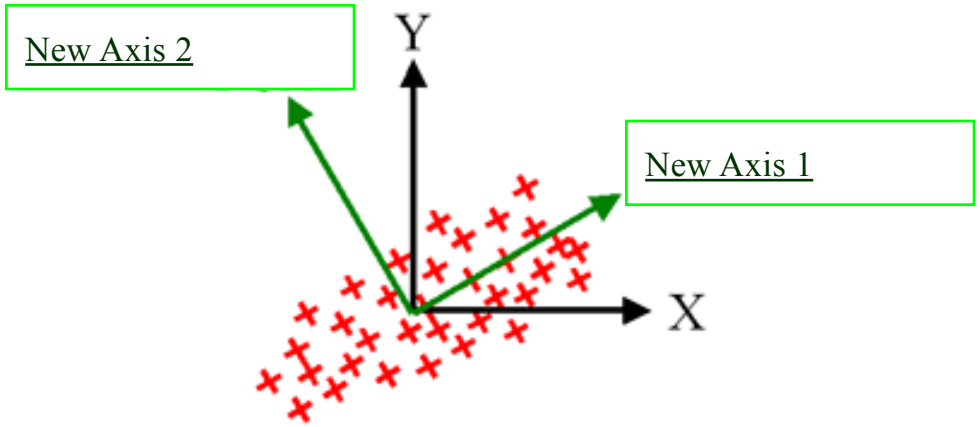


Dimension Reduction (Ordination)



Principal Components pick out the directions in the data that capture the greatest variability

Representing data in a reduced space



The first new axes will be projected through the data so as to explain the greatest proportion of the variance in the data.

The second new axis will be orthogonal, and will explain the next largest amount of variance

Interpreting an Ordination

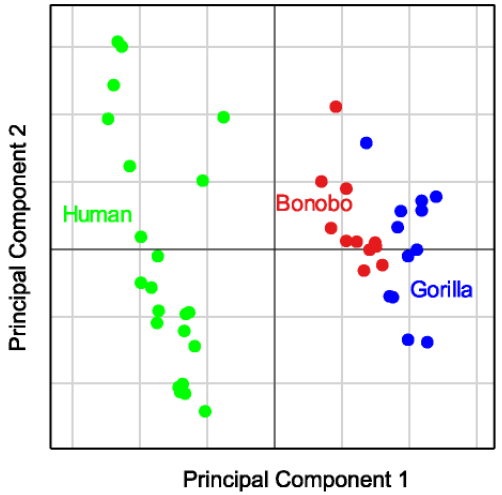
Each axes represent a different "trend" or set of profiles

The further from the origin

Greater loading/contribution

(ie higher expression)

Same direction from the origin



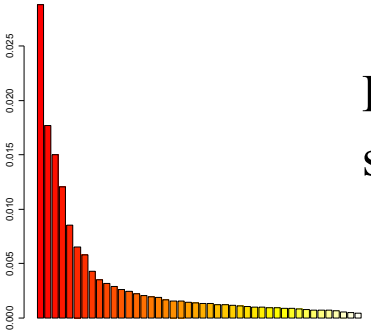
Principal Axes

- Project new axes through data which capture variance. Each represents a different trend in the data.
- Orthogonal (decorrelated)
- Typically ranked: First axes most important
- Principal axis, Principal component, latent variable or eigenvector

Typical Analysis

X

Ordination

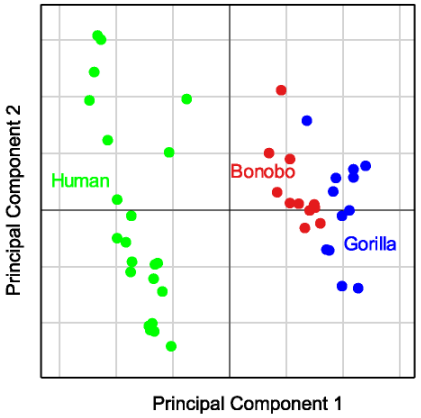


Plot of eigenvalues, select number.

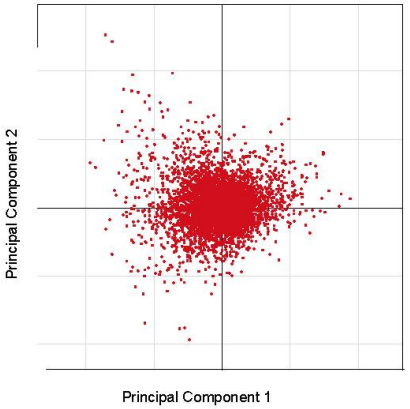
Plot PC1 v PC2

etc

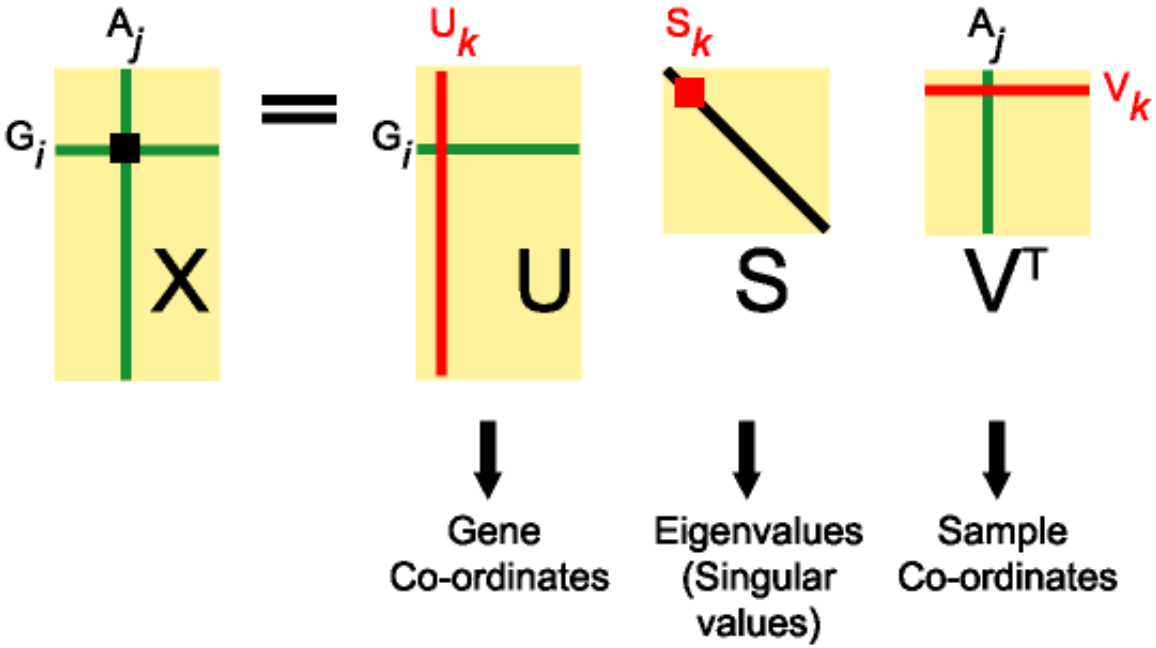
Array Projection



Gene Projection



Singular Value Decomposition $X=USV^T$



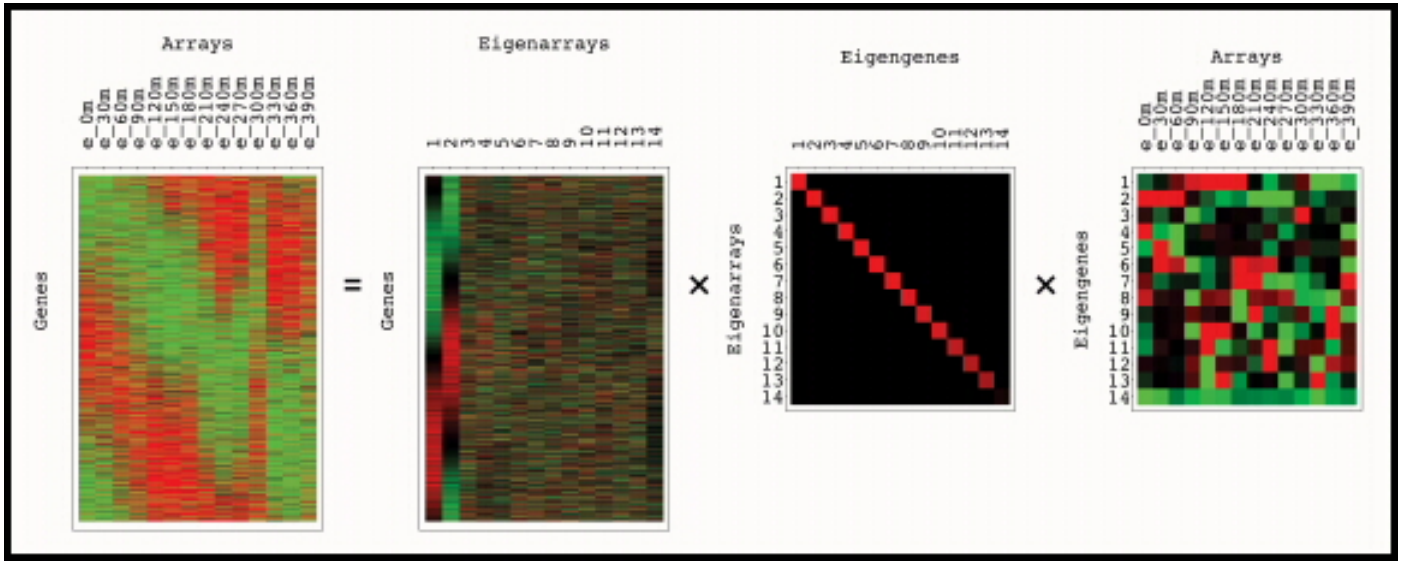


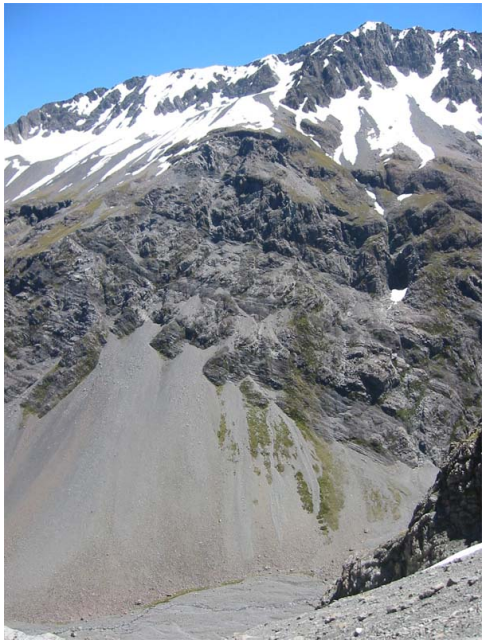
Image from

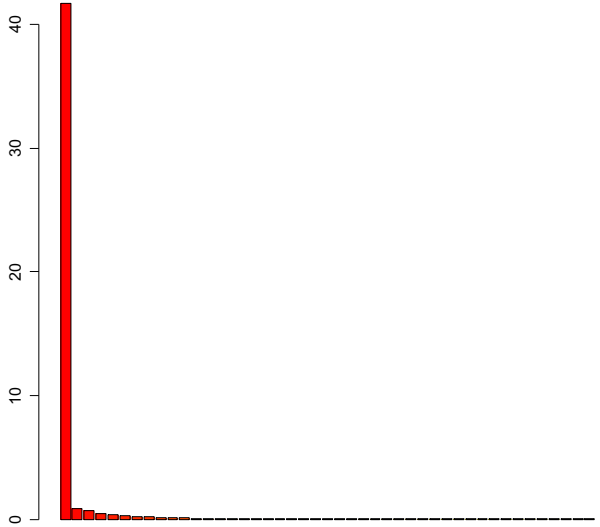
http://genome-www.stanford.edu/SVD/

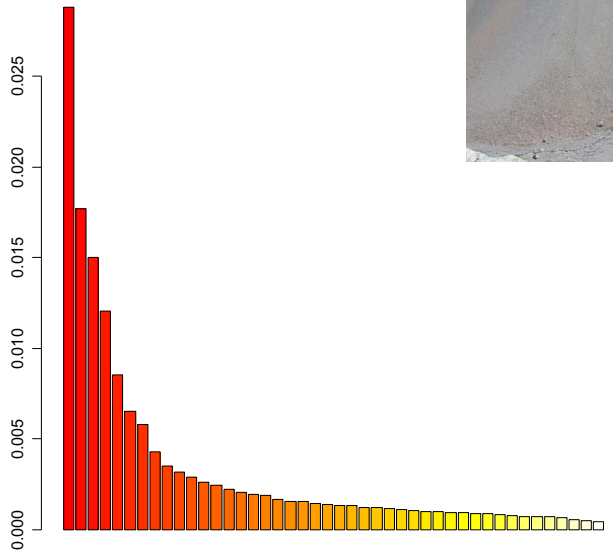
Eigenvalues

- Describe the amount of variance (information) in eigenvectors
- Ranked. First eigenvalue is the largest.
- Generally only examine 1st few components
  - scree plot

Choosing number of Eigenvalues: Scree Plot







Maximum number of Eigenvalues/Eigenvectors = min(nrow, ncol) -1

Ordination Methods

- Most common :
  - Principal component analysis (PCA)
  - Correspondence analysis (COA or CA)
  - Nonmetric multidimensional scaling (NMDS, MDS)
  - Principal co-ordinate analysis (PCoA)

Relationship

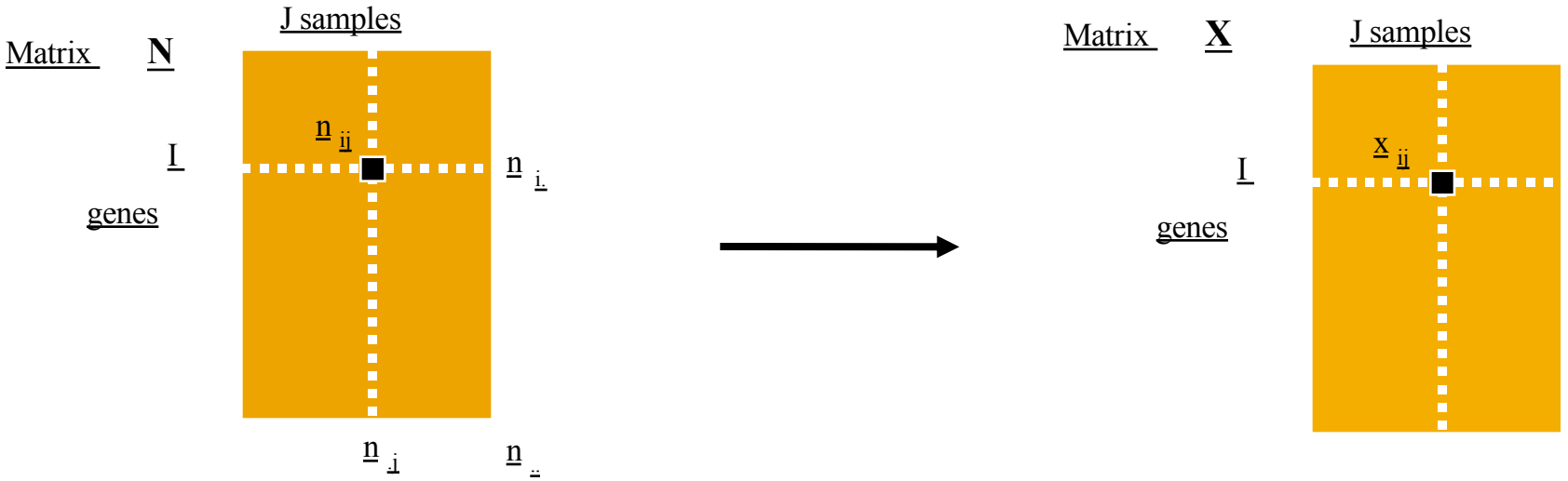
- PCA, COA, etc can be computed using Singular value decomposition (SVD)
- SVD applied to microarray data (Alter et al., 2000)
- Wall et al., 2003 described both SVD, PCA (good paper)

Principal Component Analysis (PCA)

- Probably most popular ordination method
- Eigenanalysis of a covariance matrix (most common) or correlation matrix
- Dates back to Pearson (1901)
- Applied to quantitative data.
- First applied to microarray data by Raychadhuri et al., 2000.
- PCA: prcomp(stats), princomp(stats) dudi.pca (ade4).

PCA: Initial data transformation

Column mean centred
(covariance PCA)

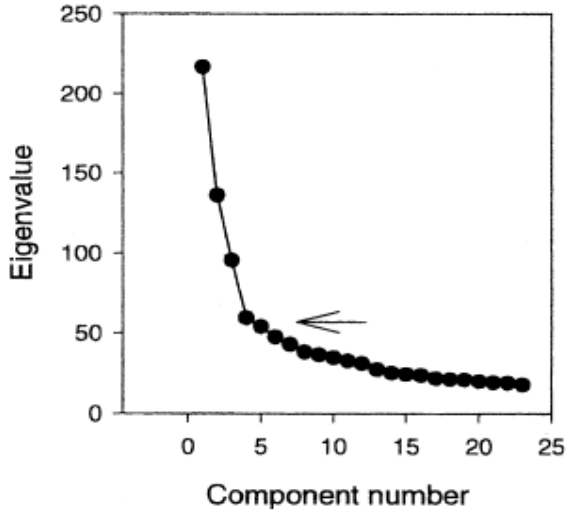


$$x_{ij} = n_{ij} - m_j$$

Where *mj* is the mean of column J

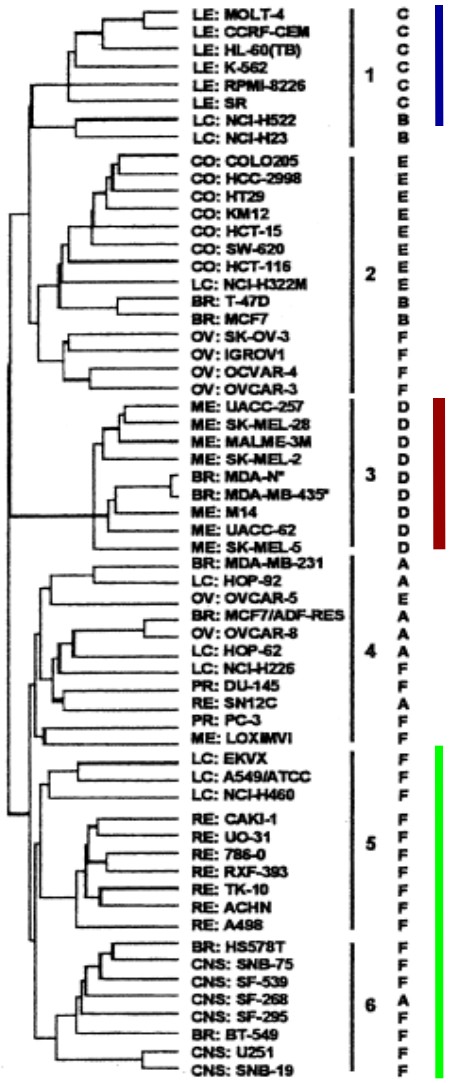
PCA of the NCI 60 cell lines

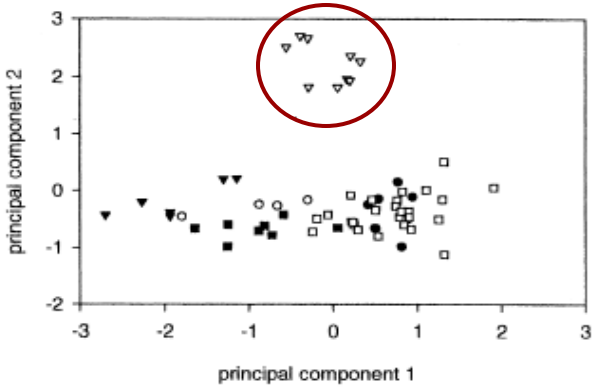
- Crescenzi and Giuliani, 2001
- Gene expression profiles of 60 cancer cell lines representing diverse cancers
- Performed PCA
- Found 5 PC's to contain most of the variance. Of which first 3 most interesting

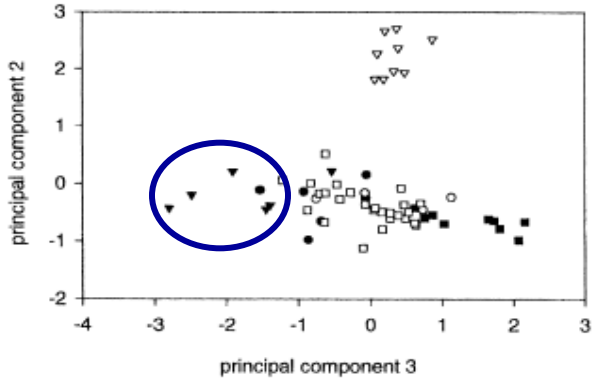


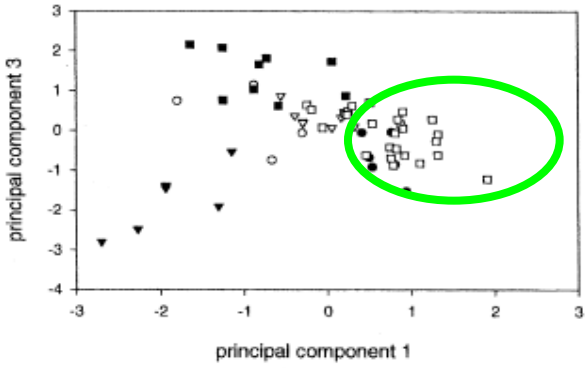
Eigenvalue distribution

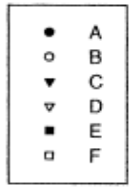
Component number	Eigenvalue	Variance (%)	Cumulative
1	216.70	15.30	15.3
2	135.85	9.59	24.9
3	95.44	6.74	31.6
4	59.42	4.20	35.8
5	53.80	3.80	39.6


LE: MOLT-4
LE: CCRF-CEM
LE: HL-60(TB)
LE: K-562
LE: RPMI-8226
LE: SR
LC: NCI-H522
LC: NCI-H23
CO: COLO205
CO: HCC-2998
CO: HT29
CO: KM12
CO: HCT-15
CO: SW-620
CO: HCT-116
LC: NCI-H322M
BR: T-47D
BR: MCF7
OV: SK-OV-3
OV: IGROV1
OV: OVCAR-4
OV: OVCAR-3
ME: UACC-257
ME: SK-MEL-28
ME: MALME-3M
ME: SK-MEL-2
BR: MDA-N
BR: MDA-MB-435
ME: M14
ME: UACC-62
ME: SK-MEL-5
BR: MDA-MB-231
LC: HOP-92
OV: OVCAR-5
BR: MCF7/ADF-RES
OV: OVCAR-8
LC: HOP-62
LC: NCI-H226
PR: DU-145
RE: SN12C
PR: PC-3
ME: LOX IMVI
LC: EKVX
LC: A549/ATCC
LC: NCI-H460
RE: CAKI-1
RE: UO-31
RE: 786-0
RE: RXF-393
RE: TK-10
RE: ACHN
RE: A498
BR: HS578T
CNS: SNB-75
CNS: SF-539
CNS: SF-268
CNS: SF-295
BR: BT-549
CNS: U251
CNS: SNB-19
1
2
3
4
5
6


principal component 2
principal component 1


principal component 2
principal component 3


principal component 3
principal component 1


A
B
C
D
E
F

Considerations when applying PCA

- Distance – Euclidean
- Robust, but designed for analysis of multi-normal distributed data
  - if very skewed data, the first few axes will only separate a few objects with extreme values instead of displaying main axes of variation
- Generally for microarray analysis: Row centre.
  - Eliminate size effect. Low abundance genes can be detected
- Problems which Correspondence analysis handles better
  - If lots zero
  - Unimodal or non-linear trends. Get distortion or artifact in plot, in which the second axis is an arched function of the first axis. Called horseshoe effect in PCA.

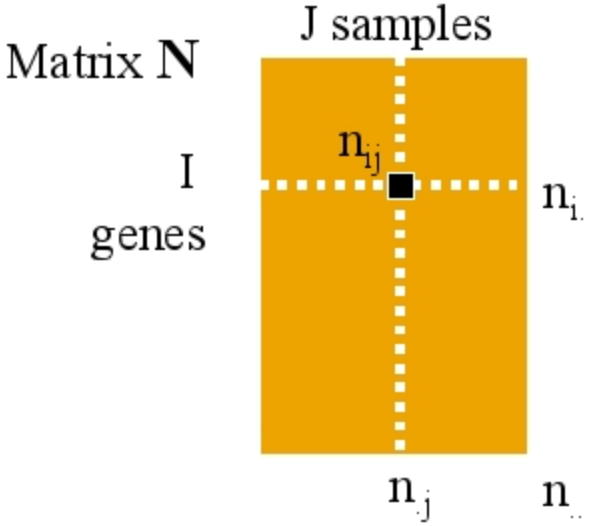

If I'd known they wanted me to use all this info – I would never have asked for it!
Julie

Correspondence Analysis

- **COA** (or CA) is an eigenanalysis of a Chi-square distance matrix.
- Measures the "strength" of association between an up-regulated gene and an array sample.
- Developed by numerous authors, also known as reciprocal averaging/ordering, dual scaling etc.
- Initially designed for analysis of 2-way contingency tables (frequency counts). Thus assumes matrix counts positive integers or zeros.

COA: Initial Transformation

Matrix **N**

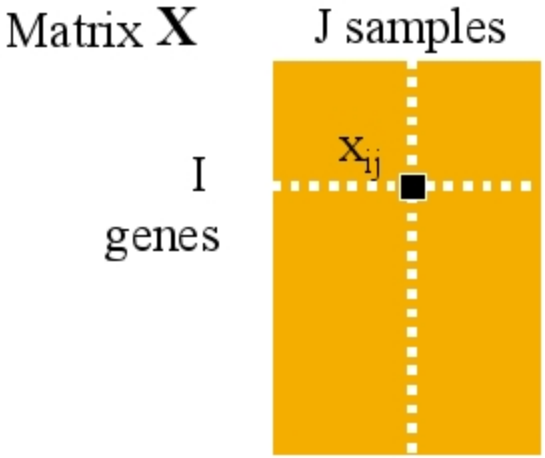


$$c_j = n_{.j}/n_{..}$$

$$r_i = n_{i.}/n_{..}$$

$$p_{ij} = n_{ij}/n_{..}$$

Matrix **X**



$$x_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j}$$

Pearson chi-square statistic O_{ij}-E_{ij} / $\sqrt{E_{ij}}$

COA of the Yeast Cell-Cycle Data

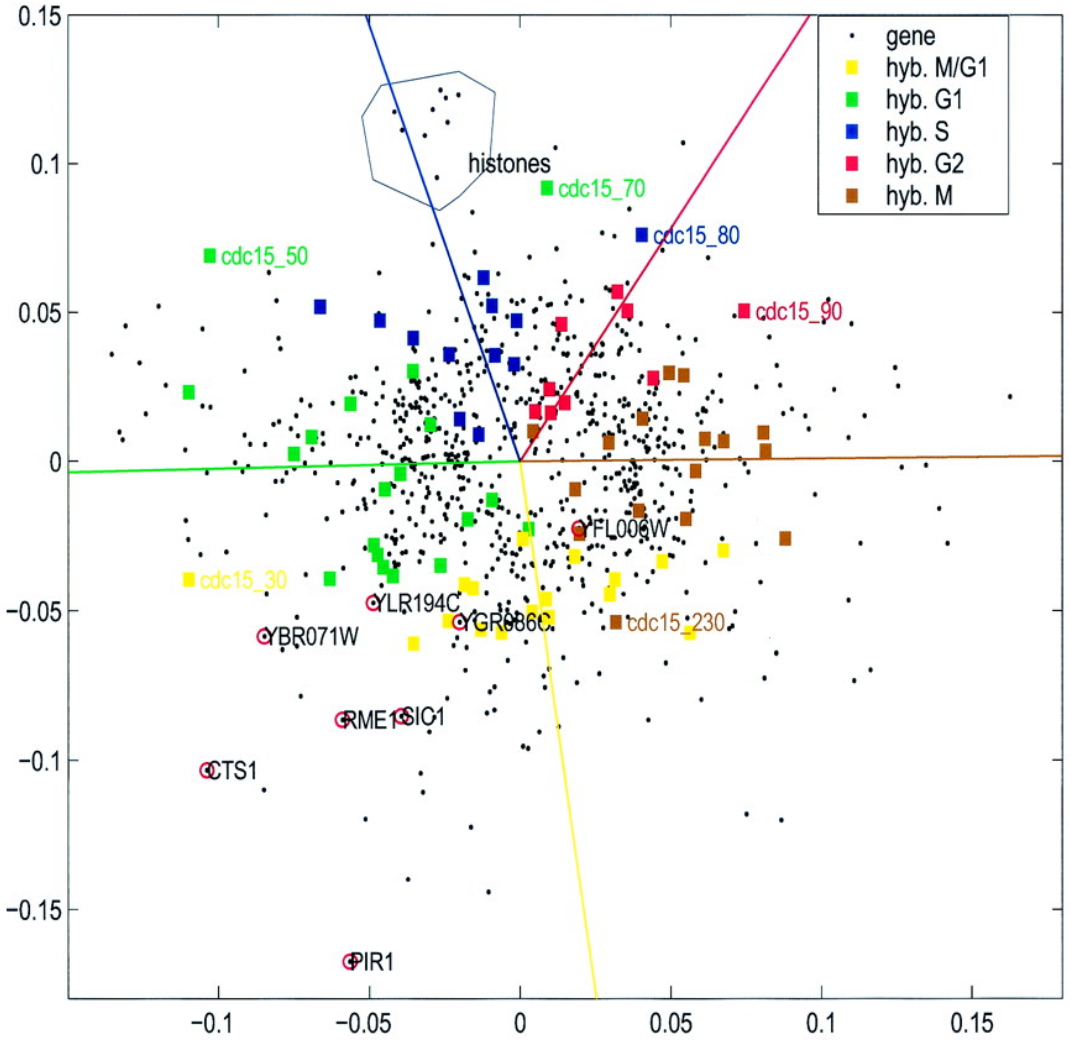
Fellenberg et al., 2001

- Dataset:
  - Gene Expression of *S. cerevisiae* arrested during cell cycle by 4 methods
    - alpha factor-, *CDC15-, CDC28*-based blocking and elutriation. (Spellman et al., 1998)
- COA
- Visualised using biplot of genes and arrays

Correspondence Analysis

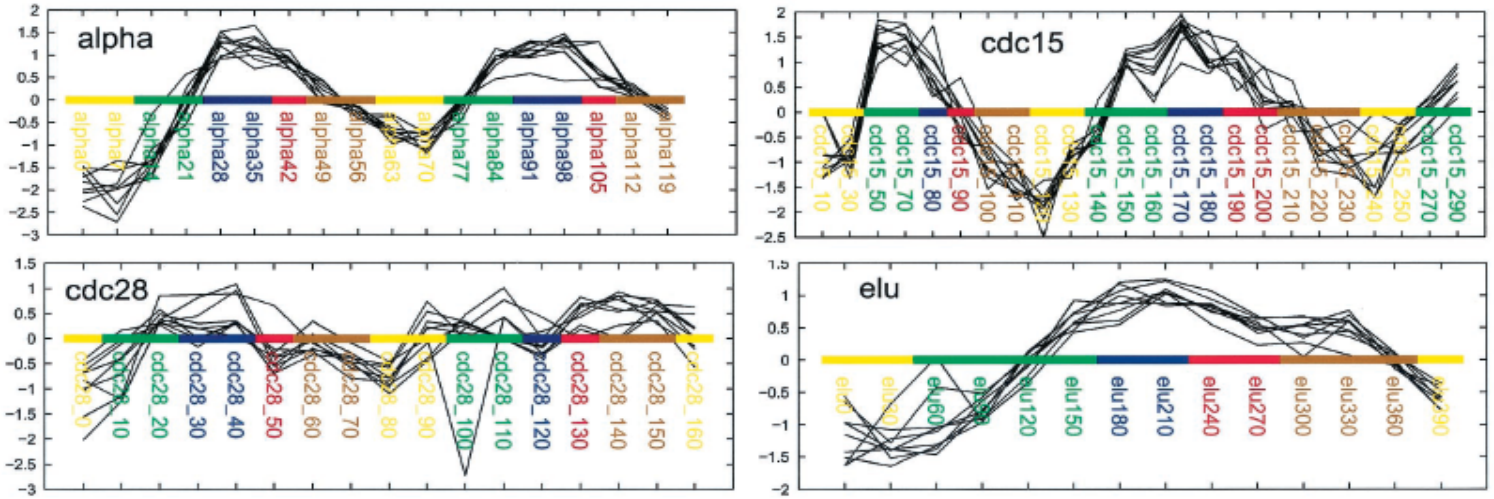
Biplot

Arrays and genes with a strong association (correspondence) are projected in the same direction from the origin.



Fellenberg et al., (2001) PNAS 98(19):10781-10786

Gene Expression of 9 histone genes



See gene expression of histones ↑ during ■ S phase

Consideration when applying COA

- Data must be in same units (so they can be added)
- Data must be non-negative or made position by translation (scalar addition)
- In case of steep gradients (many zero) COA should produce better results than PCA
- Data is dual (column and row) scaled.
- Unimodal or non-linear trends may be represented as arch (2^{nd} axis). Less serious than PCA's horseshoe effect.

Among the other related methods

Independent Component Analysis

- generalization of PCA
- does not constrain the axes to be orthogonal
- attempts to place them in the directions of statistical dependencies in the data.
- Lee & Batzoglou 2003 and Saidi et al., 2004 show ICA outperformed PCA

Spectral map analysis

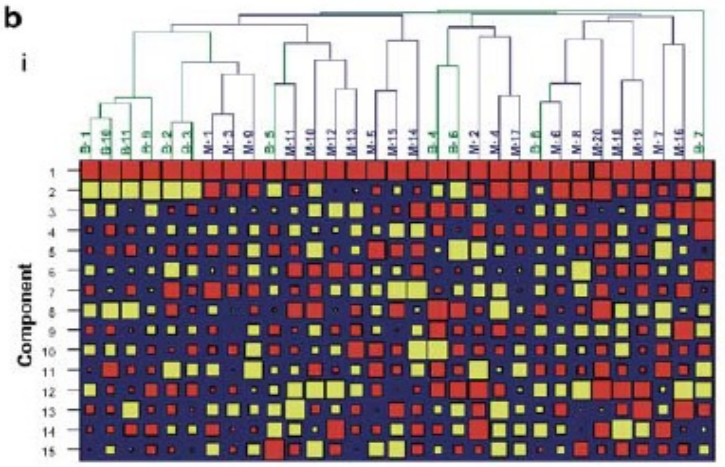
- related to COA (dual scaling of both rows + columns)
- not limited to contingency tables and cross-tabulations. possibility to use other weighting factors
- Wouters et al., 2003 showed SMA outperformed PCA, comparable to COA.

MDS

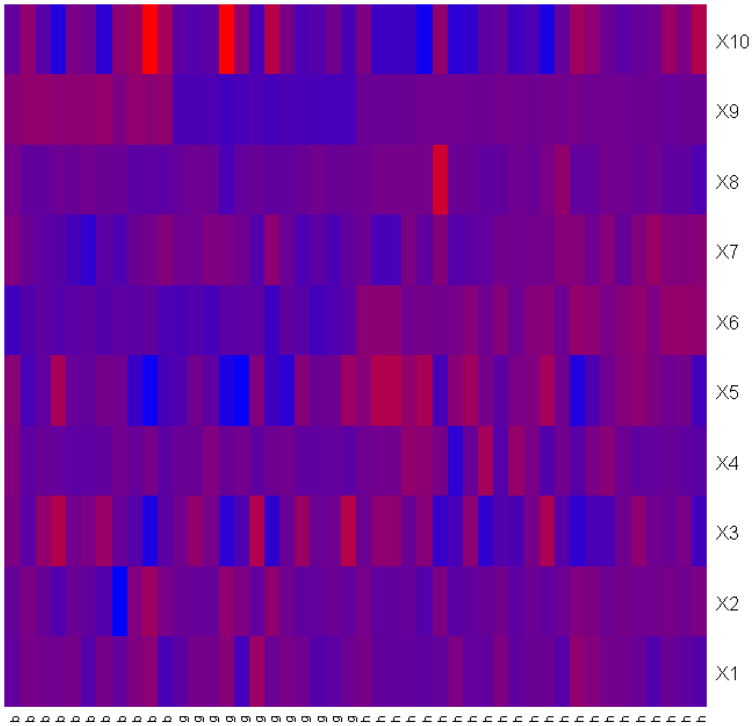
- Multidimensional scaling
  - metric and non-metric.
- Input distance matrix
- Classical MDS is identical to principal coordinates analysis (PCoA). R code cmdscale (stats), dudi.pco (ade4)
- NMDS. Iterative. isoMDS (MASS), sammon (MASS).

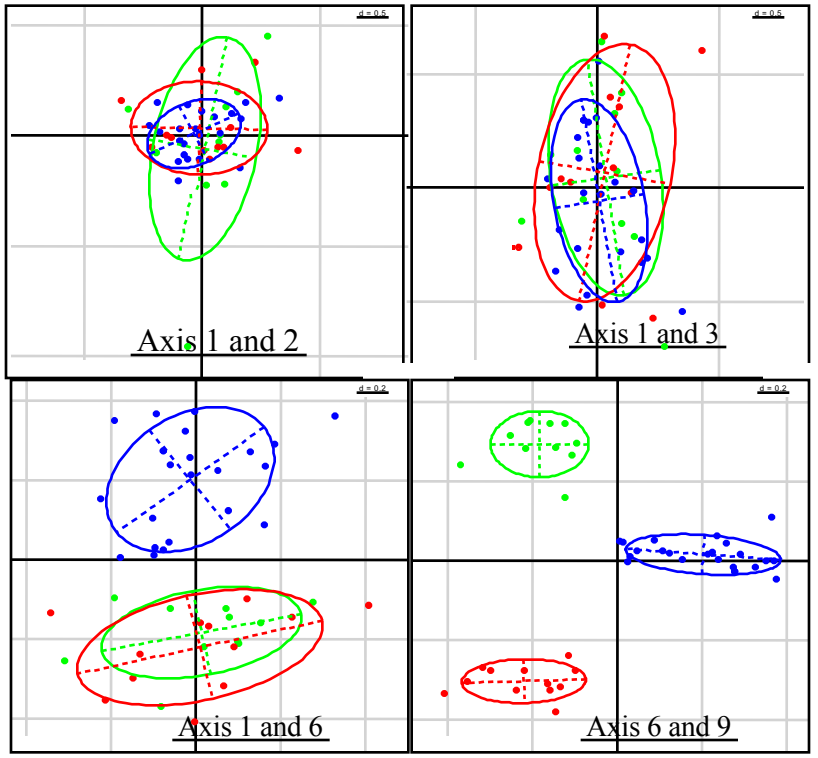
Independent Component Analysis Saudi et al., 2003

- Axes/eigenvectors are not constrained to be orthogonal
- May detect more subtle patterns in data
- More complex interpretation of axes

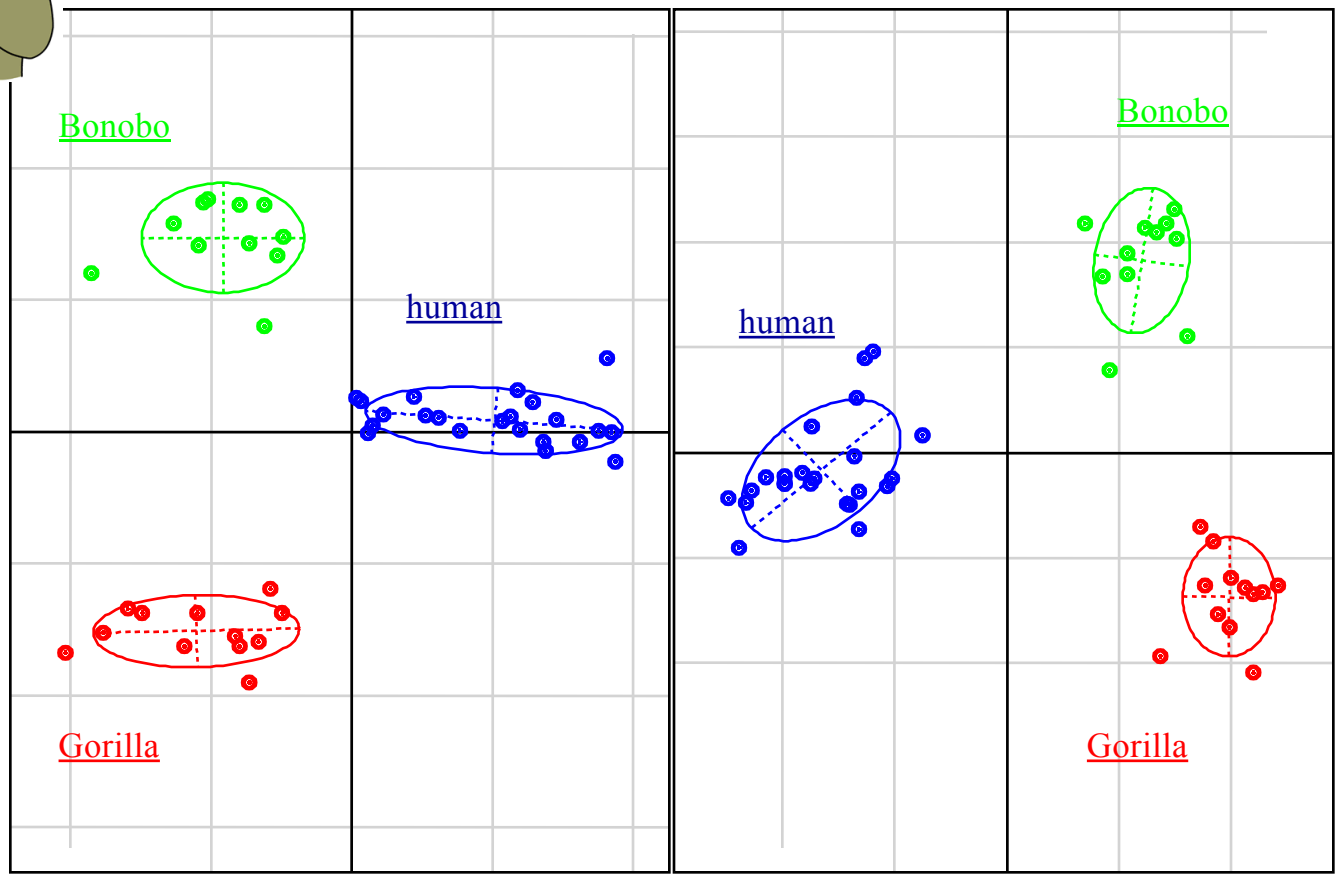


Fast ICA of Karaman Ape Data





Results from fastICA and COA



fastICA, Axis 6 and 9

COA, Axis 1 and 3

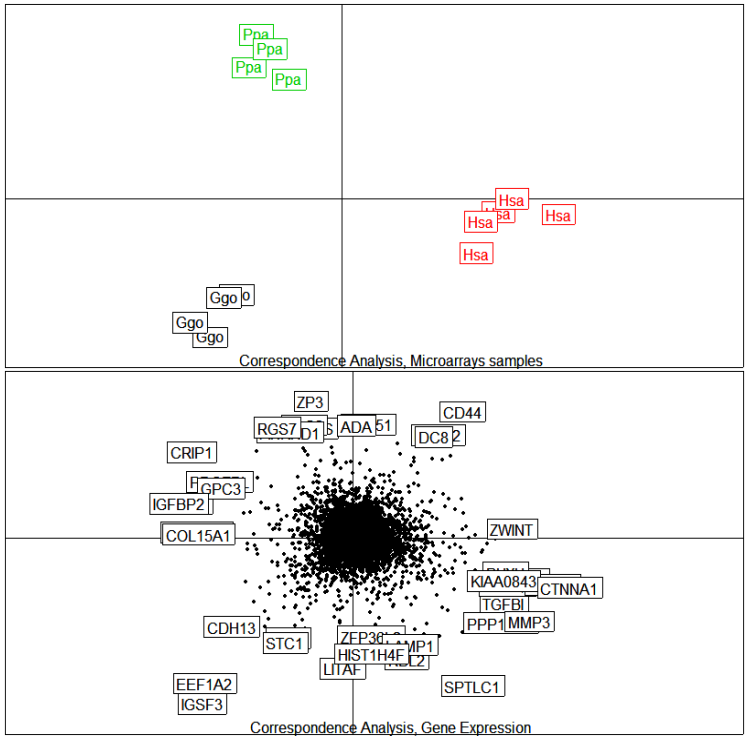
Summary: Exploration analysis using Ordination

- SVD = straightforward dimension reduction
- PCA = column mean centred +SVD
  - Euclidean distance
- COA = Chi-square +SVD
  - produces nice biplot

- Ordination be useful for visualising trends in data
- Useful complementary methods to clustering

Ordination in R

Ordination (PCA, COA)



- `library(ade4)`
- `dudi.pca()`
- `dudi.coa()`

- `library(made4)`
- `ord(data, type="pca")`
- `plot()`
- `plotarrays()`
- `plotgenes()`

Link to example 3d html file

Books/Book Chapters:

1. Legendre, P., and Legendre, L. 1998. *Numerical Ecology*, 2nd English Edition. ed. Elsevier, Amsterdam.
2. Wall, M., Rechtsteiner, A., and Rocha, L. 2003. Singular value decomposition and principal component analysis. In *A Practical Approach to Microarray Data Analysis*. (eds. D.P. Berrar, W. Dubitzky, and M. Granzow), pp. 91-109. Kluwer, Norwell, MA.

Papers:

1. Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2:** 559-572.
2. Hotelling, H., 1933. Analysis of a complex statistical variables into principal components. J. Educ. Psychol. 24, 417-441. Alter, O., Brown, P.O., and Botstein, D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* **97:** 10101-10106.
3. Culhane, A.C., Perriere, G., Considine, E.C., Cotter, T.G., and Higgins, D.G. 2002. Between-group analysis of microarray data. *Bioinformatics* **18:** 1600-1608.
4. Culhane, A.C., Perriere, G., and Higgins, D.G. 2003. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics* **4:** 59.
5. Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J.D., and Vingron, M. 2001. Correspondence analysis applied to microarray data. *Proc Natl Acad Sci U S A* **98**: 10781-10786.
6. Raychaudhuri, S., Stuart, J.M., and Altman, R.B. 2000. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*: 455-466.
7. Wouters, L., Gohlmann, H.W., Bijnens, L., Kass, S.U., Molenberghs, G., and Lewi, P.J. 2003. Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics* **59**: 1131-1139

Reviews

1. Quackenbush, J. 2001. Computational analysis of microarray data. *Nat Rev Genet* **2:** 418-427.
2. Brazma A., and Culhane AC. (2005) Algorithms for gene expression analysis. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics.* Dunn MJ., Jorde LB., Little PFR, Subramaniam S. (eds) John Wiley & Sons. London (download from http://www.hsph.harvard.edu/research/aedin-culhane/publications/)

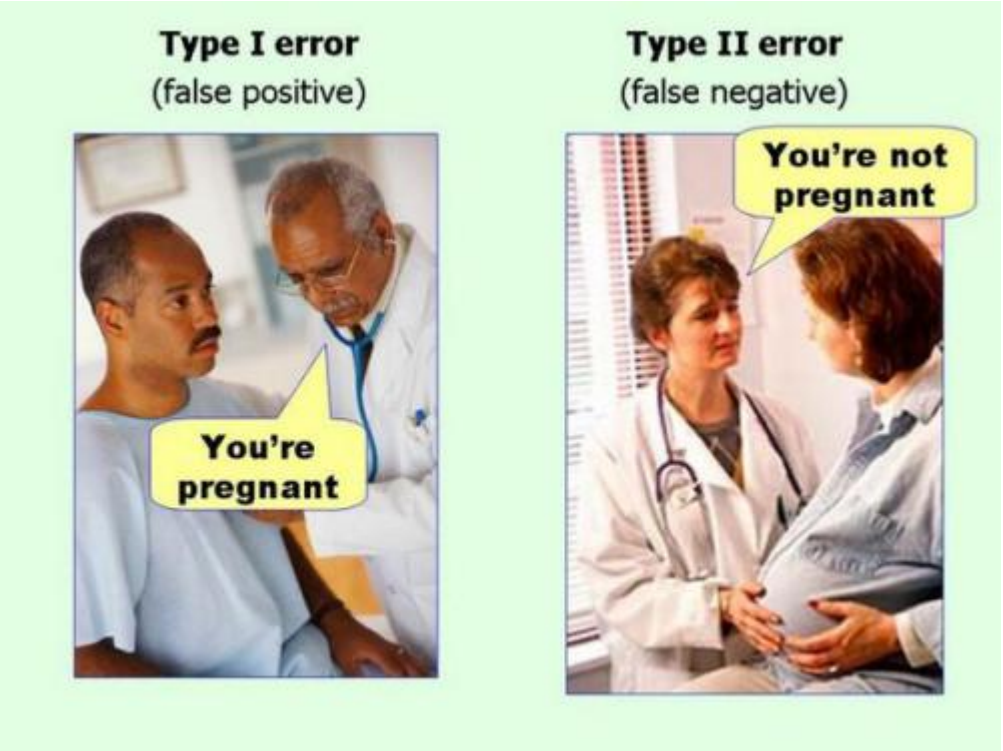
Interesting Commentary

Terry Speed's commentary on PCA download from http://bulletin.imstat.org/pdf/37/3

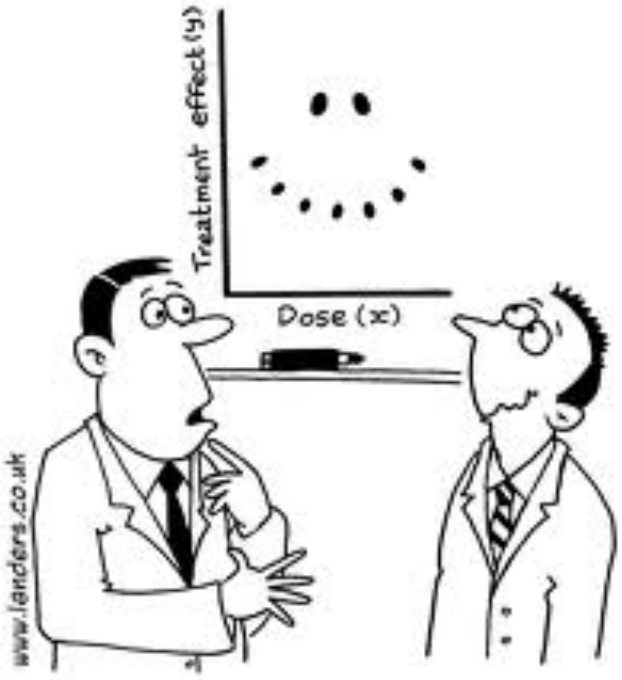
Summary: Exploratory Data Analysis

- Test your intuition about the data
- Keep open mind to discover unexpected patterns
- Be curious and skeptical

EDA is a FIRST step before predictive modelling



EDA: Finding the unexpected



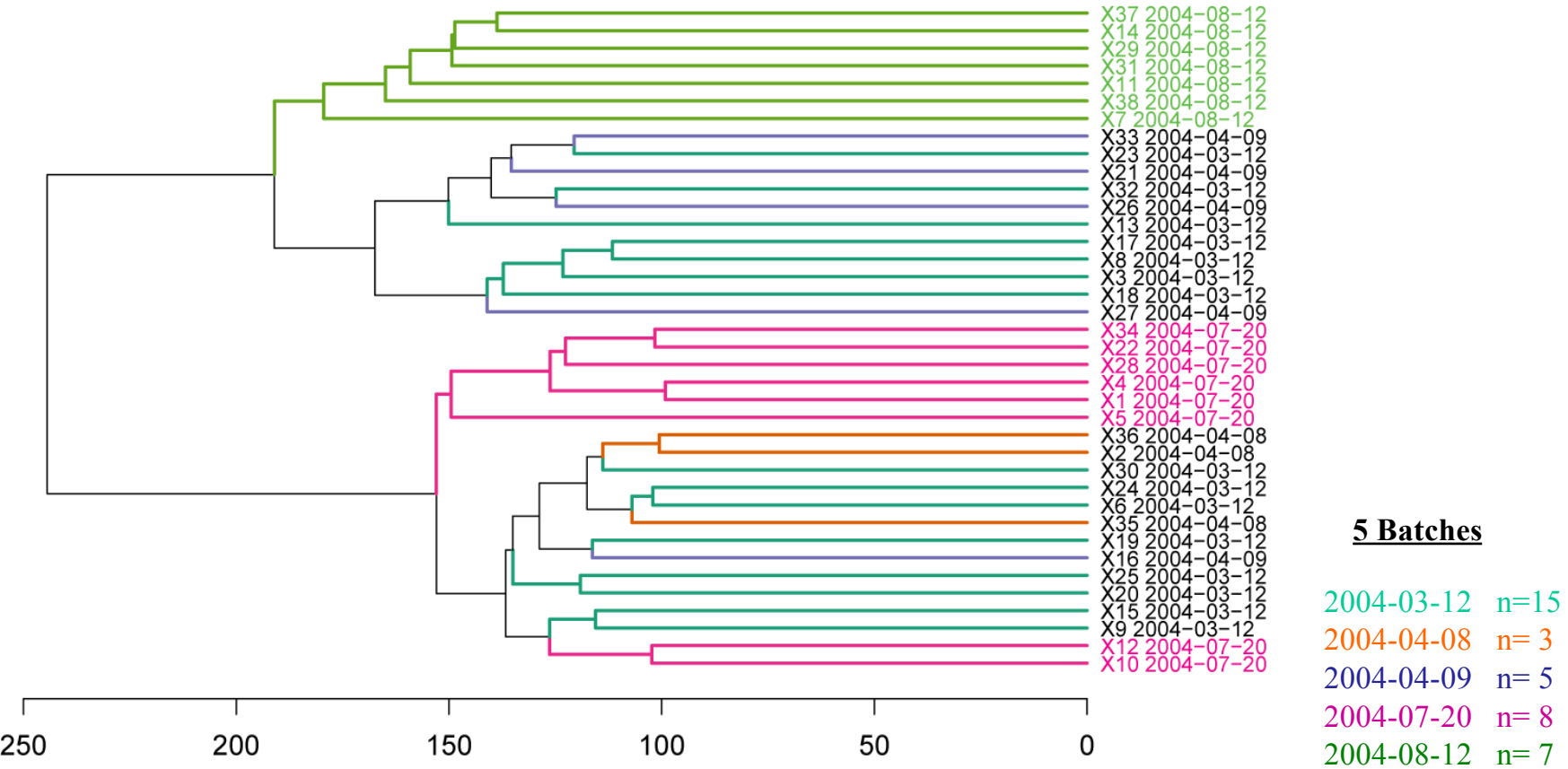
"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

Case Study 1

- Poor experimental Design

- Batch Effects
  - EDA analysis reveal samples cluster by RNA extraction data, technician, assay date, machine,
  - This are non-random biases in the data
  - Increasing the sample size, or standard normalization will not resolve this

Samples cluster by date



5 Batches

2004-03-12 n=15
2004-04-08 n= 3
2004-04-09 n= 5
2004-07-20 n= 8
2004-08-12 n= 7

How to fix batch effects?

OPINION

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry

- Rafa Irizarry will present Combat, Surrogate Variable Analysis on June 28th

Case Study 2

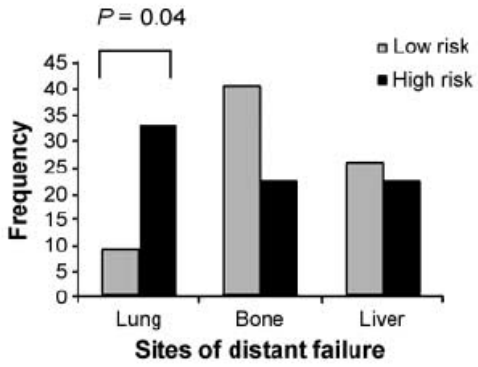
- Poor experimental Design
- Confounding Covariates

Case Study 2:
A 6 gene signature of lung metastasis

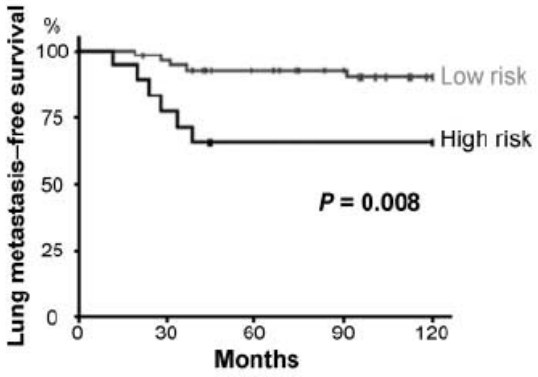
A

Characteristics	All patients (*n*=72)	High risk (*n*=18)	Low risk (*n*=54)	*P*
Metastasis	38 (53%)	10 (56%)	28 (52%)	-
Lung metastasis	11 (15%)	6 (33%)	5 (9%)	0.04
Postmenopause	40 (69%)	9 (60%)	31 (72%)	-
Macroscopic tumor size (>20 mm)	47 (67%)	13 (72%)	34 (65%)	-
Grade 3 (SBR)	18 (29%)	9 (56%)	9 (20%)	0.01
Estrogen receptor negative	28 (39%)	15 (83%)	13 (24%)	<0.001
Progesterone receptor negative	35 (49%)	14 (78%)	21 (39%)	0.004

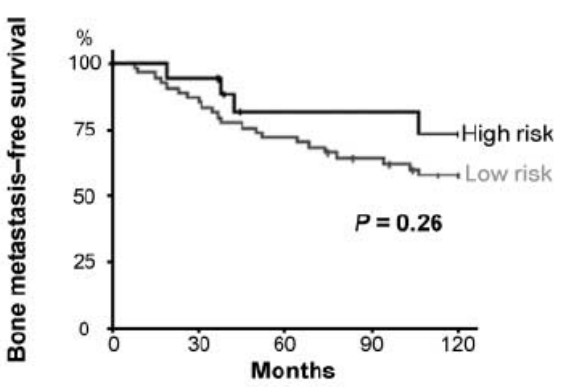
B



C



D



Landemaine T et al., Cancer Res. 2008 Aug 1;68(15):6092-9.

But metastatic profile of breast cancer differs by tumor subtype

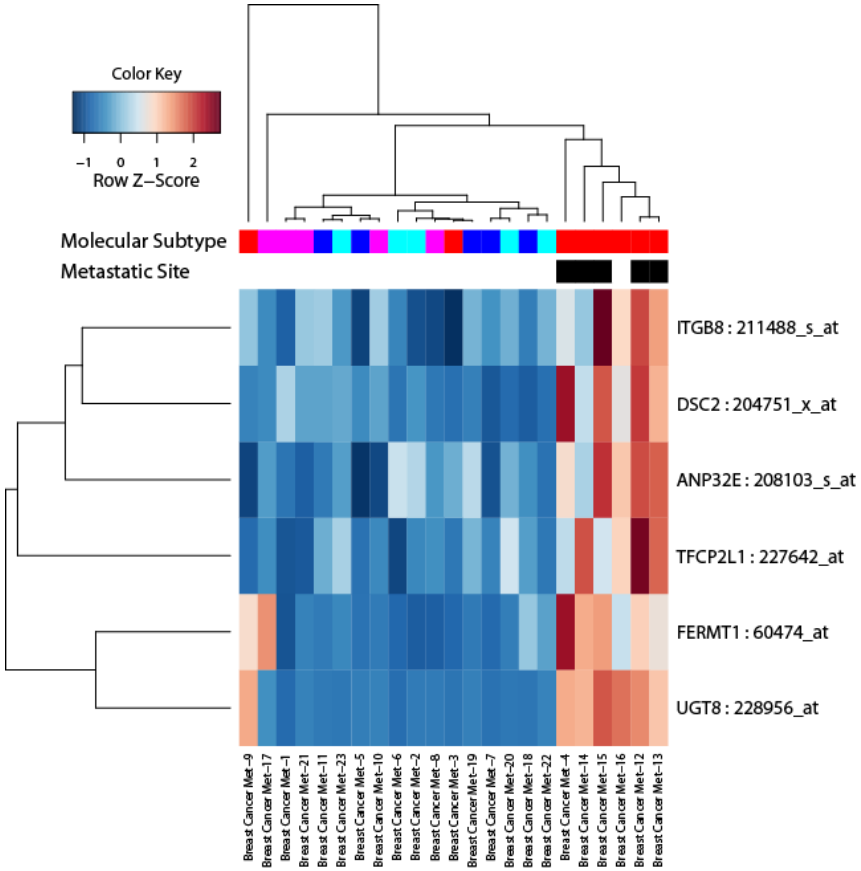
Table 1. Frequencies of site of relapse in the molecular subtypes

Subtype	Site of relapse					
	Bone	Lung	Liver	Brain	Pleura	Total
Luminal B	26 (36.6)	11 (36.7)	2 (11.1)	1 (7.1)	5 (41.7)	45
Luminal A	22 (31.0)	2 (6.7)	4 (22.2)	1 (7.1)	5 (41.7)	34
Erbb2	14 (19.7)	4 (13.3)	6 (33.3)	3 (21.4)	0 (0.0)	27
Normal	4 (5.6)	1 (3.3)	2 (11.1)	1 (7.1)	1 (8.3)	9
Basal	5 (7.0)	12 (40.0)	4 (22.2)	8 (57.1)	1 (8.3)	30
Total	71	30	18	14	12	145

NOTE: Numbers between parentheses are column percentages, e.g., 36.6% of bone relapses are in the luminal B subtype.

Smid et al., 2008 Cancer Res 68(9):3108–14

Confounding Covariates

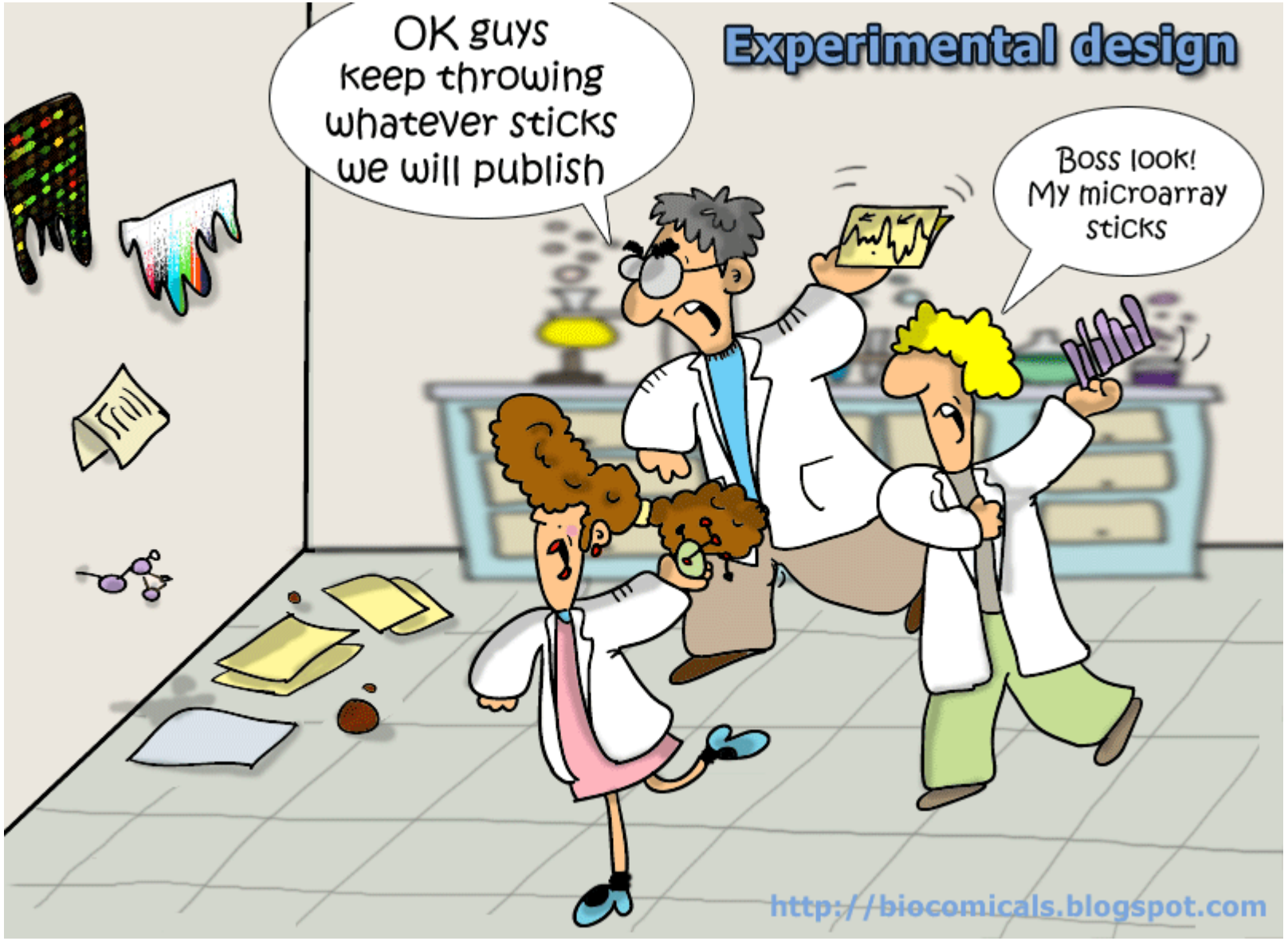


Confounding Covariates

Supplementary Table 4. Results of Analysis of Global Test and GlobalAncova analysis of MSK dataset (p-value)

Method	globaltest	globaltest	GlobalAncova	GlobalAnova
Number of Probesets tested *	4	10	4	10
Q1: Are the genes associated with metastases status?	0.048	0.100	0.015	0.023
Q2: Are the genes associated with molecular subtype ?	<0.00000001	<0.0000001	0	0
Q3: Is metastases status significant independent of molecular subtype?	0.720	0.694	0.630	0.696
Q4: Is molecular subtype significant independent of metastases status ?	<0.000001	<0.000001	0	0
Q5: Are the genes associated with metastases status in the basal-like tumors?	0.514	0.168	0.380	0.190

Culhane AC & Quackenbush J. 2009 Cancer Research (In Press)


Experimental design
OK guys keep throwing whatever sticks we will publish
Boss look! My microarray sticks
http://biocomicals.blogspot.com

Experimental Design Pop Quiz

Example: benzopyrene toxicity

- Study: toxic effect of Benzo(a)pyrene on rats
- 8 rats are to be treated with BP and 8 rats with a control compound
- Each array will be hybridised against a reference sample
- 16 arrays in experiment

Experimental Design

- There are 2 batches of 8 slides, from 2 different print runs
- The hybridisation will be done by 2 different researchers, Alison and Brian
- What is the best way to arrange the experiment?

Design 1

- Alison prepares all 8 BP samples and hybridises them to the arrays of print run 1

- Brian prepares all 8 control samples and hybridises them to the arrays of print run 2

Design 2

- Alison chooses 8 rats and treats 4 with BP and 4 with control substance
- She prepares and hybridises 2 BP samples to arrays from print run 1 and 2 BP samples to arrays from print run 2
- She prepares and hybridises 2 control samples to arrays from print run 1 and 2
- Brian does the same with the other 8 rats

Design 3

- 8 rats are randomly assigned to Alison, along with 4 BP preps and 4 control preps - she is not told which preps are which
- She prepares and hybridises samples to randomly prearranged arrays so that 2 BP samples and 2 control samples are hybridised to 4 arrays from each of print runs 1 and 2
- Brian does the same with the other 8 rats

What is wrong with design 1?

- Treatment, researcher and print run are CONFOUNDED variables
- We cannot tell whether differences between the two groups of rats result from treatment, researcher or print run
- Use blocking, in designs 2 and 3 to deconfound the variability of interest (treatment) from the extraneous variables
- Designs 2 and 3 are also BALANCED, which increases the power of the analysis

What is wrong with Design 2?

- Alison's choice of rats may be BIASED
- e.g. she may choose the healthiest rats, confounding potential treatment effects with researcher variability
- Use randomisation and blinding in design 3 to avoid bias

Blocking, Randomization and Blinding

- Arrangement of experimental design that minimises problems from extraneous sources of variability
- Use blocking to avoid CONFOUNDING (extraneous variables)
- Use randomisation and blinding to avoid BIAS

Good Experimental Design & Sample Processing is Critical

Dr. Frederick Frankenstein: Igor, would you mind telling me whose brain I did put in?

Igor: And you won't be angry?

Dr. Frederick Frankenstein: I will NOT be angry.

Igor: Abby someone.

Dr. Frederick Frankenstein: Abby someone. Abby who?

Igor: Abby Normal.

Dr. Frederick Frankenstein: Abby Normal?

Igor: I'm almost sure that was the name.

Dr. Frederick Frankenstein: Are you saying that I put an abnormal brain into a seven and a half foot long, fifty-four inch wide GORILLA? IS THAT WHAT YOU'RE TELLING ME?

From the film
Young Frankenstein, 1974

http://www.youtube.com/watch?v=NOe_4mgmyyA