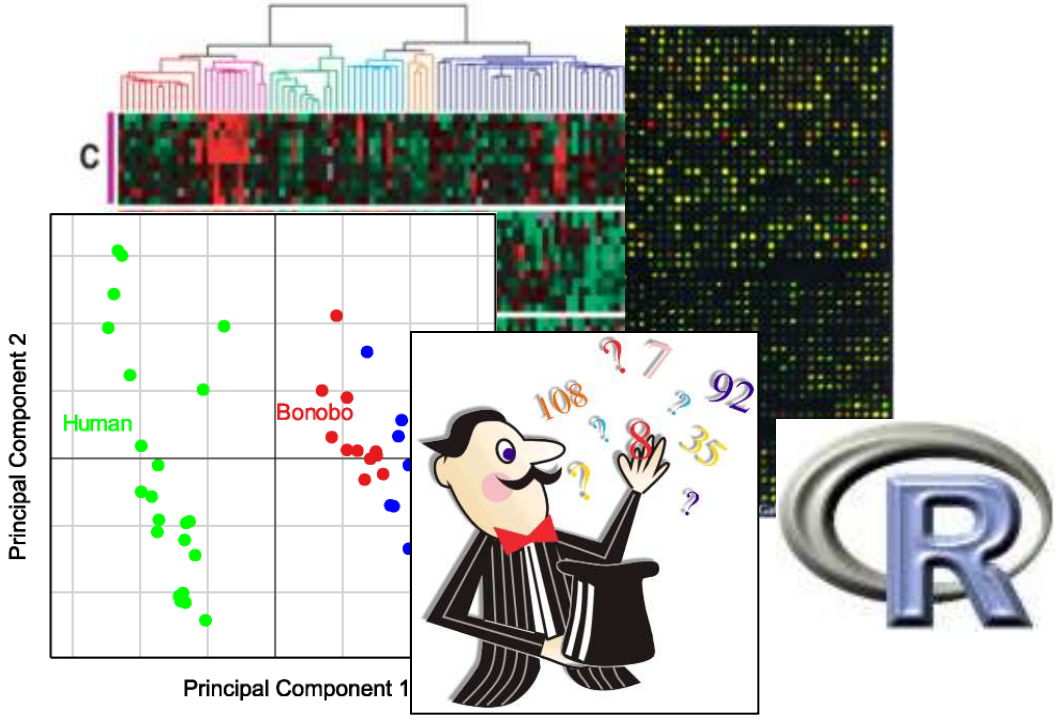
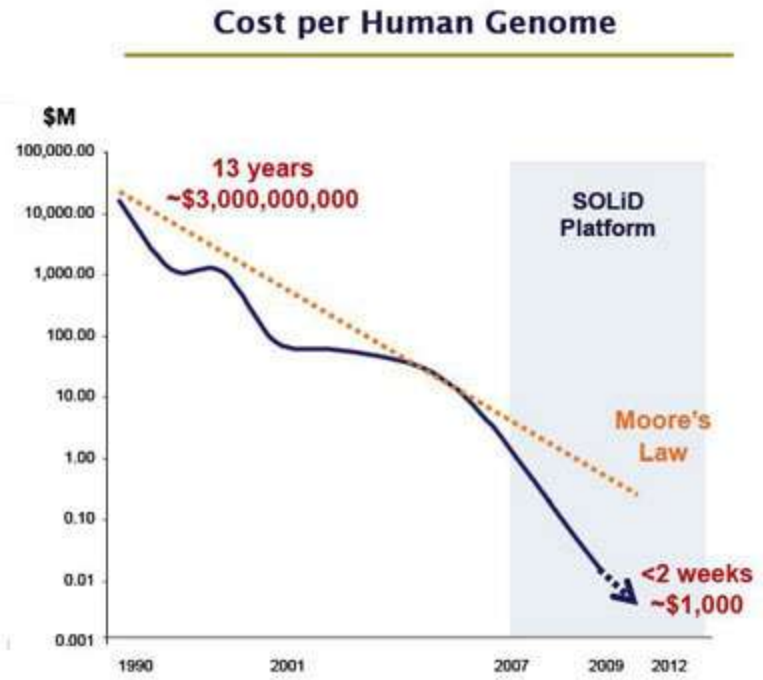


Exploratory multivariate analysis of genome scale data ...

Aedín Culhane
aedin@jimmy.harvard.edu

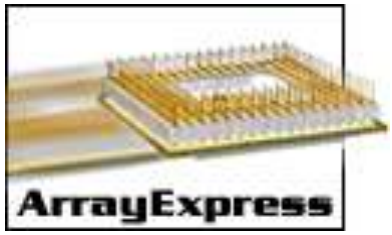
Dana-Farber Cancer Institute/Harvard School of Public Health.

Genome Information



Complete genomics

Gene Expression Data Repositories



ArrayExpress

- 21,997 Studies (622,617 profiles)

GEO Gene Expression Omnibus

GEO

- 22,735 Studies (558,074 profiles)

SMD

CIBEX

Statistics May 2011

How to get the data: GEO

- http://www.ncbi.nlm.nih.gov/geo/

- Accessions:
- GSE – Data Series
- GDS - Datasets
- GPL - Platform
- GSM - Sample

example

GEOquery

Download data directly from GEO into R

```
library(GEOquery)

geoD<- getGEO('GSE6324')  # processed data

cels<- getGEOSuppFiles('GSE6324 ')  # raw
   data
```

ArrayExpress

- AE take data from GEO, and implement MIAME more stringently
- Have experiment factor ontology for complex searches
- AE has nice browse function
- Searching AE - http://www.ebi.ac.uk/fg/doc/help/ae_help.html
- Many datasets are in Gene Expression Atlas (GXA) which is a fab resource ;-)) with a nice API

http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-6236

ArrayExpress

Download directly from ArrayExpress

> library (ArrayExpress)

queries the ArrayExpress database with keywords

> queryAE(“breast”)

> arrayexpress("E-TABM-1“)

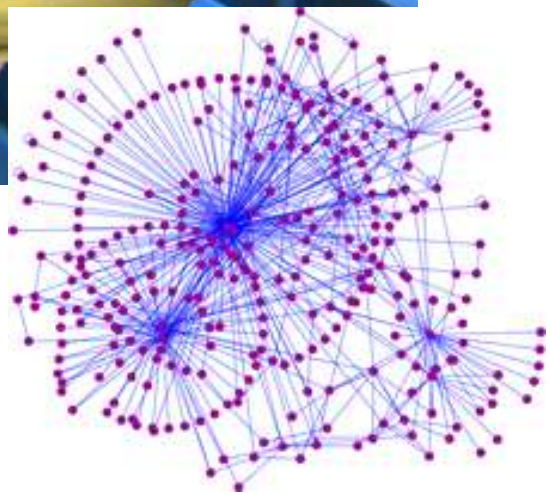
ArrayExpress

```
AEData<-getAE("E-TABM-1", type =
   "processed")

AERawData<-getAE("E-TABM-1", type =
   “raw")

AERawData<-getAE("E-TABM-1", type =
   “full")
```

So you got the data……

How do you start to analyze it?

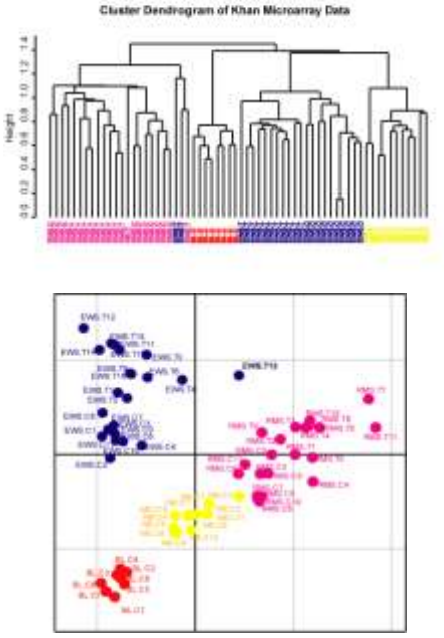
Why do we do exploratory data analysis?



- Genome scale data
- 10,000's variables
- Multivariate
- Essential to use exploratory data analysis to "get handle" on data

Exploration of Data is Critical

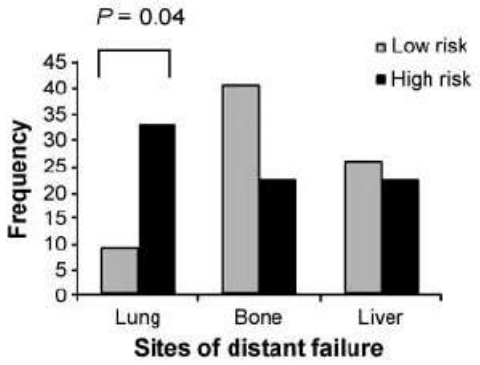
- Detect unpredicted patterns in data
- Decide what questions to ask
- Can also help detect cofounding covariates

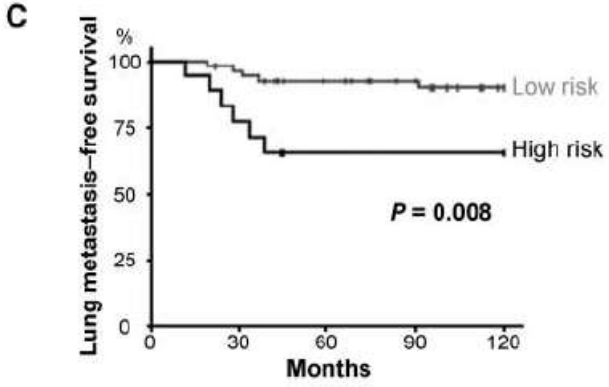


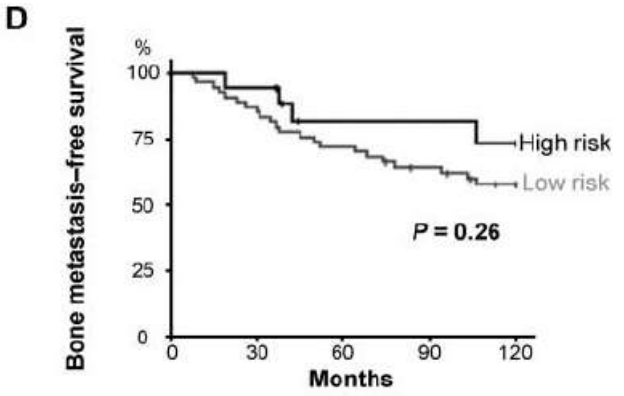
A 6 gene signature of lung metastasis

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

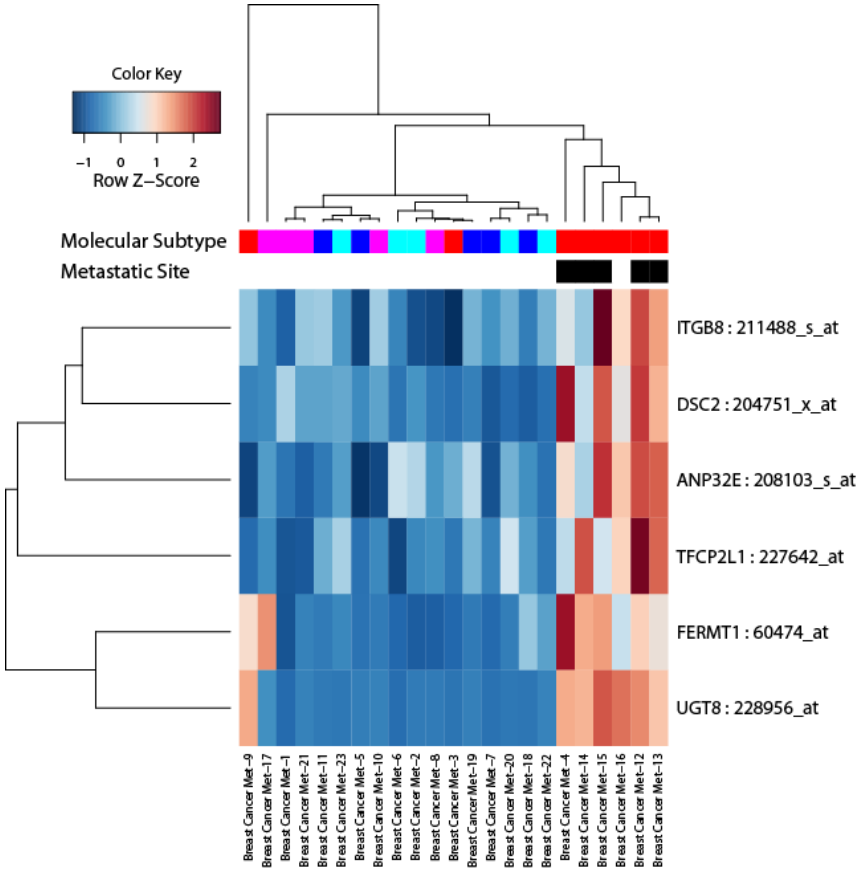






Landemaine T et al., Cancer Res. 2008 Aug 1;68(15):6092-9.

Confounding Covariates



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

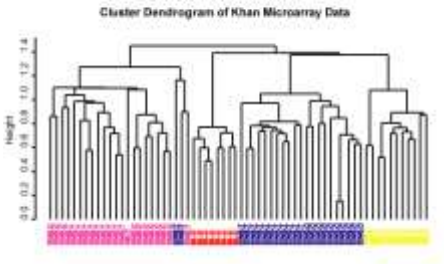
Supplementary Table 4. Results of Analysis of Global Test and GlobalAncova analysis of MSK dataset (p-value)

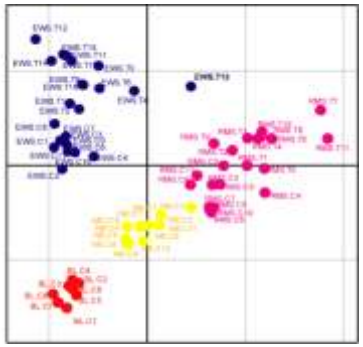
Method	globaltest	globaltest	GlobalAncova	GlobalAnova
Number of Probesets tested *	4	10	4	10
Q1: Are the genes associated with metastases status?	0.048	0.100	0.015	0.023
Q2: Are the genes associated with molecular subtype ?	<0.00000001	<0.0000001	0	0
Q3: Is metastases status significant independent of 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

Importance of Data Exploration

- Exploration of Data is Critical
- Clustering
  - Hierarchical
  - Flat (k-means)
- Ordination (Dimension Reduction)
  - Principal Component analysis, Correspondence analysis



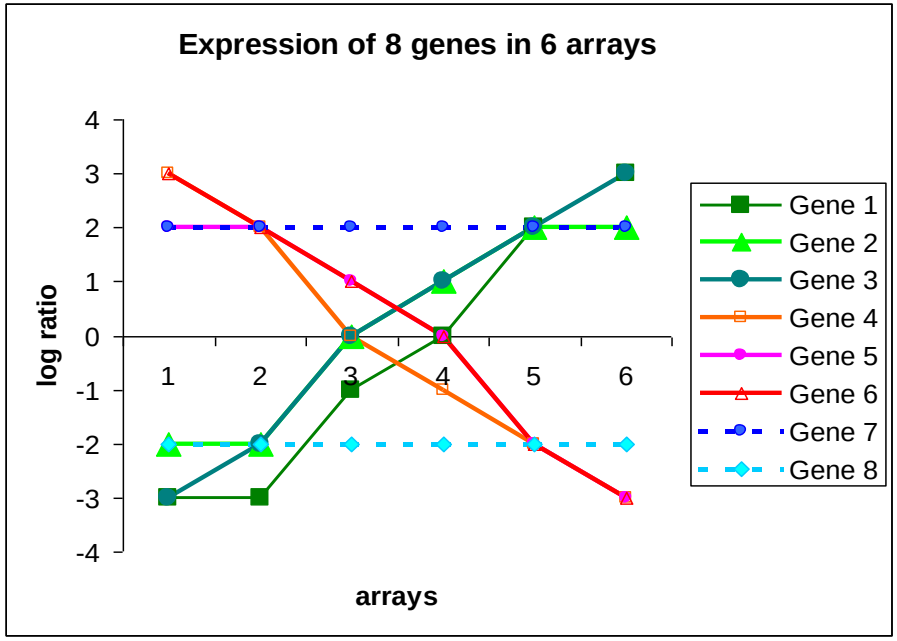


A Distance Metric

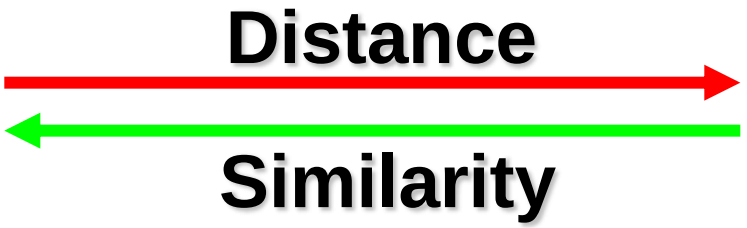
- In exploratory data analysis
  - only discover where you explore..

- The choice of metric is fundamental

8 Genes: Which is “closest”?



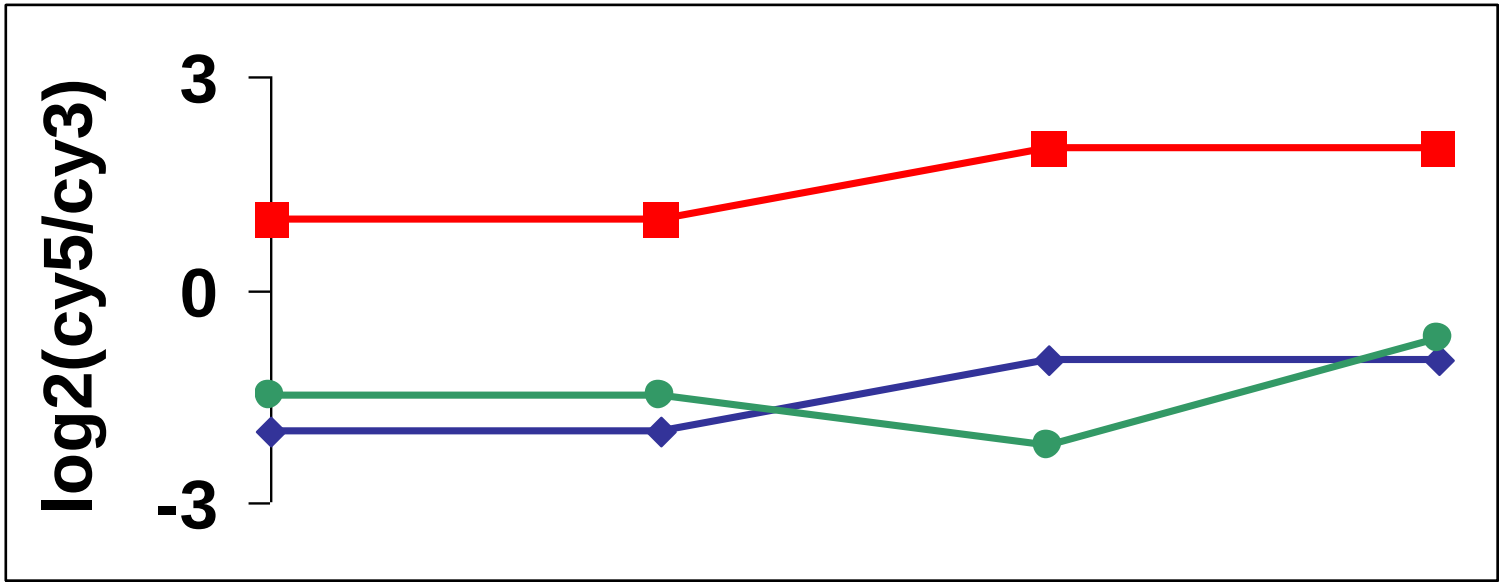
Distance Metrics



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

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

Distance Is Defined by a Metric



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

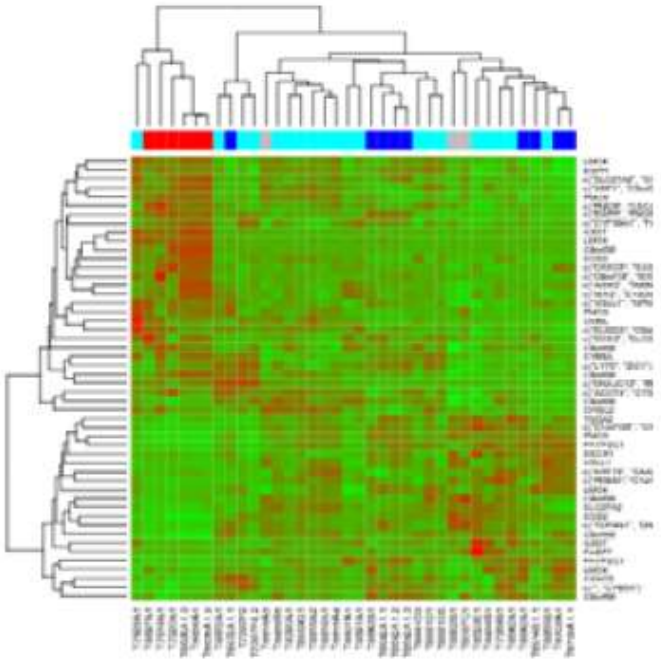
Cluster Analysis

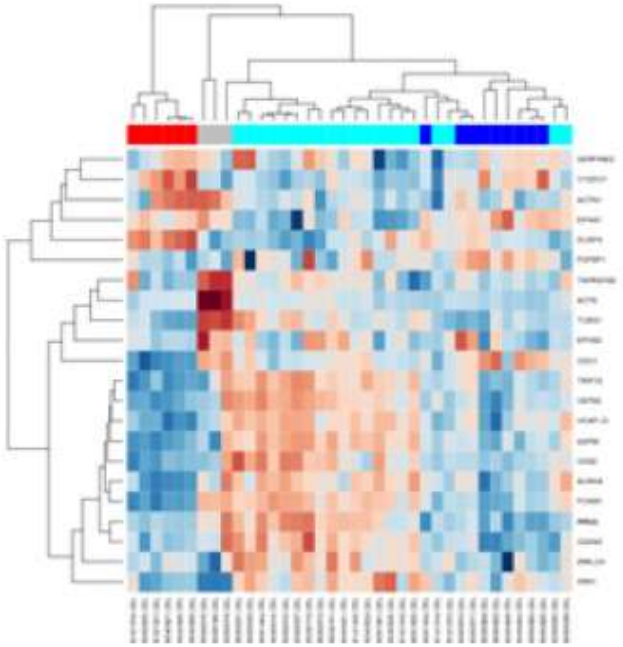
`dist()`

`hclust()`

`heatmap()`

`library(heatplus)`

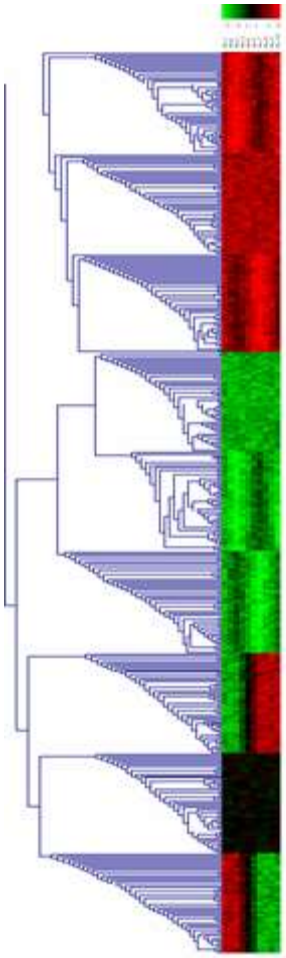


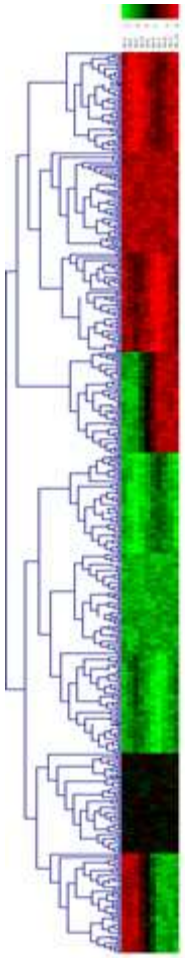


Relationships between these pairwise distances- Clustering Algorithms

- Different algorithms
  - Agglomerative or divisive
  - Popular **hierarchical agglomerative clustering** method
  - The distance between a cluster and the remaining clusters can be measured using minimum, maximum or average distance.
  - **Single lineage algorithm** uses the **minimum distance**.

Comparison of Linkage Methods



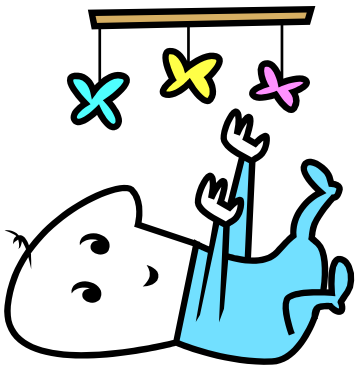


	Single	**Average**	**Complete**
Join by	min	average	max

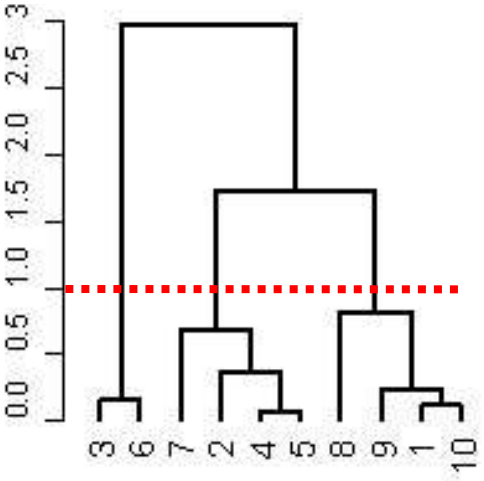
Quick Aside: Interpreting hierarchical clustering trees

Hierarchical analysis results viewed using a dendrogram (tree)

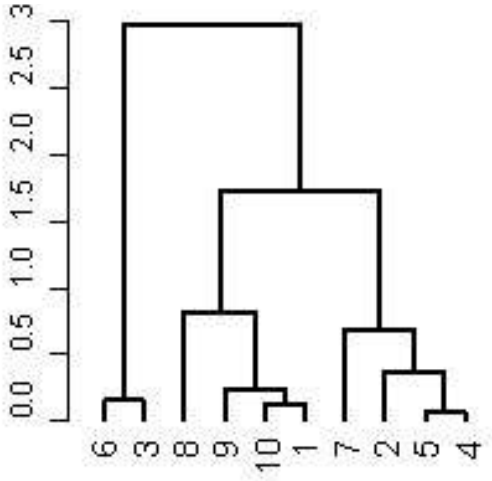
- Distance between nodes (Scale)
- Ordering of nodes not important (like baby mobile)



A



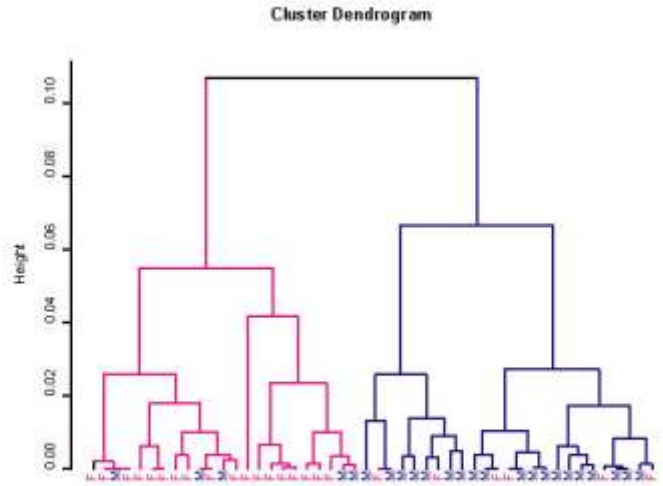
B

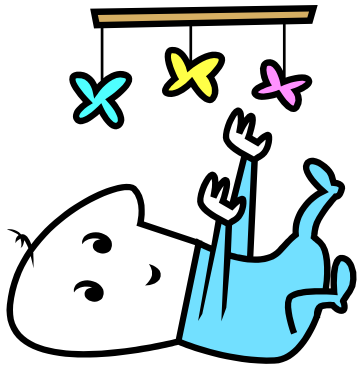


Tree A and B are equivalent

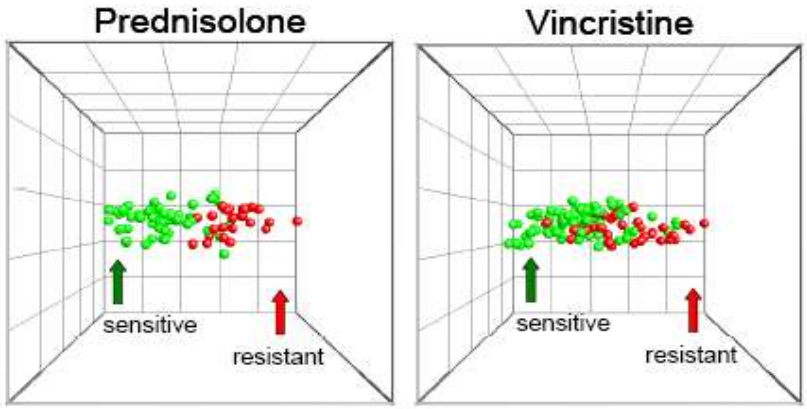
Limitations of hierarchical clustering

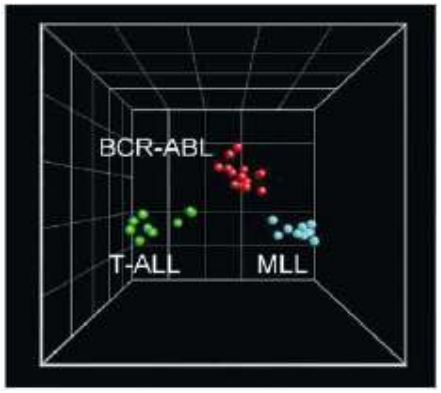
- Samples compared in a pair wise manner
- Hierarchy forced on data
- Sometimes difficult to visualise if large data
- Overlapping clustering or time/dose gradients ?

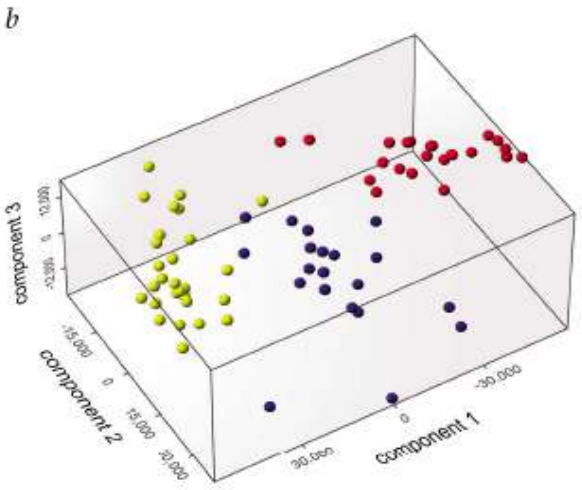


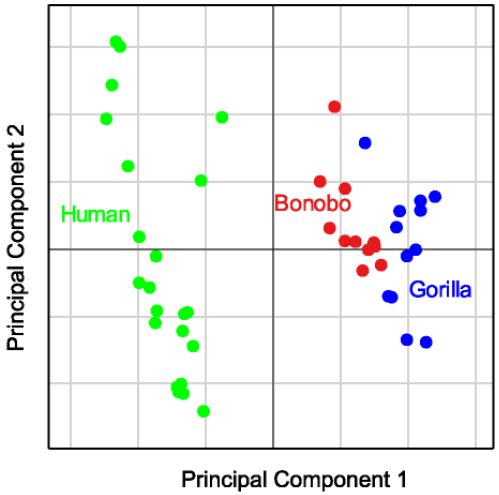


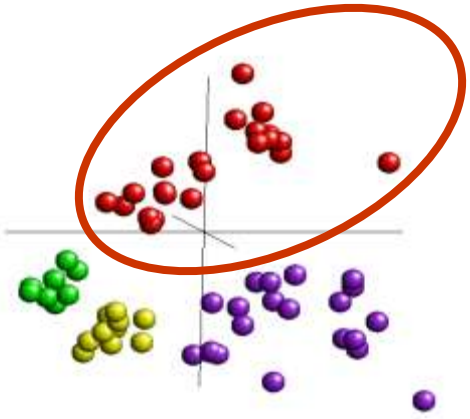
Ordination of Gene Expression Data



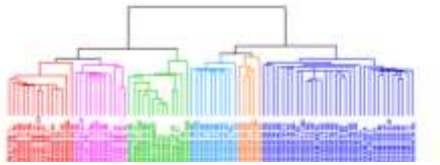




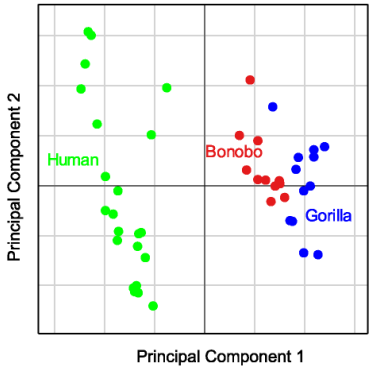




Complementary methods



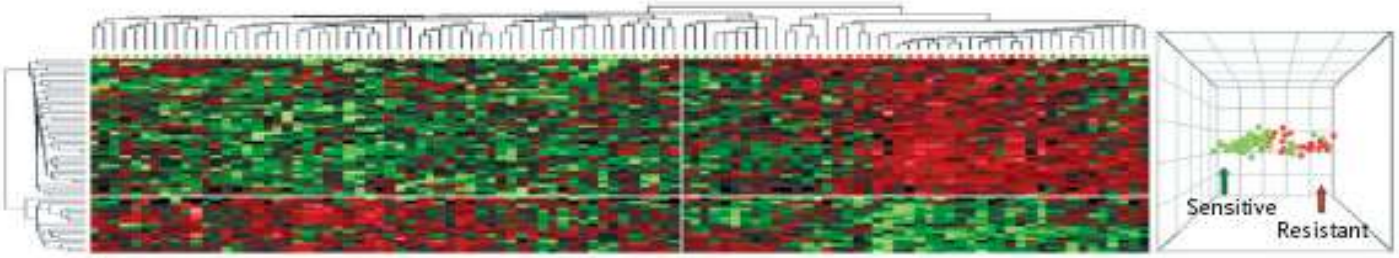
Cluster analysis generally investigates pairwise distances/similarities among objects looking for fine relationships

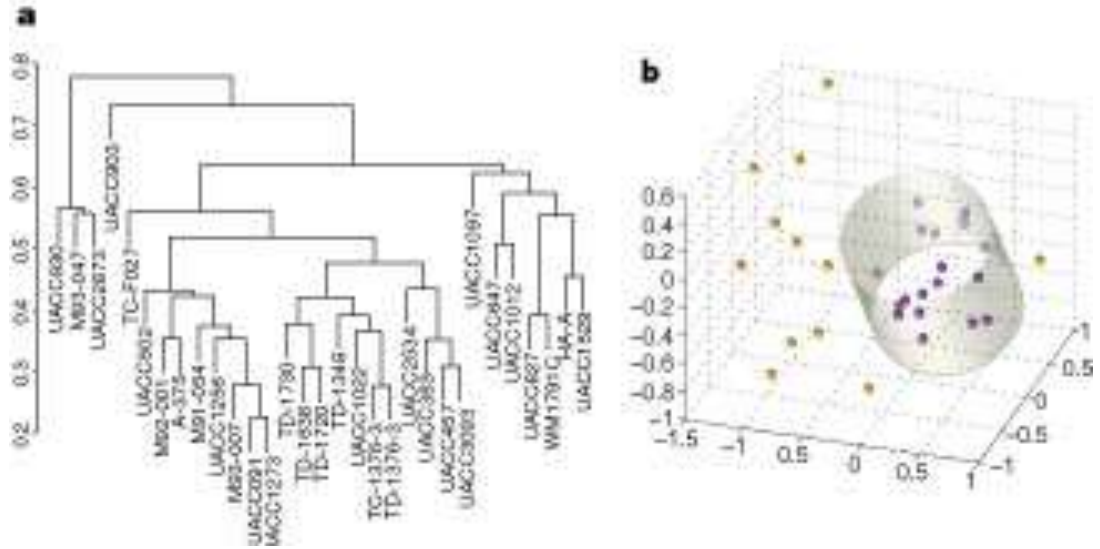


Ordination in reduced space considers the variance of the whole dataset thus highlighting general gradients/patterns

(Legendre and Legendre, 1998)

Many publications present both





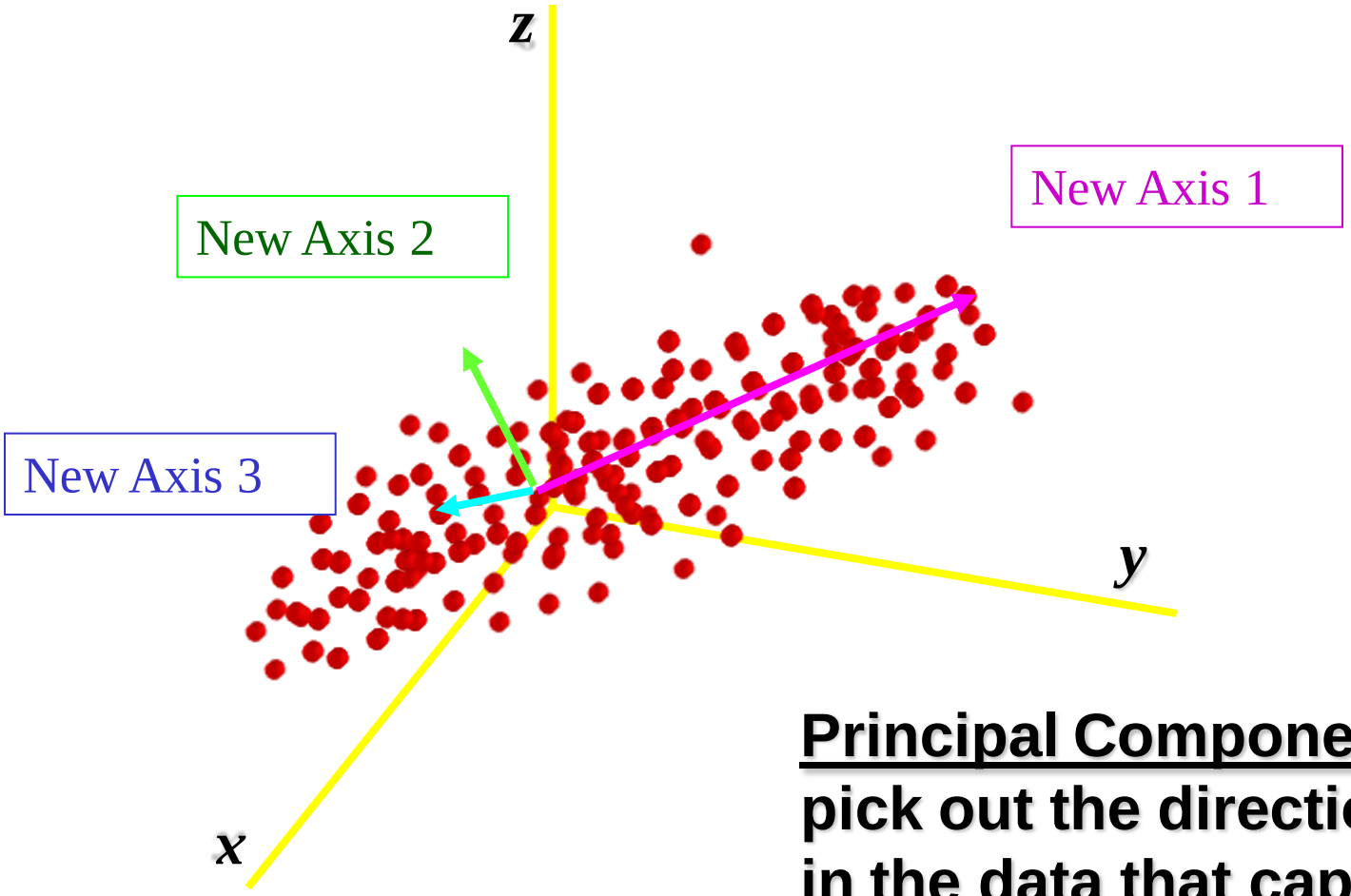
Ordination

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

- Aim:

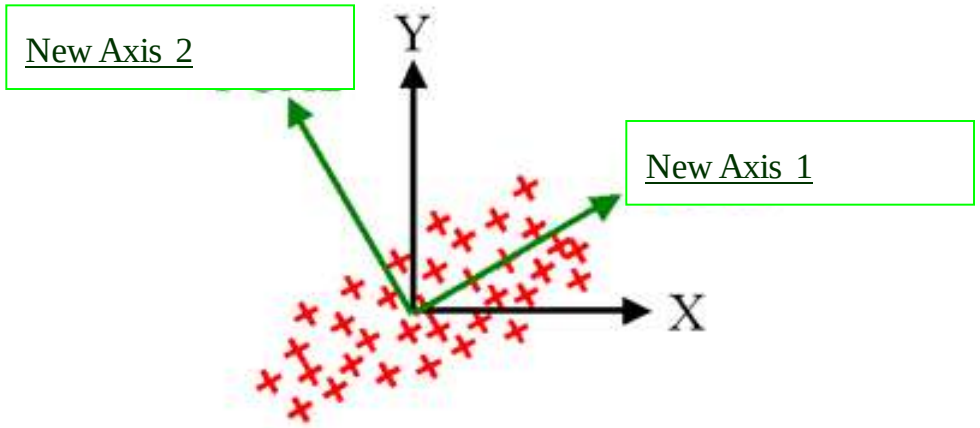
  Find axes onto which data can be project so as to explain as much of the variance in the data as possible

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

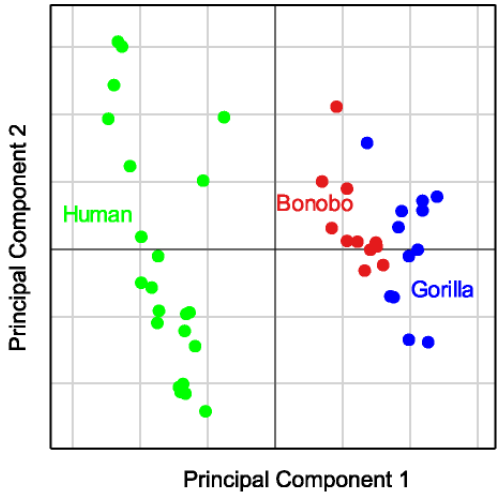
Interpreting an Ordination

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



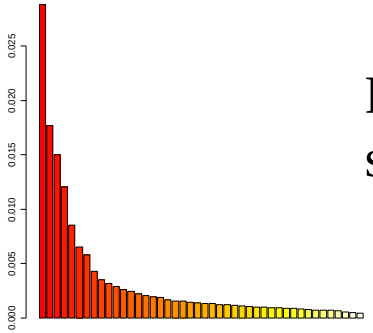
Principal Axes

- 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

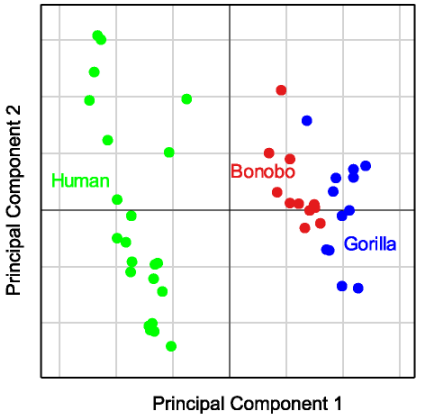


Plot of eigenvalues, select number.

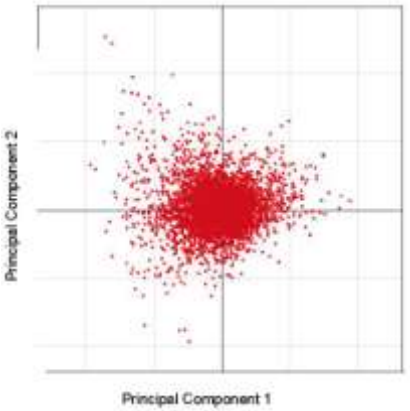
Plot PC1 v PC2

etc

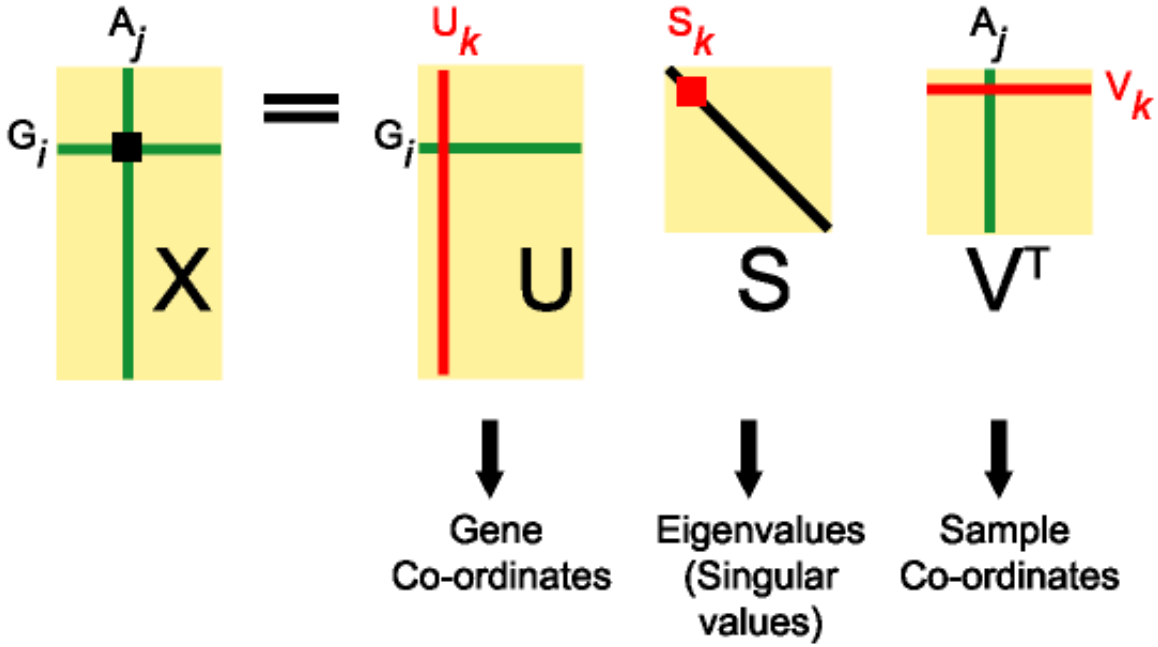
Array Projection



Gene Projection



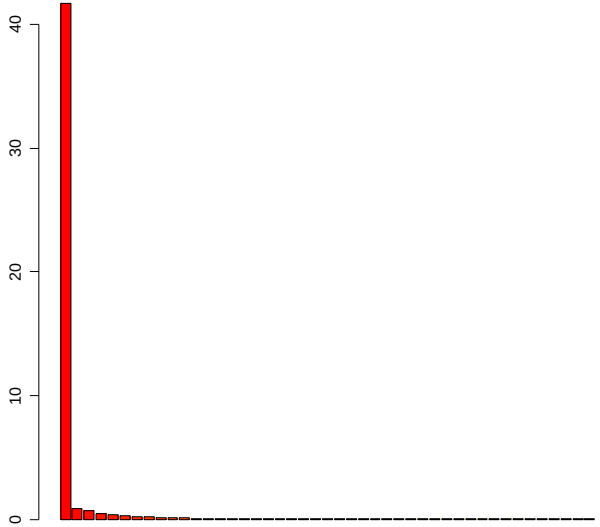
Singular Value Decomposition $X=USV^T$

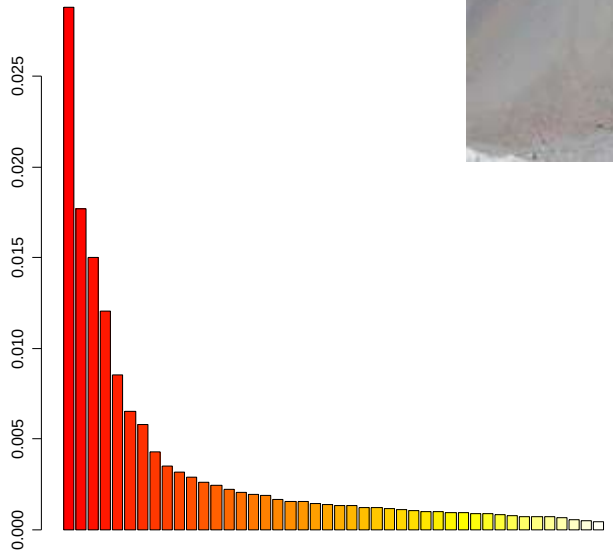


Eigenvalues

- 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



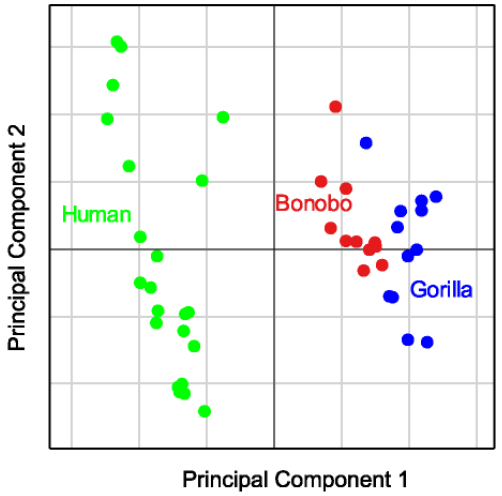


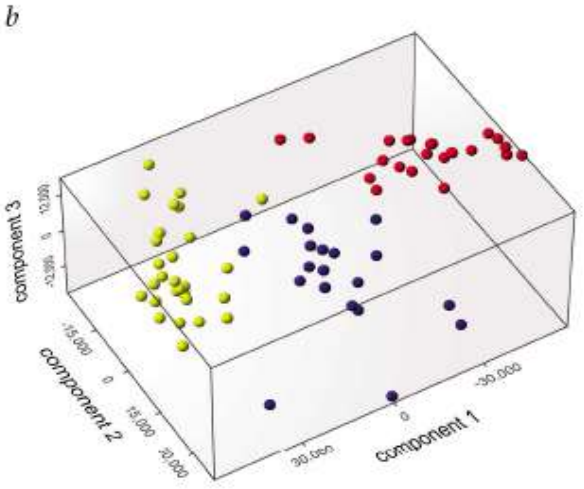
Maximum number of Eigenvalues/Eigenvectors = min(nrow, ncol) -1

Ordination Methods

Most common :

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





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)

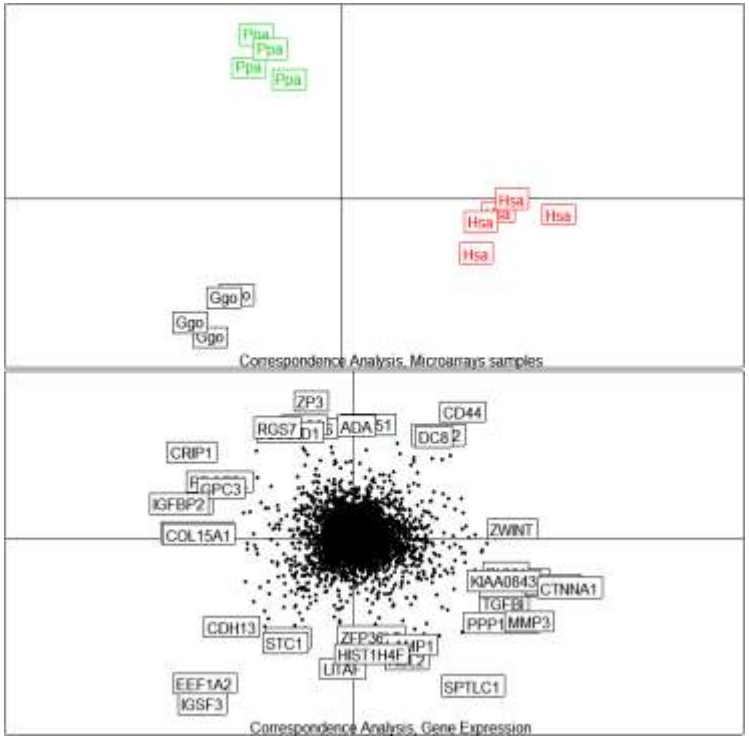
Summary: Exploration analysis using Ordination

- 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)



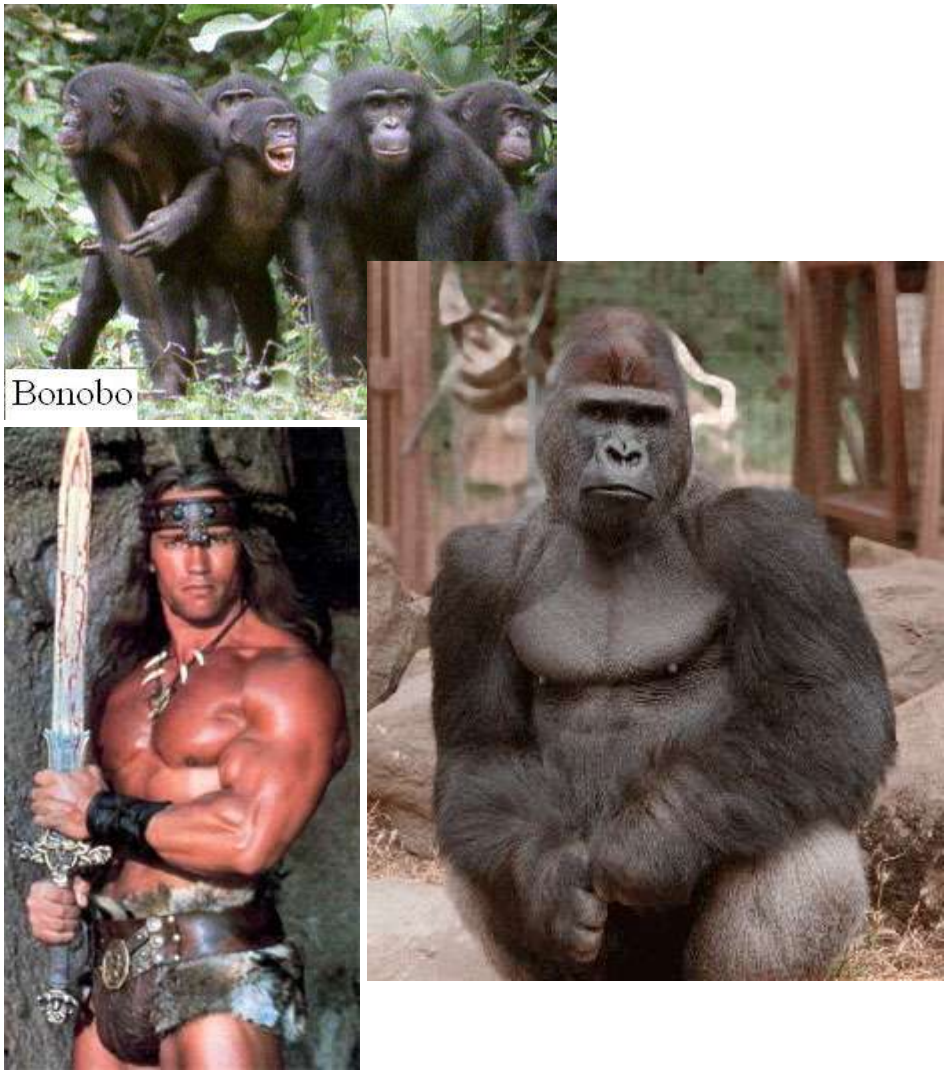
- `library(ade4)`
- `dudi.pca()`
- `dudi.coa()`

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

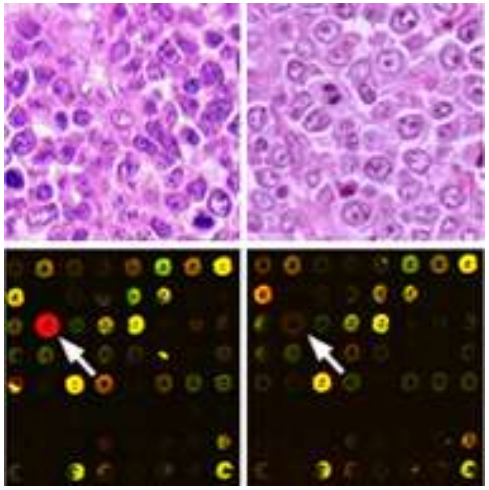
Link to example 3d html file

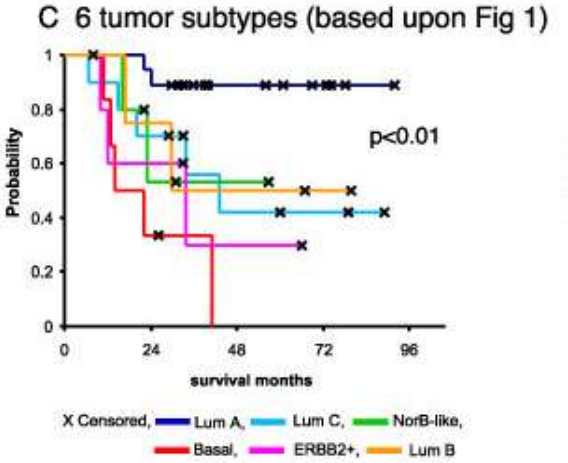
An Example and Comparison

- Karaman, Genome Res. 2003 13(7):1619-30.
- Compared fibroblast gene signature from 3 species

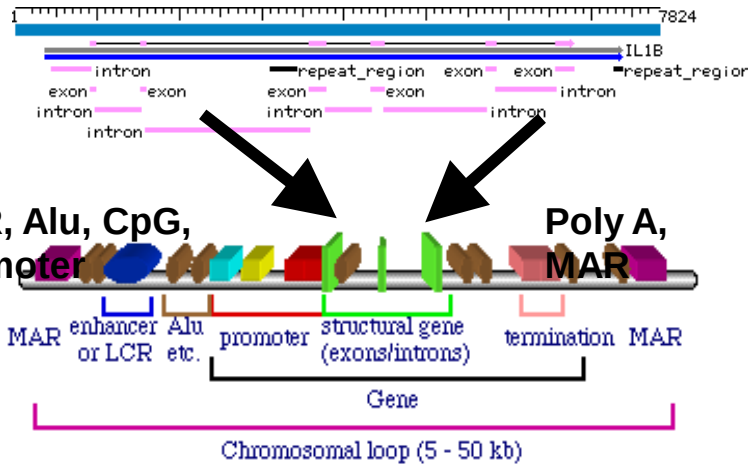


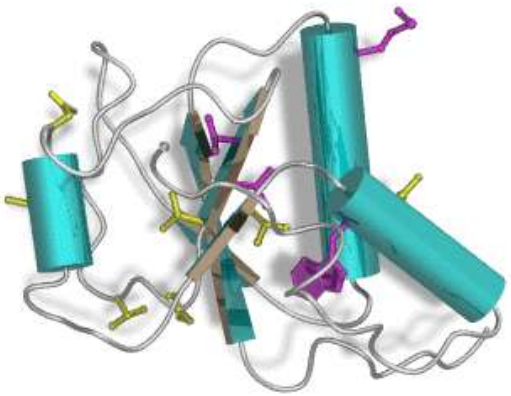
Integrate Data Sets?

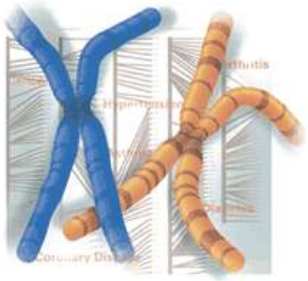




Coding Region : Introns, Exons and Internal Repeats







Multivariate Methods to detecting co-related trends in data

- Canonical correlation analysis
- Partial least squares
- Co-inertia analysis

Coinertia Analysis

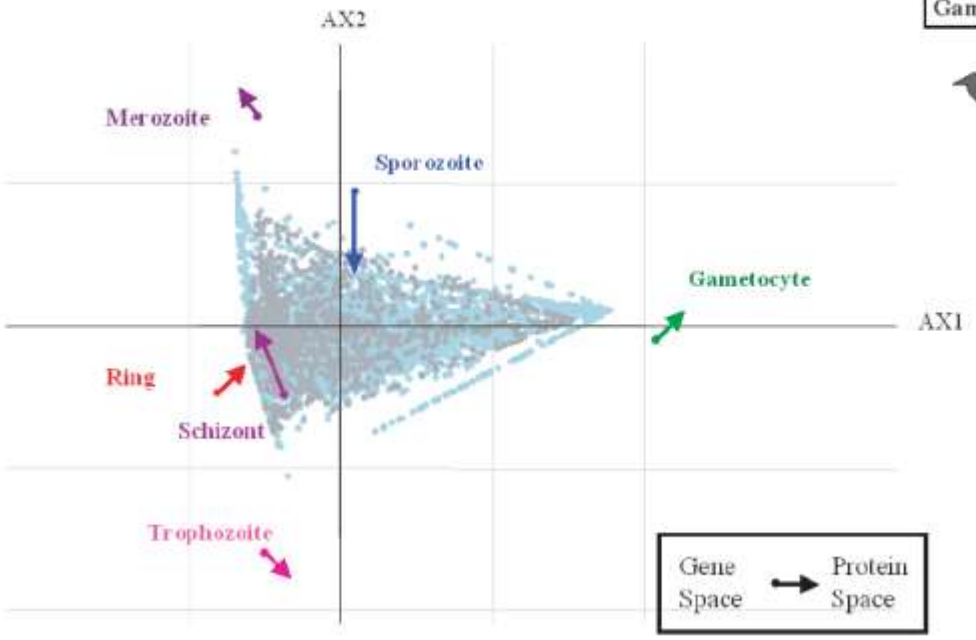
- Useful for cross-platform comparison where the same samples have been arrayed.
- Identifies correlated “trends” in data
- Consensus and divergence between gene expression profiles from different DNA microarray platforms are graphically visualised.
- Not dependent on annotation thus can extract important genes even when there are NOT present across all datasets.

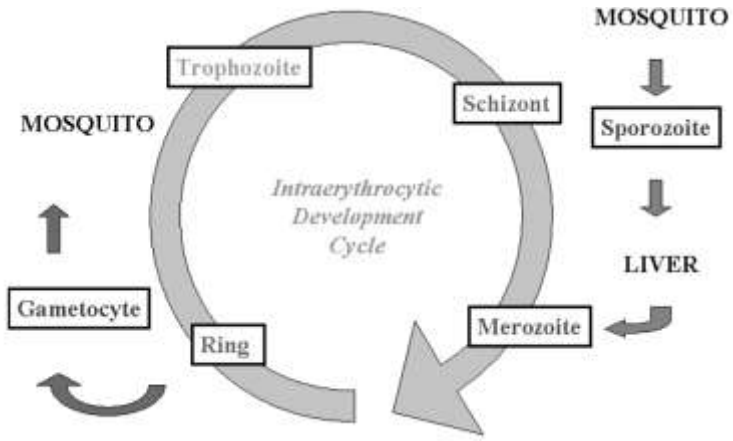
Culhane, A.C., Perriere, G., Higgins D.G., (2003) Cross platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, 4:59

Gene expression and proteomics data from the life cycle of the malarial parasitic.

Sample with variables (tri-plot)

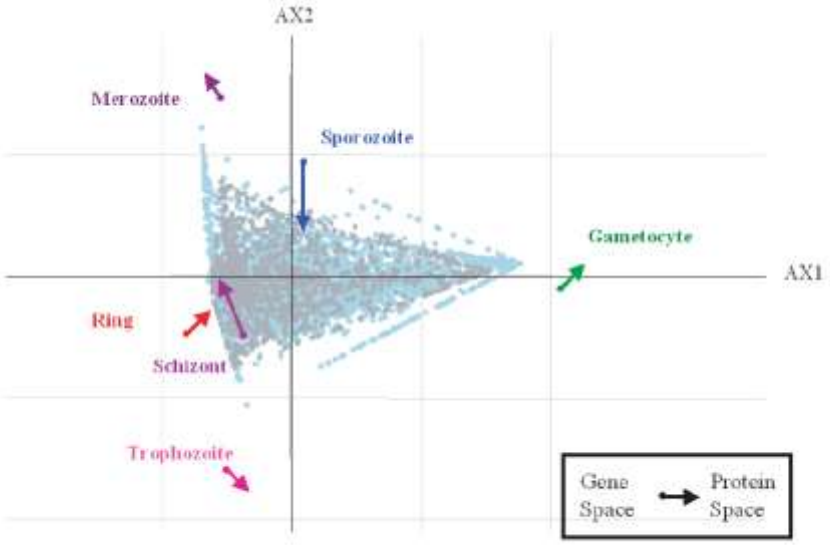
RV coefficient = 0.88.



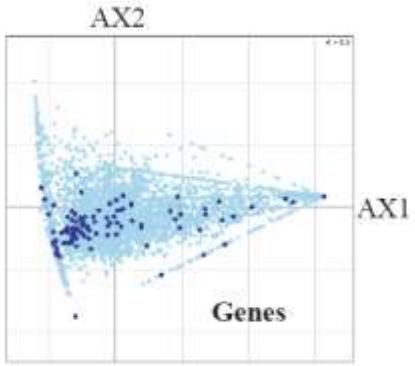


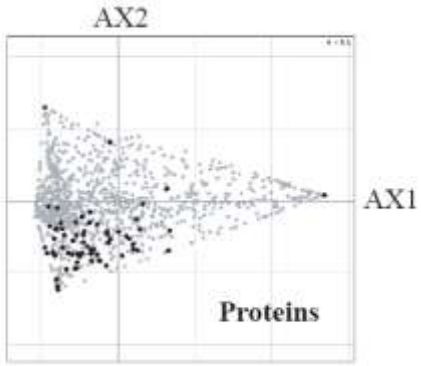
Project GO terms on Genes & Proteins space.

Sample with variables (tri-plot)

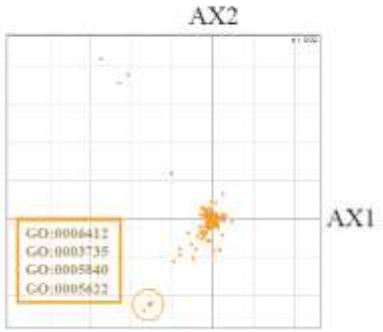


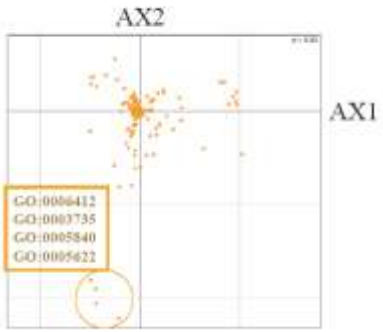
Variables





GO Terms





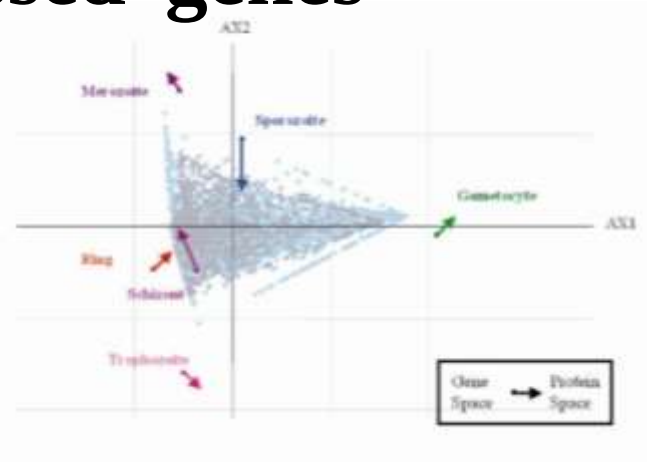
Axis 1 (horizontal) Accounts for 24.6% variance. Splits sexual & asexual life stages

Axis 2 (vertical) 4.8% variance. Splits invasive stages (Merozoite and Sporozoite stages which invade red blood)

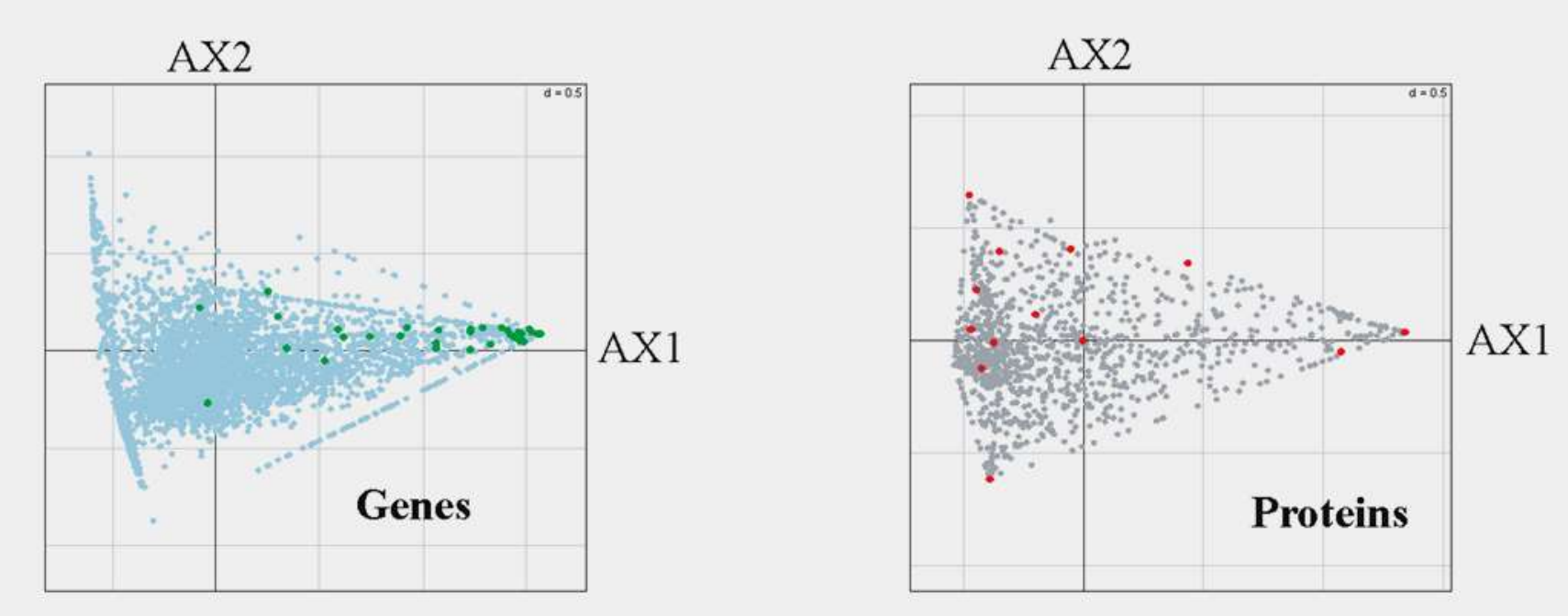
Detecting translationally repressed genes

Known: translationally repressed in female Gametocyte stage of *Plasmodium berghei*. These genes silence in the gametocyte stage but once ingested by mosquito, undergo translation into their respective proteins.

Examined *Plasmodium falciparum* orthologs



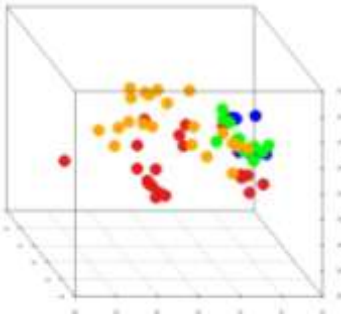
CIA: See genes transcriptionally active but their protein product is absent in the gametocyte stage.



Visualising Genes, Proteins and GO terms

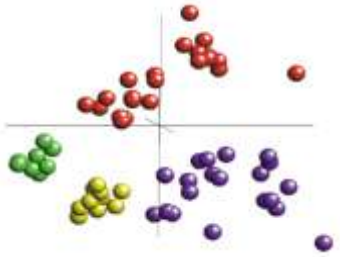
- CIA useful particularly to visualize variant "opposing" trends
- Addition of GO terms may assist when lack protein annotation (MS/MS data)
- Can be extended to supplement any annotation terms.

Fagan A, **Culhane AC**, Higgins DG. (2007) A Multivariate Analysis approach to the Integration of Proteomic and Gene Expression Data. *Proteomics.* 7(13):2162-71.



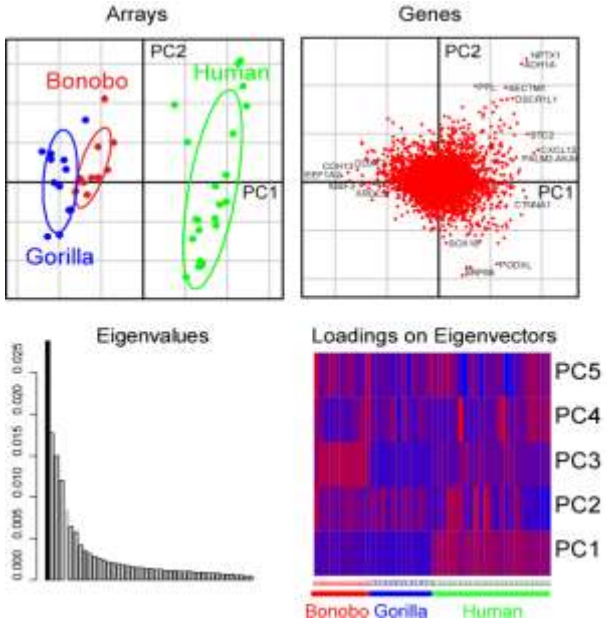
MADE4

An extension to the multivariate statistical package ade4 for microarray data analysis



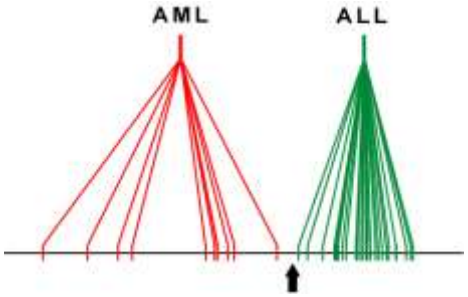
Exploratory Analysis Ordination

Correspondence Analysis, Principal Component Analysis



Supervised Class Prediction

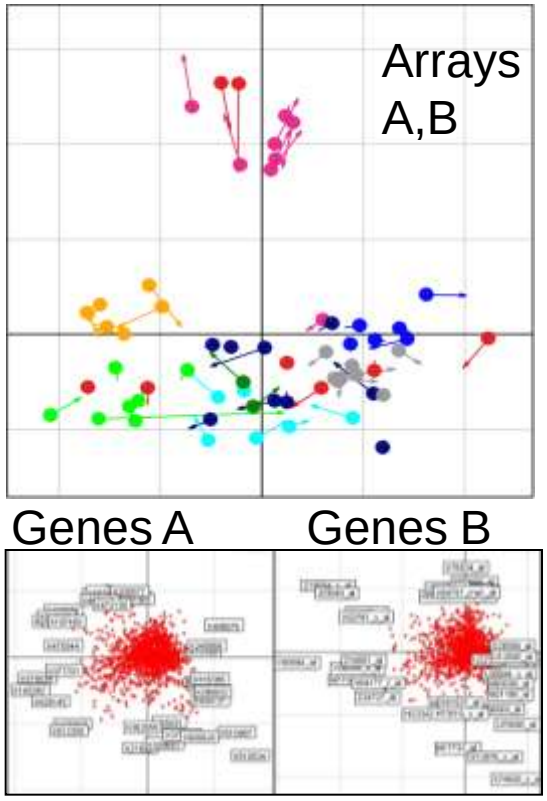
Between Group Analysis



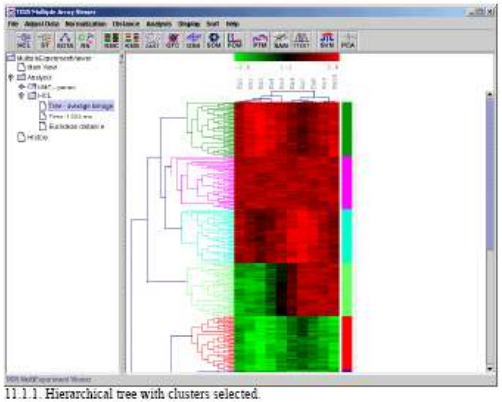
Culhane AC, Thioulouse J, Perriere G, Higgins DG. 2005 *Bioinformatics* **21(11):**2789-90.

Visualisation and integration of datasets

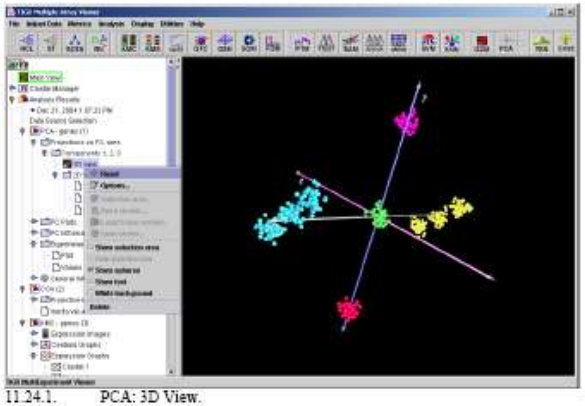
Coinertia Analysis



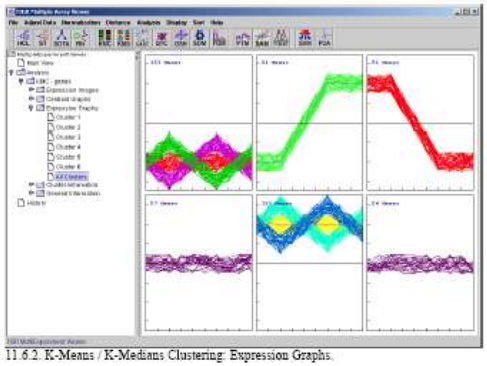
Desktop Package: mev
www.tm4.org



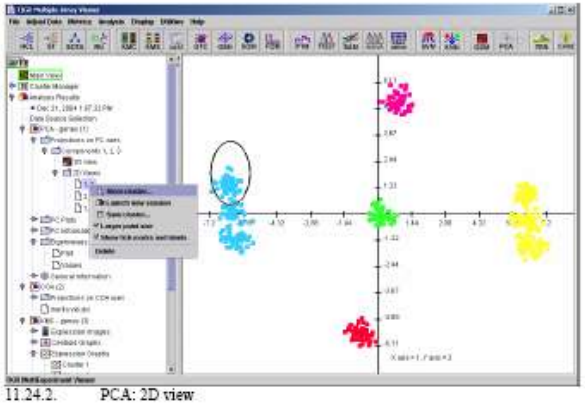
11.1.1. Hierarchical tree with clusters selected.



11.24.1. PCA: 3D View.



11.6.2. K-Means / K-Medians Clustering: Expression Graphs.



11.24.2. PCA: 2D view

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