Integrative Omics Data Analysis

Alex Sánchez

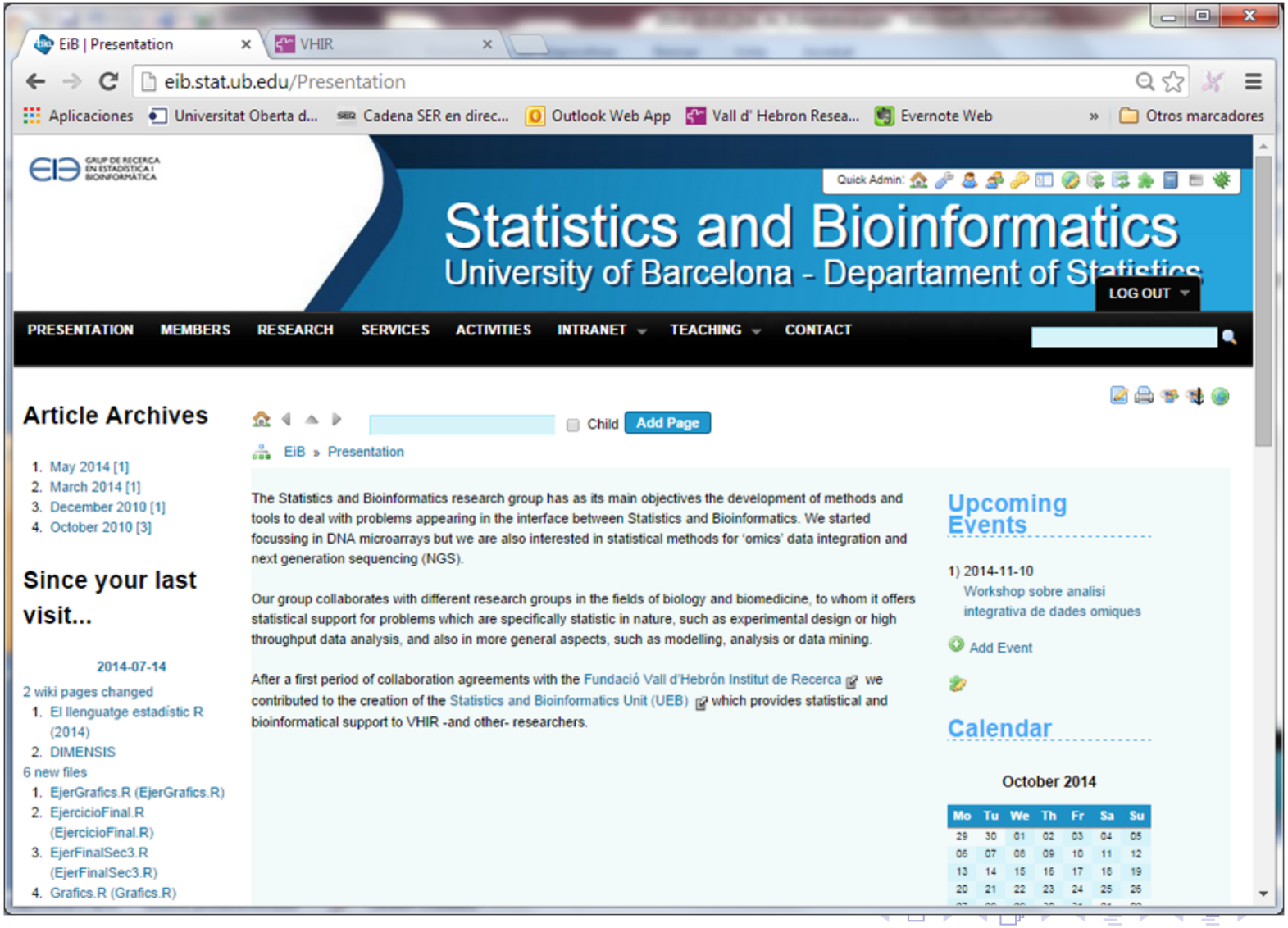
Statistics and Bioinformatics Research Group
Statistics department, Universitat de Barelona.
Statistics and Bioinformatics Unit Vall d'Hebron Institut de Recerca

November 3, 2015

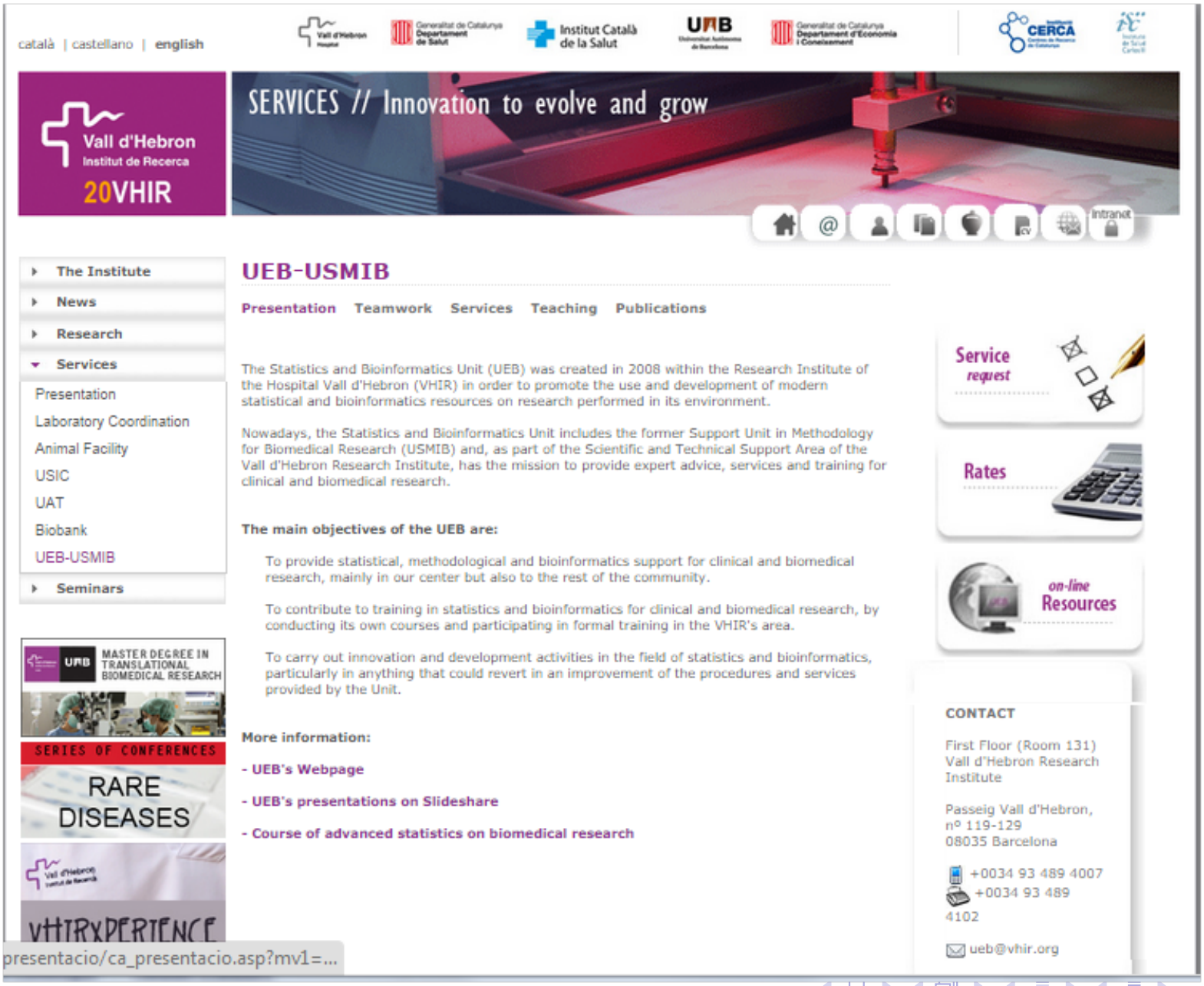
Table of Contents

Presentation and Objectives

Who, where, what?




$t = \frac{R_g}{\sqrt{\frac{d_0 \cdot SE_0^2 + d \cdot SE_g^2}{d_0 + d}}}$
Kernel PCA
Gene
2nd Principal Component
1st Principal Component
BIOCONDUCTOR
markdown
ubuntu
SerbGO
Searching the best GO Tool


català | castellano | english
Vall d'Hebron Institut de Recerca 20VHIR
SERVICES // Innovation to evolve and grow
Intranet
The Institute
News
Research
Services
Presentation
Laboratory Coordination
Animal Facility
USIC
UAT
Biobank
UEB-USMIB
Seminars
UEB-USMIB
Presentation Teamwork Services Teaching Publications
The Statistics and Bioinformatics Unit (UEB) was created in 2008 within the Research Institute of the Hospital Vall d'Hebron (VHIR) in order to promote the use and development of modern statistical and bioinformatics resources on research performed in its environment.
Nowadays, the Statistics and Bioinformatics Unit includes the former Support Unit in Methodology for Biomedical Research (USMIB) and, as part of the Scientific and Technical Support Area of the Vall d'Hebron Research Institute, has the mission to provide expert advice, services and training for clinical and biomedical research.
The main objectives of the UEB are:
To provide statistical, methodological and bioinformatics support for clinical and biomedical research, mainly in our center but also to the rest of the community.
To contribute to training in statistics and bioinformatics for clinical and biomedical research, by conducting its own courses and participating in formal training in the VHIR's area.
To carry out innovation and development activities in the field of statistics and bioinformatics, particularly in anything that could revert in an improvement of the procedures and services provided by the Unit.
More information:
- UEB's Webpage
- UEB's presentations on Slideshare
- Course of advanced statistics on biomedical research
Service request
Rates
on-line Resources
CONTACT
First Floor (Room 131)
Vall d'Hebron Research Institute
Passeig Vall d'Hebron, nº 119-129
08035 Barcelona
+0034 93 489 4007
+0034 93 489 4102
ueb@vhir.org
MASTER DEGREE IN TRANSLATIONAL BIOMEDICAL RESEARCH
SERIES OF CONFERENCES
RARE DISEASES
VHIR EXPERIENCE
presentacio/ca_presentacio.asp?mv1=...


Vall d'Hebron

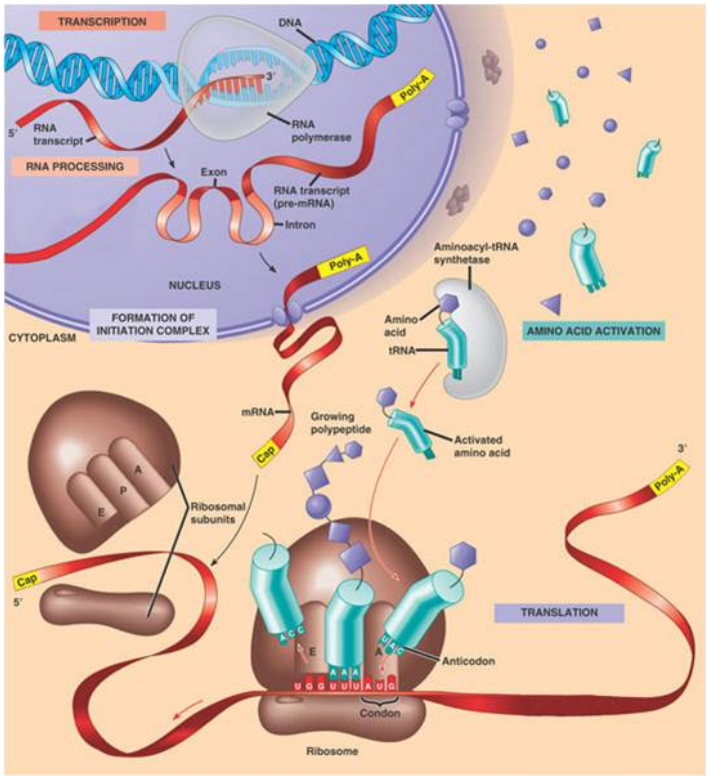
Interaction promotes opportunities

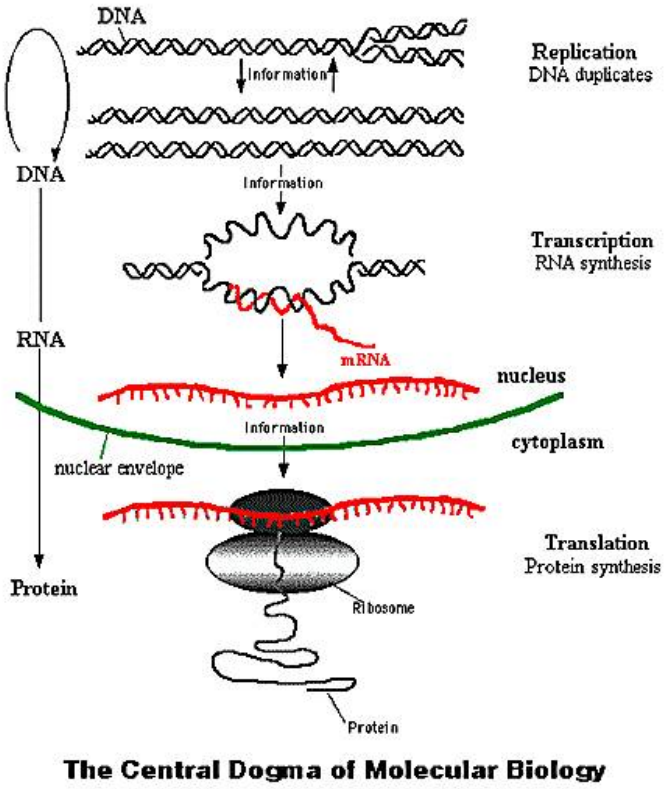
Omics, Omics Data and Omics Integration

What is "omics"?

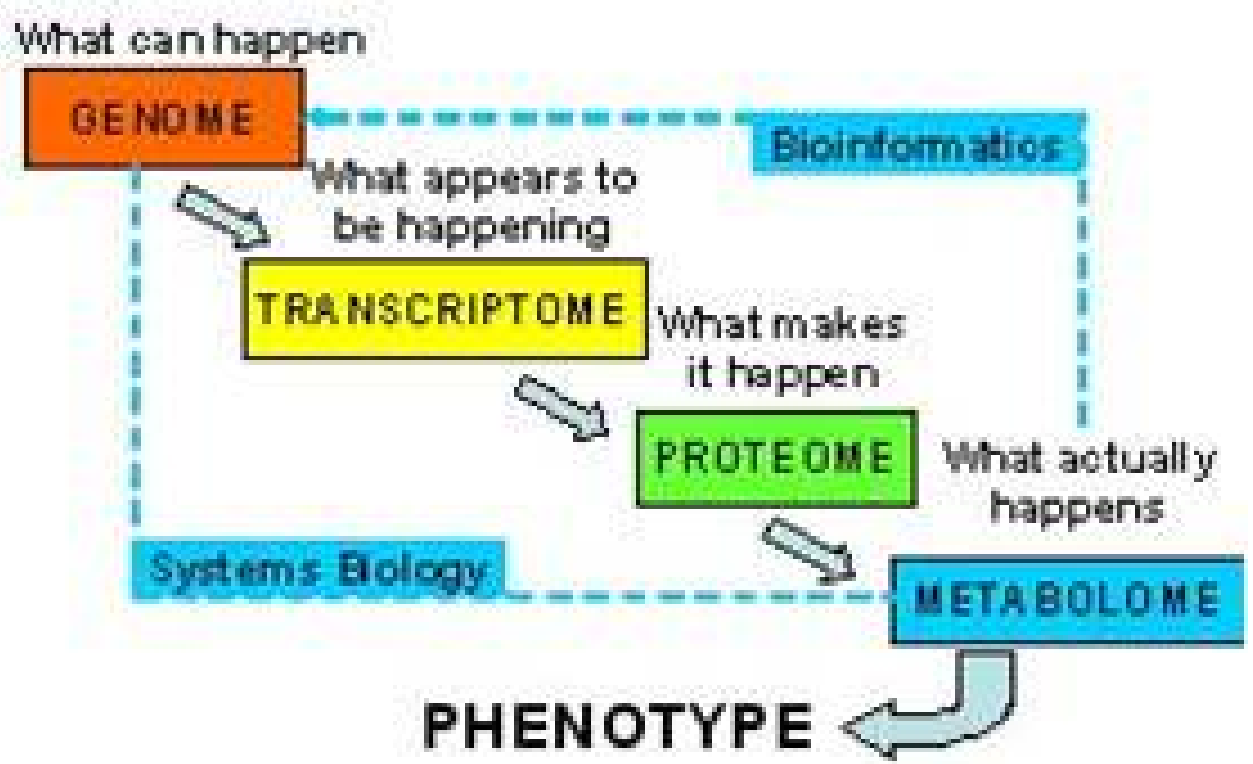
- In biological context , the suffix "omics" is used to refer to the study of large sets of biological molecules (Smith et al., 2005)
- The study of different components participating and/or regulating complex biological processes, triggered the development of several fields that, together, are described with the term OMICS.

The central dogma

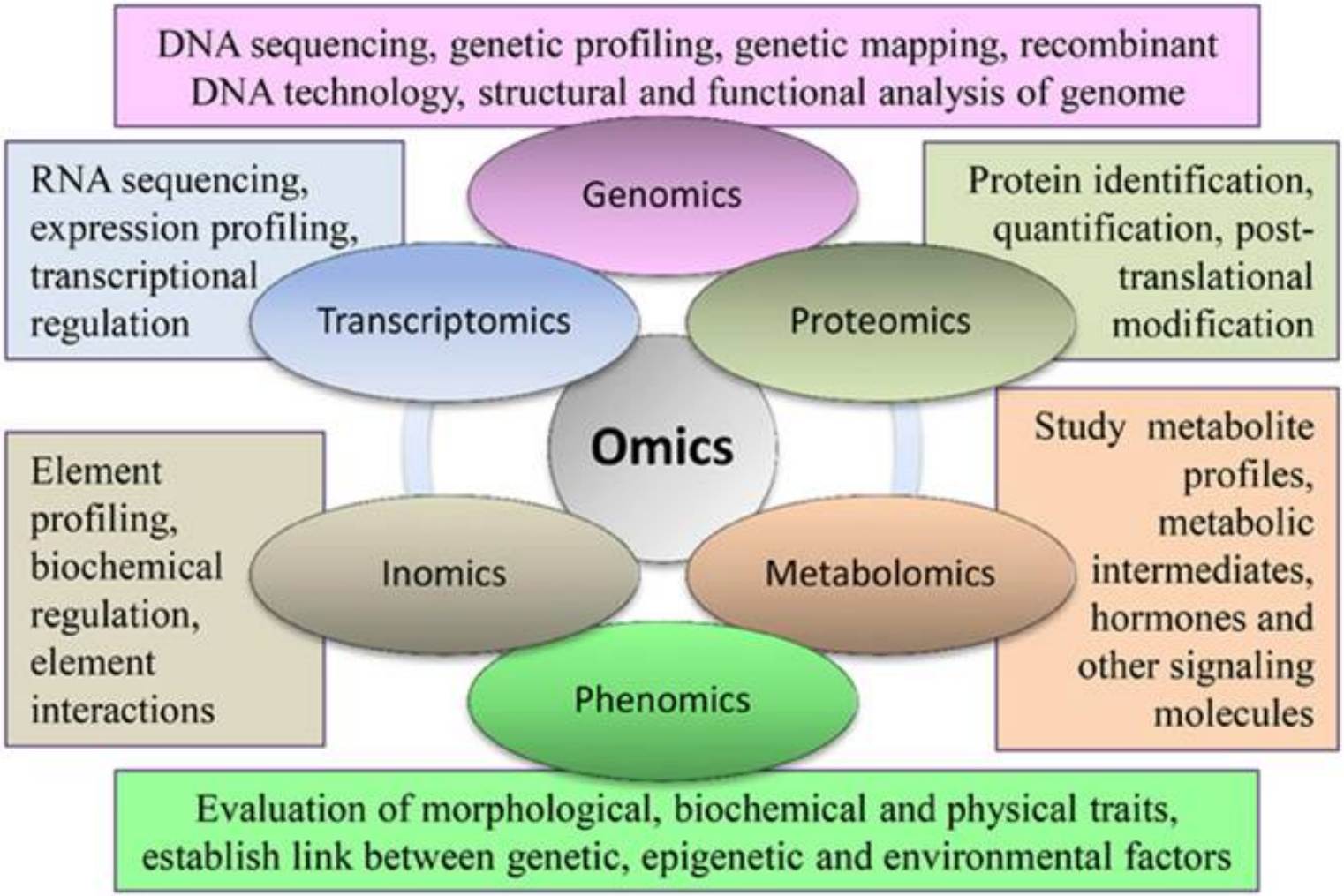




The Omics Cascade (1): "omes"

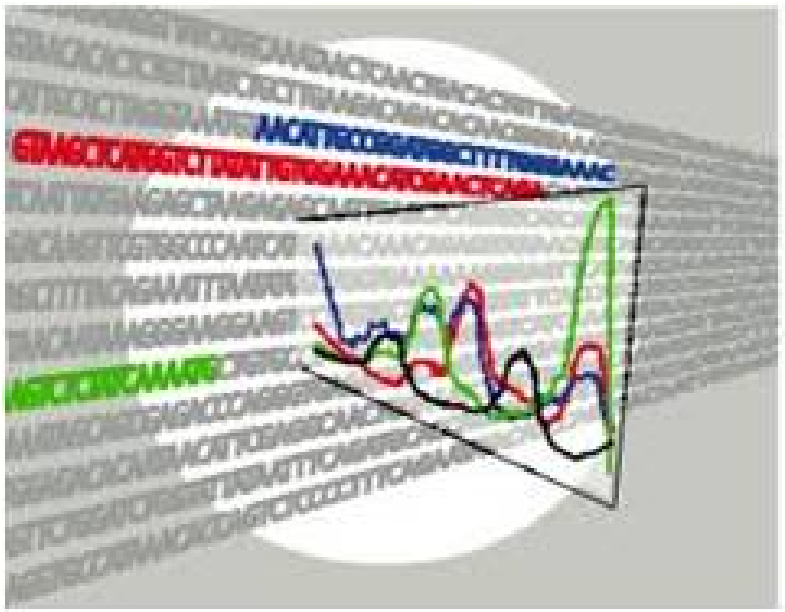


We study omes with omics



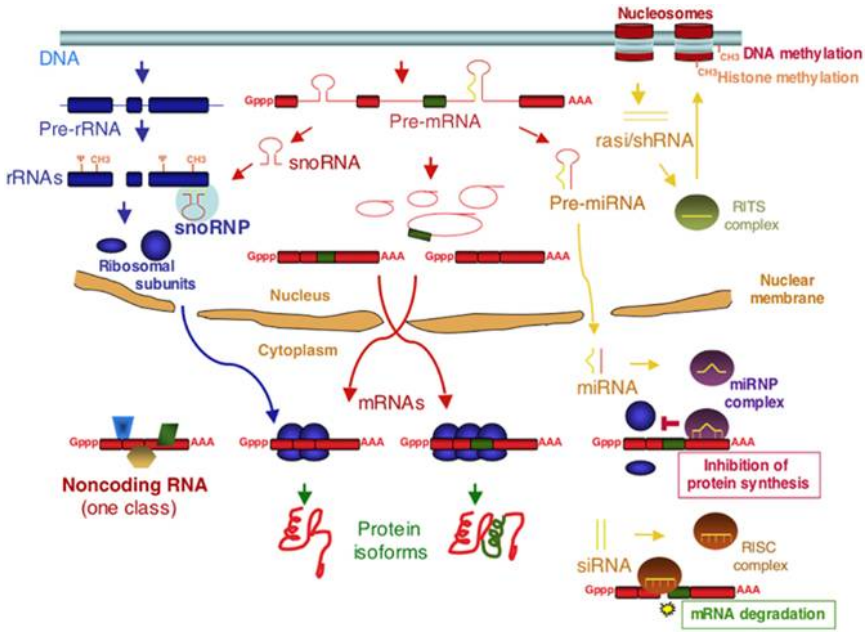
Genomics

Genomics is a discipline in genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of genomes (the complete set of DNA within a single cell of an organism)



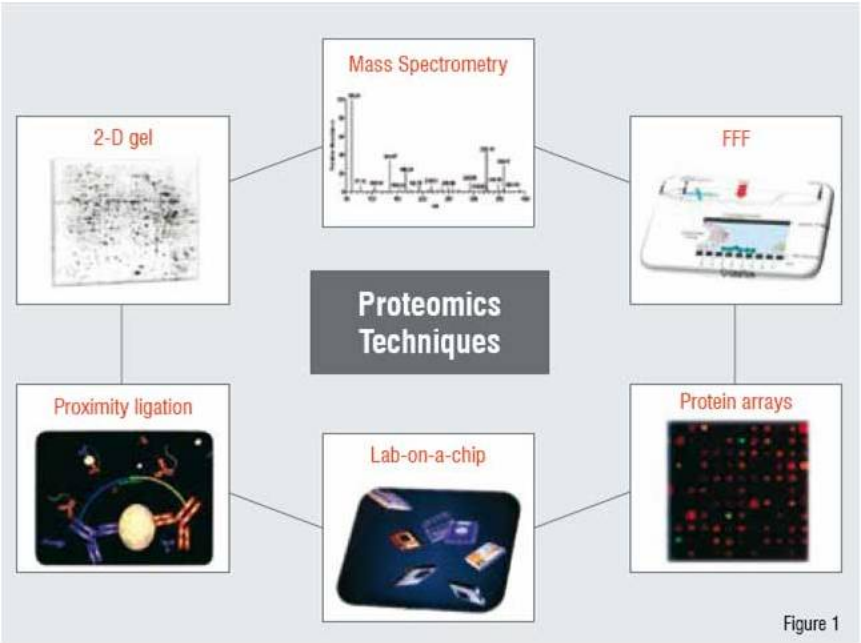
Transcriptomics

- The transcriptome is the set of all RNA molecules, in one or a population of cells.
- Transcriptomics, examines expression levels of mRNAs in a given cell population, often using high-throughput techniques: microarrays or NGS.



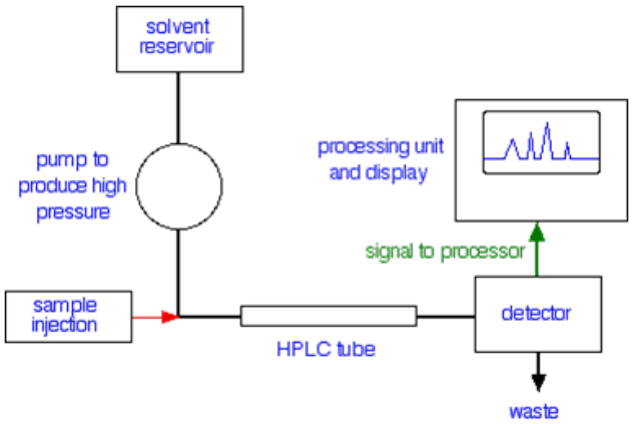
Proteomics

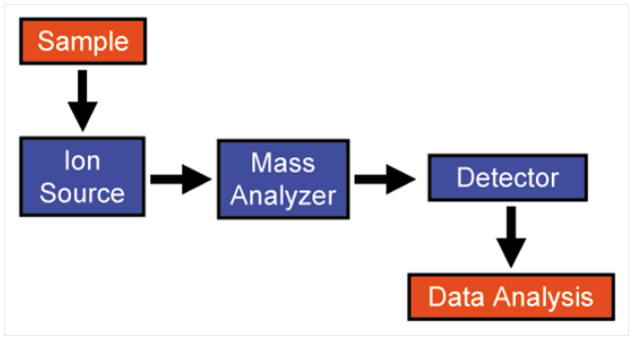
- The large–scale study of proteins (the proteome), particularly their structure and function.
- Relies on a wide spectra of techniques
  - 2D gel based
  - Mass Spectrometry (MS)
  - Seldi-TOF (MS)
  - Protein Arrays
  - ...



Metabolomics

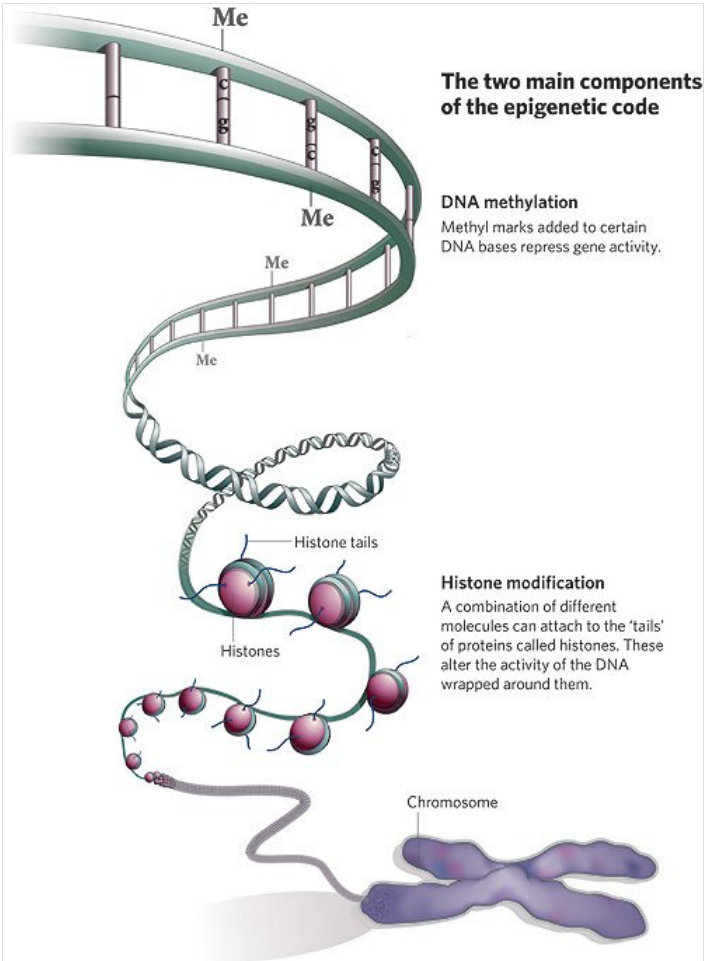
- Comprehensive and simultaneous systematic determination of
  - metabolite levels in the metabolome and
  - their changes over time as a consequence of stimuli.
- Relies on
  - Separation techniques: GC, CE, HPLC, UPLC
  - Detection techniques: NMR, MS



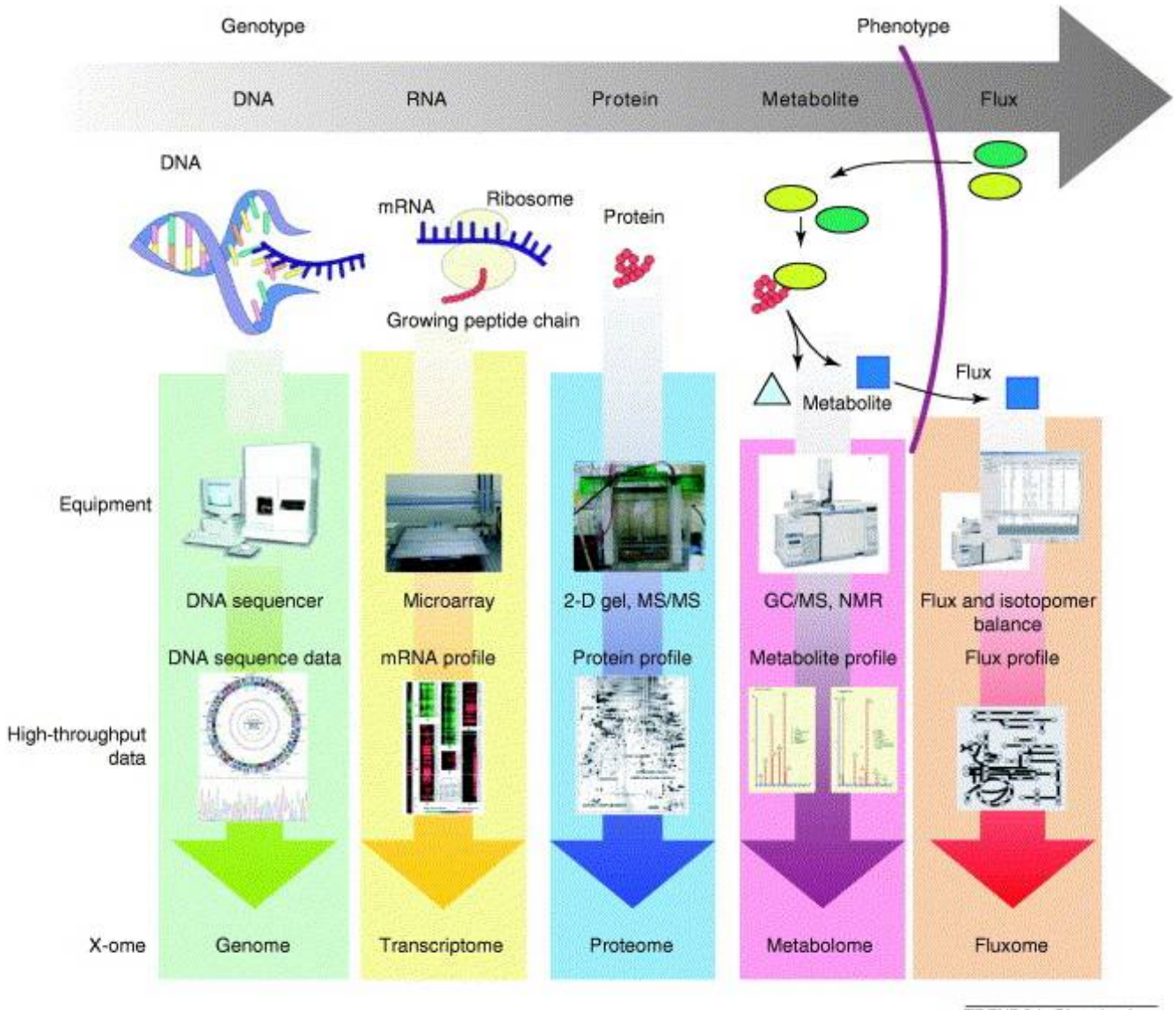


Epigenetics and Epigenomics

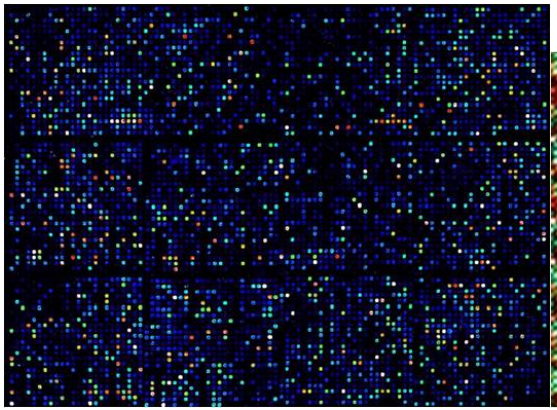
- *Epigenetics* is the study of changes in the phenotype or gene expression caused by other mechanisms than changes in the underlying DNA sequence.
  - DNA methylation
  - Histone modifications
- Epigenetics refers to the study of single genes or sets of genes. Epigenomics refers to global analyses of epigenetic changes across the entire genome

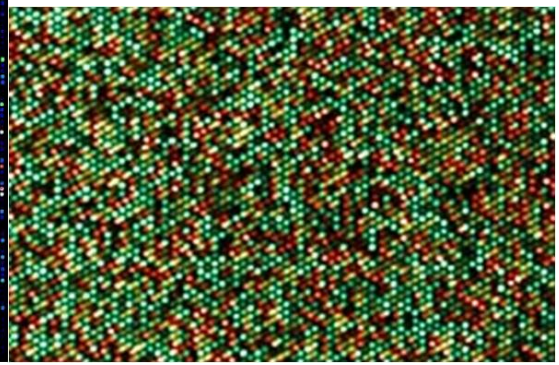


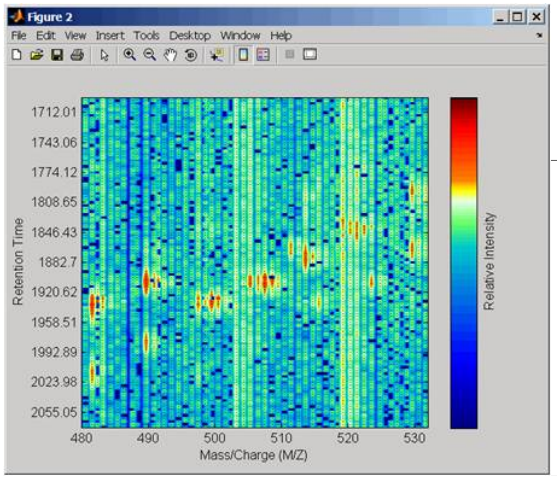
In summary omic to stuy omes

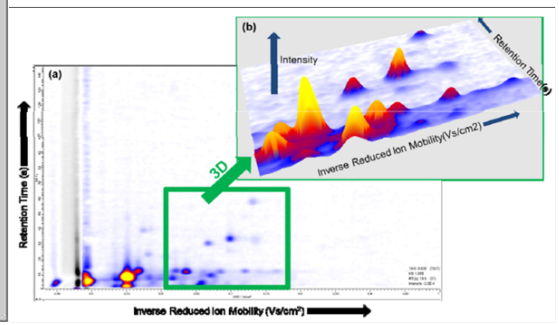


Omics data are high throughput









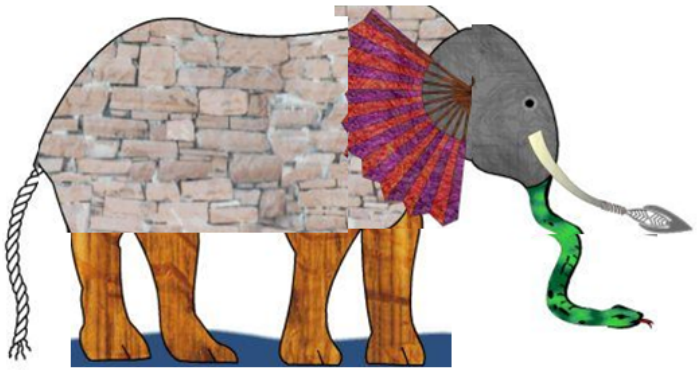
Bioinformatics and Biostatistics is essential

"Data don't make any sense,
we will have to resort to statistics."

Integrative Analysis and Data integration: methods, types, tools, challenges

Why should we integrate data?

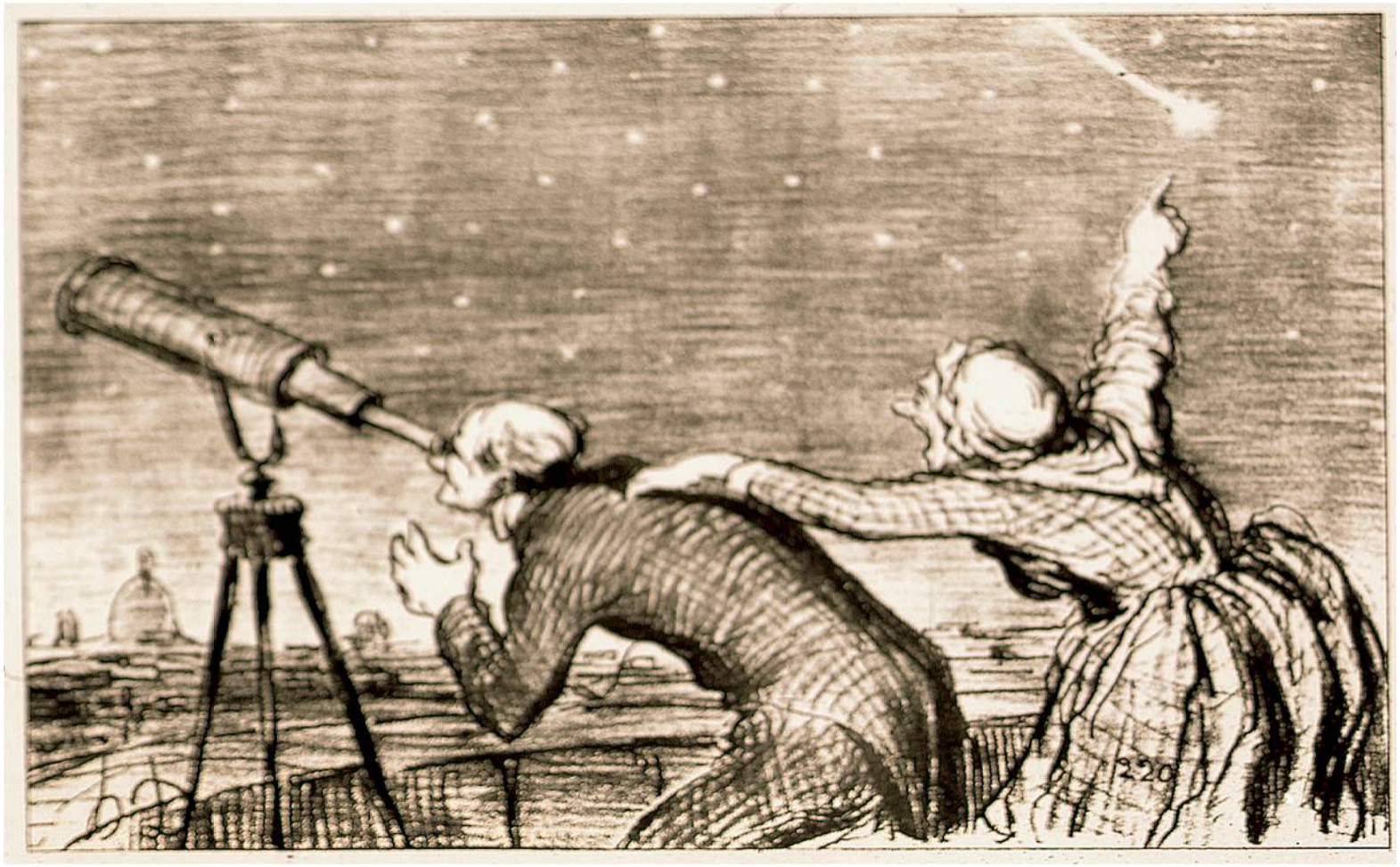
The Blind Men and the Elephant



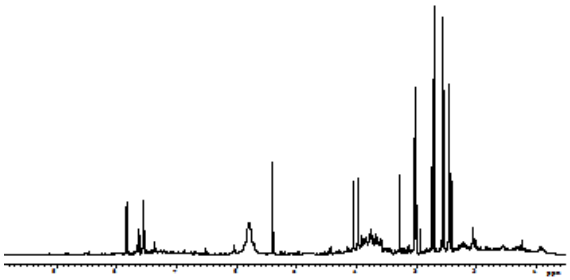
What we learn from an experiment may depend on where we look, how we look, and the scope of our view!

`http://www.noogenesis.com/pineapple/blind_men_elephant.html`

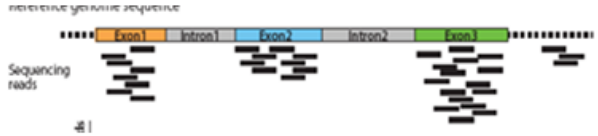
Focussing only on one platform risks missingan obvious signal



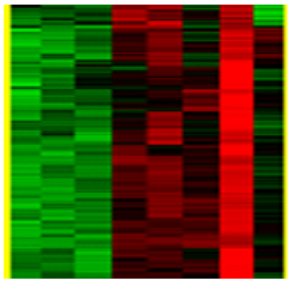
So let's measure as many as possible



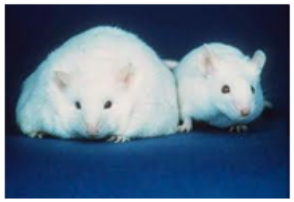
H NMR metabolites

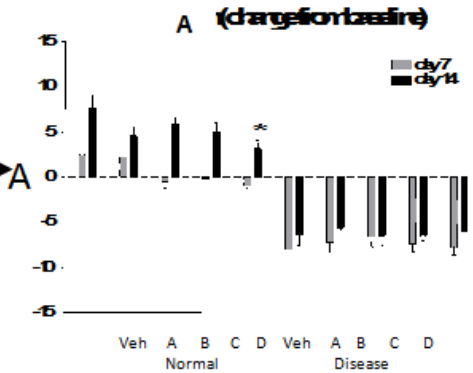


"NGS-Sequences

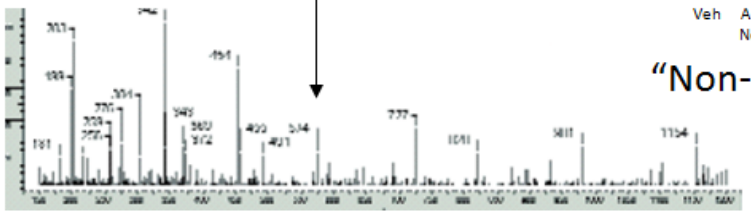


Affy Transcriptome



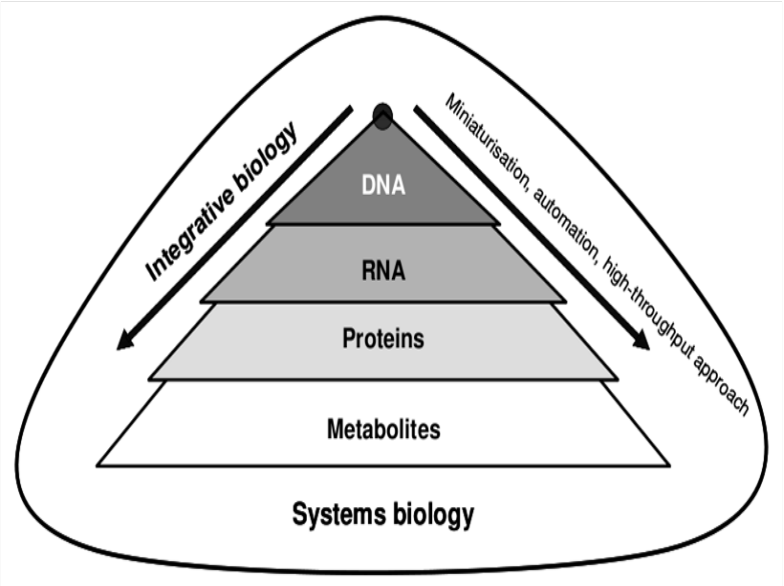


"Non-omic" markers



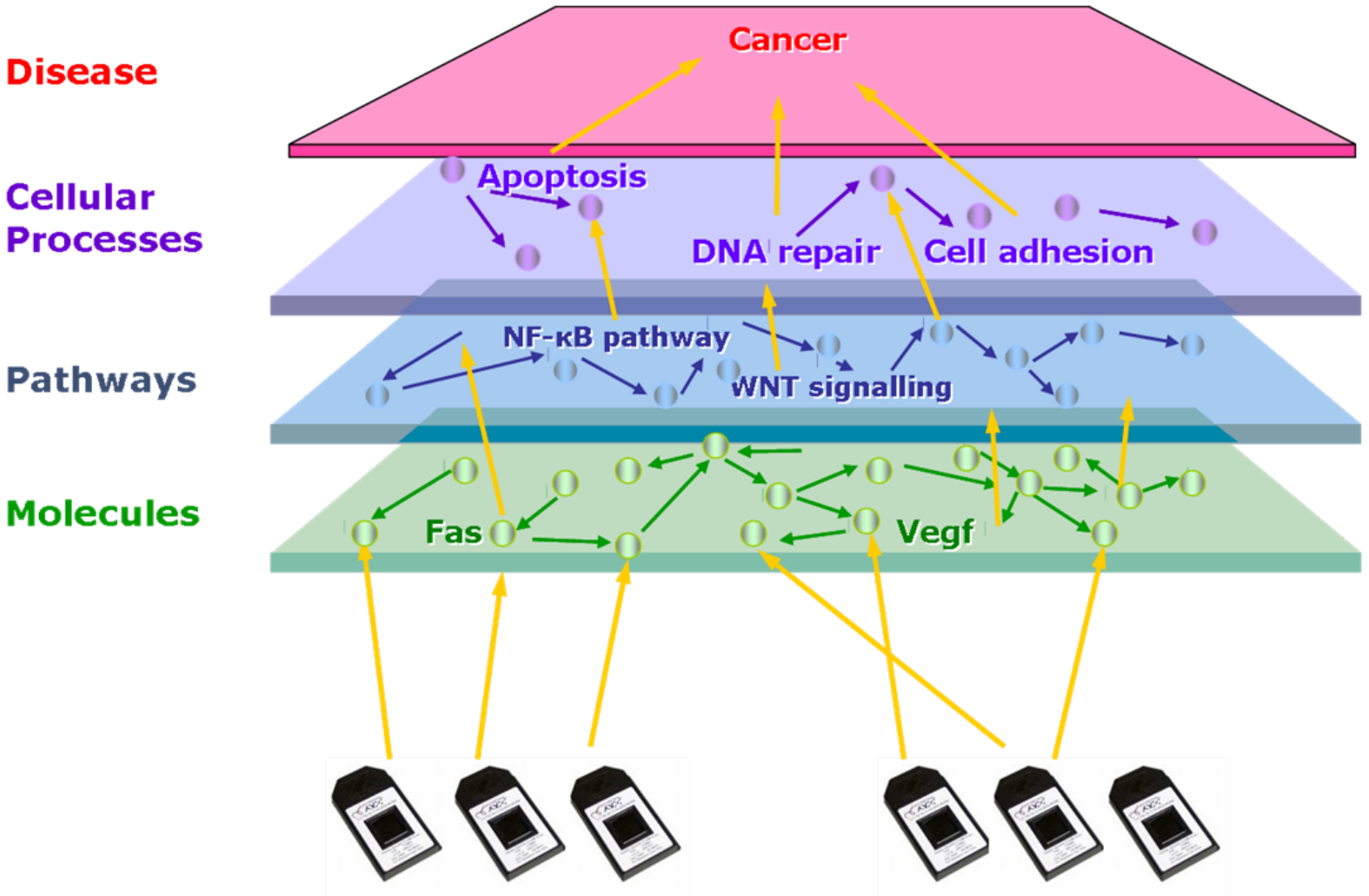
LC-MS proteomicss

From componentwise to global approaches

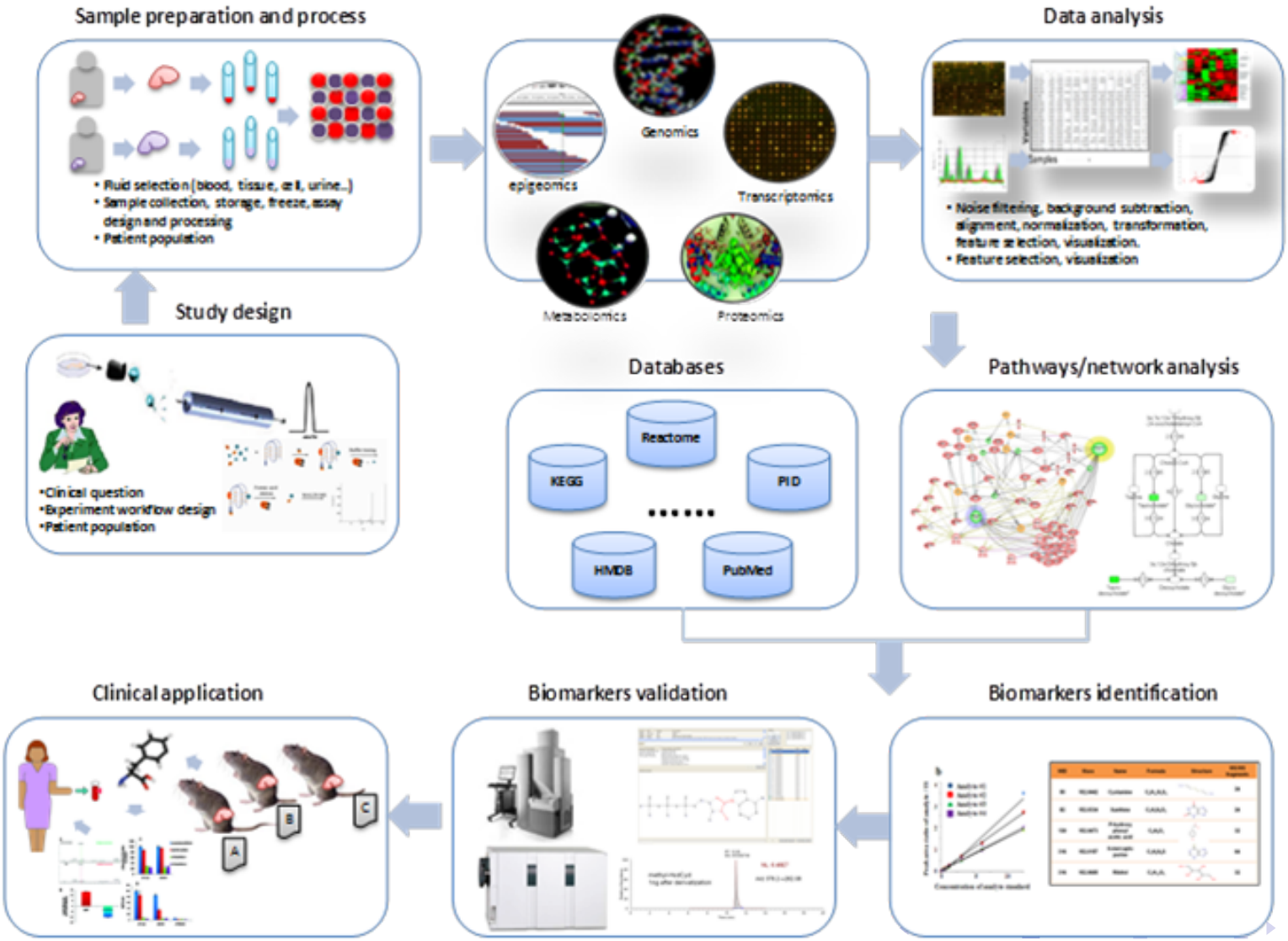


- It is expected that the integrated collection and analysis of diverse types of data,
- jointly modelled and analyzed in a systems biology approach
- can shed light on the global functioning of biological systems.

Ultimate Goal (1): understanding of complex processes

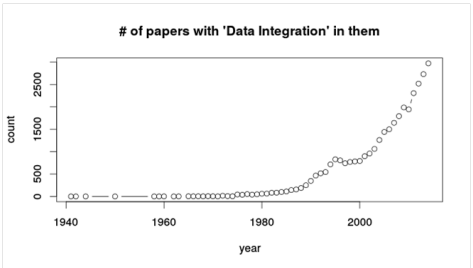


Ultimate Goal (2): Improved Robust Biomarkers



Data Integration is cool

- Everywhere nowadays in Biology, Medicine, Bioinformatics, [BTW much less in Statistics]
  - Meetings: Barcelona (Feb. 2013), Leiden (Apr. 2013), Ascona (May 2013), Crete (Nov 2014) and many more
  - Financing (FP7): projects with $> 10^6$ each (Stategra, MimOmics).
  - Increasing use of 'data integration' in publication titles.



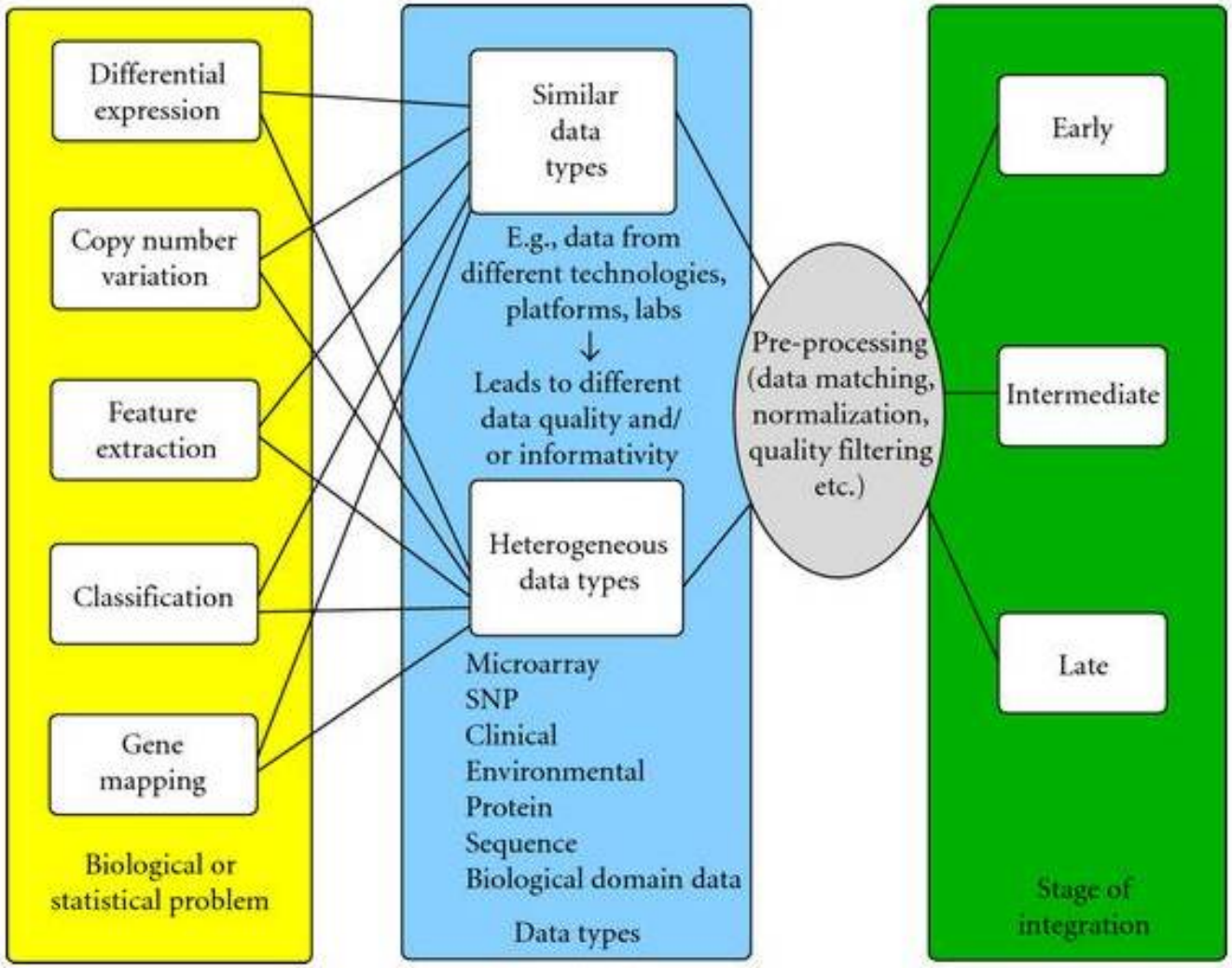
But what is Data Integration?

- "*Data integration*" may mean different things...
  - Computational combination of data
  - Combination of studies performed independently
  - Simultaneous analysis of multiple variables on multiple datasets.
  - Not to mention any possible approach for homogeneously querying heterogeneous data sources
- **Integrative analysis** may be preferable

Integrative Omics Data Analysis

- The idea that efficient integration of data from different OMICS can greatly facilitate the discovery of true causes and states of disease is rapidly pervading the biomedical community
- The aims of integrative analysis is the deciphering of complex biological relationships empowered by the combined use of distinct pieces of information that represent a, probably partial, view of the different levels at which these processes happen

There are many types of integrative analysis



There are many methods ...

BMC Bioinformatics

BioMed Central

Research article

Open Access

A structured overview of simultaneous component based data integration

Katrijn Van Deun*[1], Age K Smilde[2], Mariët J van der Werf[3], Henk AL Kiers[4] and Iven Van Mechelen[1]

Vol. 26 ISMB 20
doi:10.1093/bi

Multivariate multi-way analysis of multi-source data

Ilkka Huopaniemi[1,*], Tommi Suvitaival[1], Janne Nikkilä[1,2], Matej Orešič[3] and Samuel Kaski[1,*]

Advances in Integrative Causal Analysis

Presenter: Ioannis Tsamardinos, Ph.D., Associate Professor and Sofia Triantafillou, Ph.D. candidate.

Low level data fusion methods searching for common and distinctive biological information in hetero omics data-sets

Frans M. van der Kloet[1], Patricia Sebastián-León[2], Ana Conesa[2], Age K. Smilde[1] and Johan A. Westerhuis[1]

A myriad of tools (1): R-based

Exploration and
Integration of
Omics datasets

& Made4
& Omicade4

pwOmics

A myriad of tools (2): Others

Cytoscape

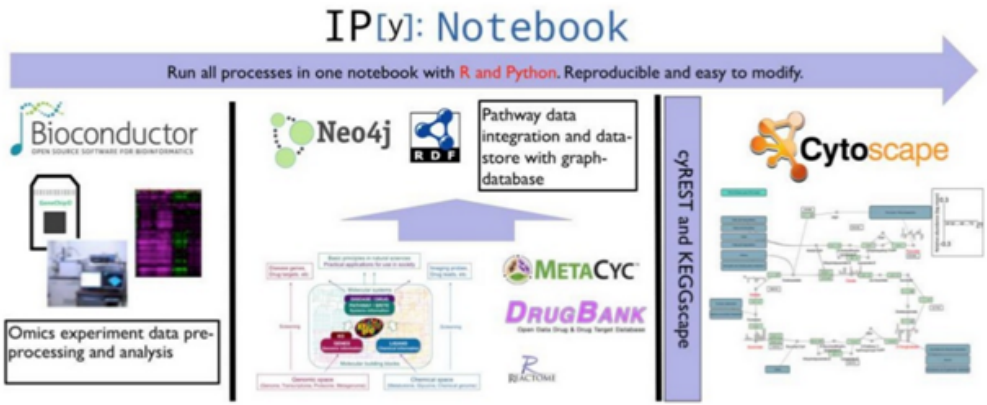
InterMine

gitools

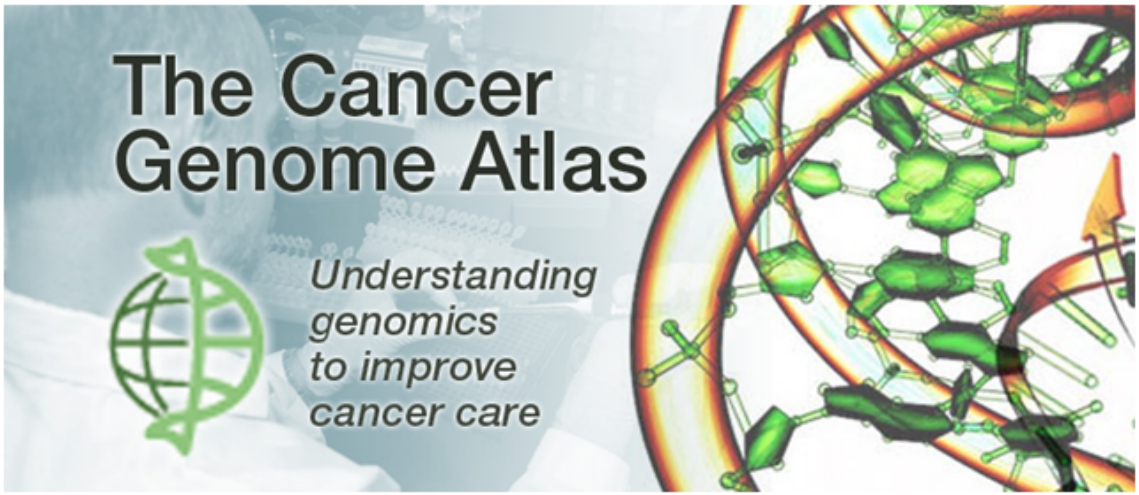
3Omics

Ondex
Data integration and visualisation

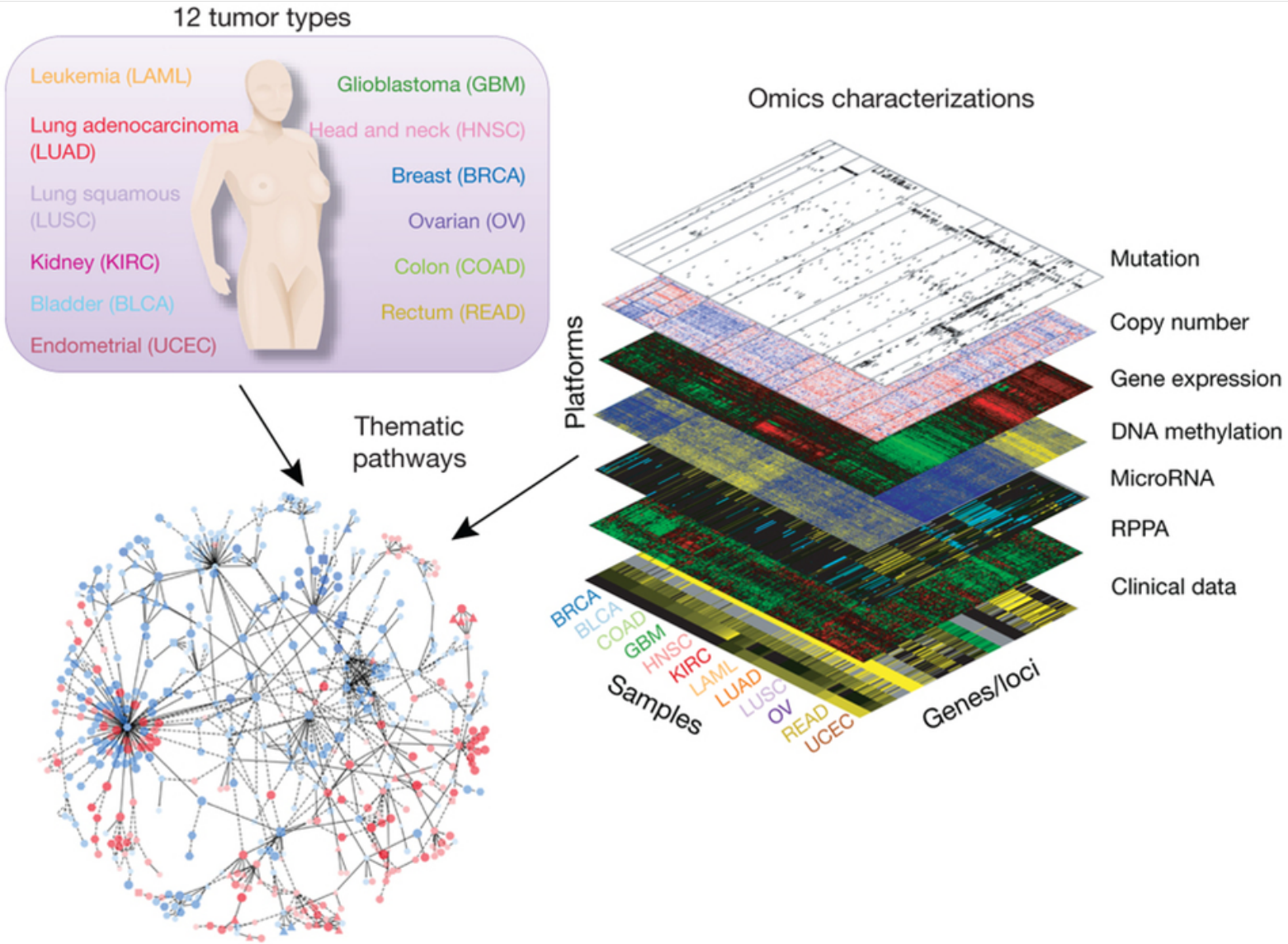
Painting omics data in biological pathways
Version 2.0



Multidisciplinary projects



Data sources: TCGA



So what?

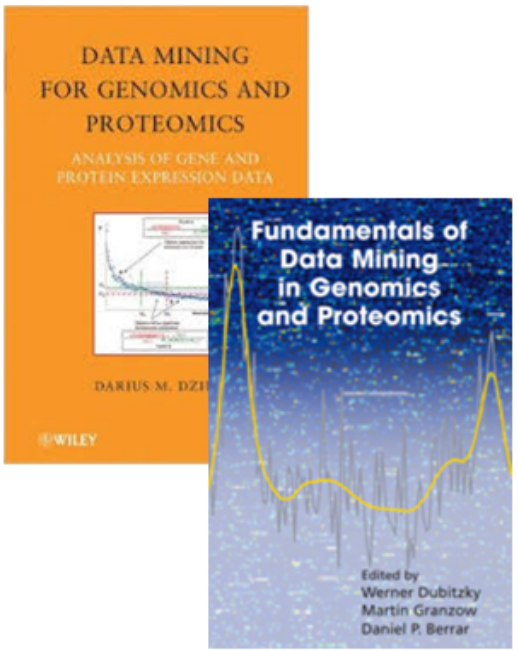
- We willl restrict to arbitrarily chosen situation
  - Multivariate statistical methods, classic and extensions
- for which we will sketch basic ideas,
  - General concept
  - Basic formulation
- and provide some examples of use
  - Applications and tools

Methods for Integrative Omics Data Analysis

Multivariate statistics in genomics

- Multivariate methods have pervaded the field of genomics since its very beginning
  - The (in)fammous clustering (HC, heatmaps)
  - Matriz factorizations / Dimension reduction (PCA, SVD, CoA)
  - Discriminant Analyses (LDA $->$ DLDA, ...)

To cite but a few.



Best approach for omics data analysis?

- Classical Statistics
  - Multiple regression
  - Discriminant analysis
  - ANOVA
- Data tables are long and lean

- Assumptions
  - Independent variables
  - More observations than variables
  - Multivariate normality
  - Interested in one dependent
  - Few missings
- DO NOT hold for many omics data

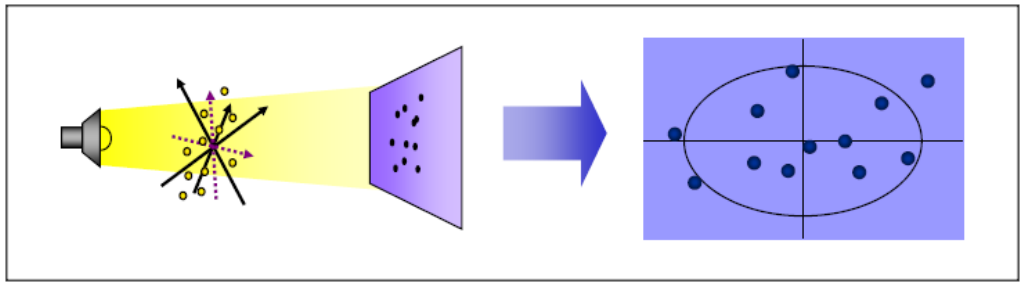
The nature of omics data

- Omics data are diverse
  - They measure distinct characteristics
    - GC/MS spectrum, Expression, Concentration
- Although they have aspects in common
  - Most of them are high throughpu
    - Many variables (K) measured simultaneously
  - Relatively expensive, ethical limitations, regulations
    - Few samples (N) analyzed

Figure: $K >> N$

A Better Way

- Multivariate analysis by projection methods
  - Looks at ALL the variables together
  - Avoids loss of information
  - Finds underlying trends = "latent variables"
  - More stable models



A not-so random walk by classical MVA

- "Classical multivariate statistical methods" (whatever this means) have been thoroughly used in the analysis of omics data.
- We review some methods and applications
  - Principal components analysis (PCA)
  - Singular value decomposition (SVD)
  - Correspondence analysis (CoA)
  - Factor analysis (FA)

Principal Components Analysis

- PCA is an statistical method that can be traced back to Pearson (1901) or Hotelling (1933).
- Its main goals are
  - Capture the information provided by the original variables using a smaller number of new variables
  - Provide a representation of the data in reduced dimension

What does PCA do?

- Given a $k \times n$ data matrix containing k (probably correlated) measurements on n samples (objects/individuals...), PCA decomposes this matrix in new k components that they ...
  - account for different sources of variability in the data,
  - are uncorrelated, that is each component accounts for a different source of variability,
  - have decreasing explanatory ability: each component explains more than the following
  - allow for a lower dimensional representation of the data in terms of scores on principal components.
  - get an overview of the dominant patterns and major trends in the data (visualize clusters, identify outliers)

Mathematical formulation

We search for linear combinations of the original variables

$$Z_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$

$$Z_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$$

$$...$$

$$Z_p = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p$$

Subject to:

- $Z_1, Z_2, ..., Z_p$ uncorrelated
- $Var(Z_1)$ maximal
- $Var(Z_1) \geq Var(Z_2) \geq ...Var(Z_p)$.

Mathematical formulation

- All coeficients and eigenvalues can be obtained by the spectral decomposition of the covariance matrix

$$\Sigma = AD_\lambda A'$$

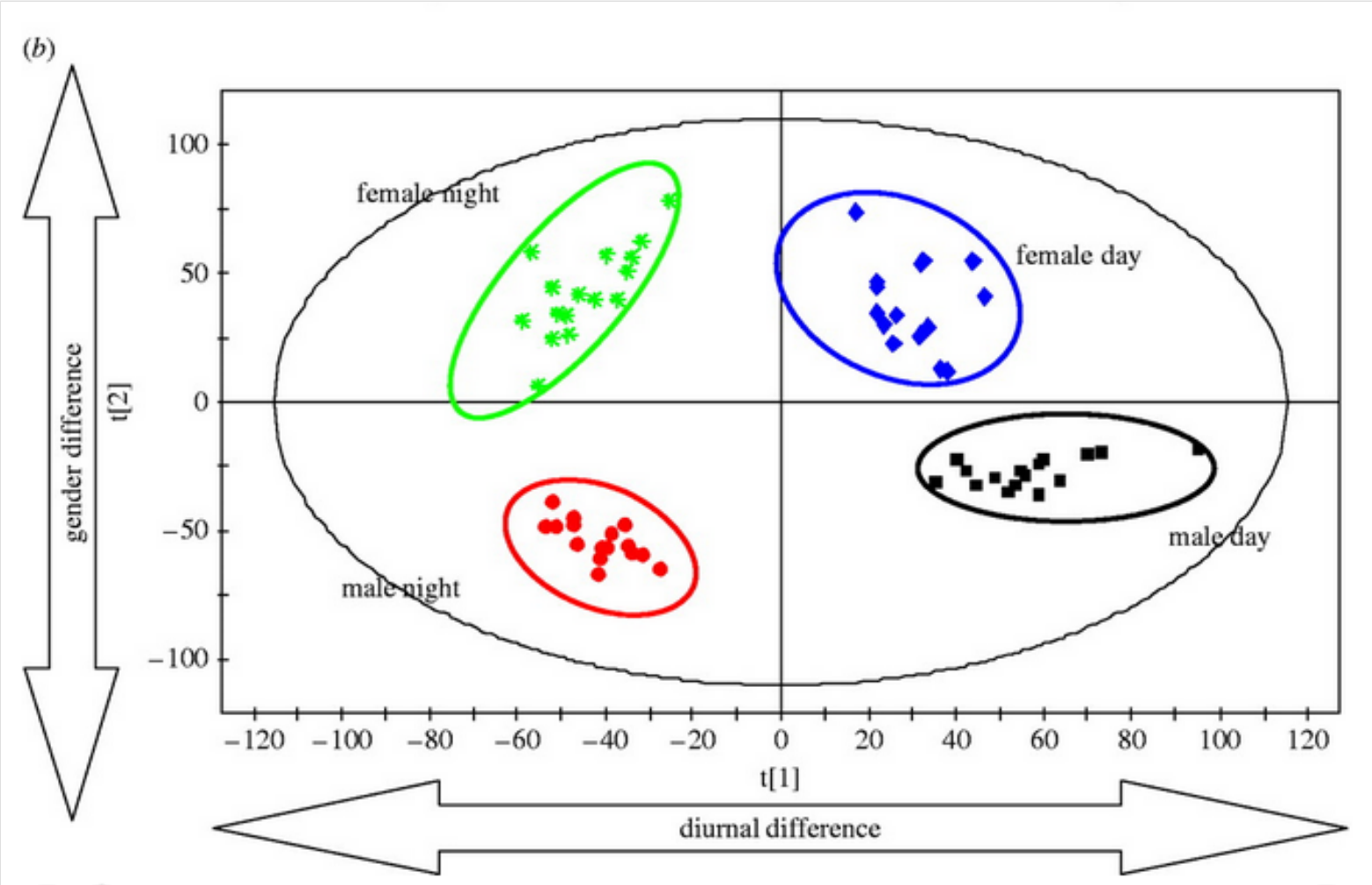
- We will use the sample covariance matrix S to estimate Σ

$$S = AD_\lambda A'$$

- and estimate the components by sample components

$$\underset{(nxp)}{Z} = \underset{(nxp)}{X} \underset{(pxp)}{A}$$

Data visualization in reduced dimension



Applications of PCA

- PCA has been used in a wealth of applications
- The best known
  - Batch effect detection
  - Pattern recognition
- Besides standard applications there exist many extensions
  - Probabilistic PCA, Bayesian PCA, Inverse non-linear PCA, Nipals PCA, Robust PCA

Examples

Science
AAAS.ORG | FEEDBACK | HELP | LIBRARIANS
All Science Journals
Enter Search Term
CRAI. UNIVERSITAT DE BARCELONA ALERTS | ACCESS RIGHT
AAAS NEWS SCIENCE JOURNALS CAREERS MULTIMEDIA COLLECTIONS
Science The World's Leading Journal of Original Scientific Research, Global News, and Commentary.
Science Home | Current Issue | Previous Issues | Science Express | Science Products | My Science | About the Journal
Home > Science Magazine > 23 October 1998 > Chu et al., 282 (5389): 699-705
Article Views
Abstract
Full Text
Full Text (PDF)
A correction has been published
Science 23 October 1998:
Vol. 282 no. 5389 pp. 699-705
DOI: 10.1126/science.282.5389.699
‹ Prev | Table of Contents | Next ›
RESEARCH ARTICLE
The Transcriptional Program of Sporulation in Budding Yeast
S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, I. Herskowitz

Pac Symp Biocomput. Author manuscript; available in PMC 2009 Apr 17.
Published in final edited form as:
Pac Symp Biocomput. 2000 : 455–466.
PMCID: PMC2669932
NIHMSID: NIHMS97353
Author Manus
PRINCIPAL COMPONENTS ANALYSIS TO SUMMARIZE MICROARRAY EXPERIMENTS: APPLICATION TO SPORULATION TIME SERIES
Soumya Raychaudhuri, Joshua M. Stuart, and Russ B. Altman
Author information ► Copyright and License information ►

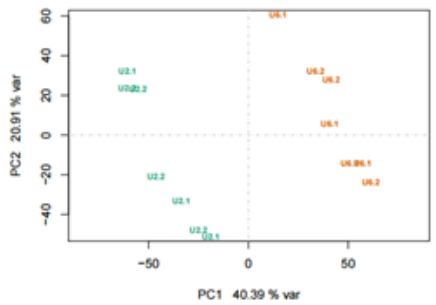
Examples

Journal of Proteomics
Volume 75, Issue 13, 16 July 2012, Pages 3938–3951

Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics

Josep Gregori[a, b], Laura Villarreal[a], Olga Méndez[a], Alex Sánchez[b, c], José Baselga[a], Josep Villanueva[a,

Show more



R Packages alpha Home Blog All packages Search R packages Searc

Exploratory Data Analysis of LC-MS/MS data by spectral counts

Exploratory data analysis to assess the quality of a set of LC-MS/MS experiments, and visualize de influence of the involved factors.

Author	Josep Gregori, Alex Sanchez, and Josep Villanueva
Date of publication	None
Maintainer	Josep Gregori <josep.gregori@gmail.com>
License	GPL-2
Version	1.2.0

View on Bioconductor

Singular value Decomposition

- When Things Get Messy...
  - PCA is fine when initial dimension is not too big
    - Space and time complexity are of $O(d^2)-$ size of covariance matrix
  - Otherwise – we have a problem...
    - E.g. when $d = 10^4 =$ time/space complexity is $O(10^8)$...
  - An alternative approach: SVD

Mathematical formulation

Any real $n \times p$ matrix X can be decomposed as

$$X = UDV'$$

- U $n \times r$ matrix of orthonormal left singular vectors. $U'U = I_r$
- D $r \times r$ diagonal matrix of non−increasing positive singular values $(d_{11} \geq d_{22} \geq ... \geq d_{rr})$
- V $p \times r$ matrix of orthonormal right singular vectors $V'V = I_r$

A rank k approximation $\hat{X}$ to matrix X optimal in the least squares sense, is obtained as

$$\hat{X} = U_{[,1:k]} D_{[1:k,1:k]} V'_{[,1:k]}$$

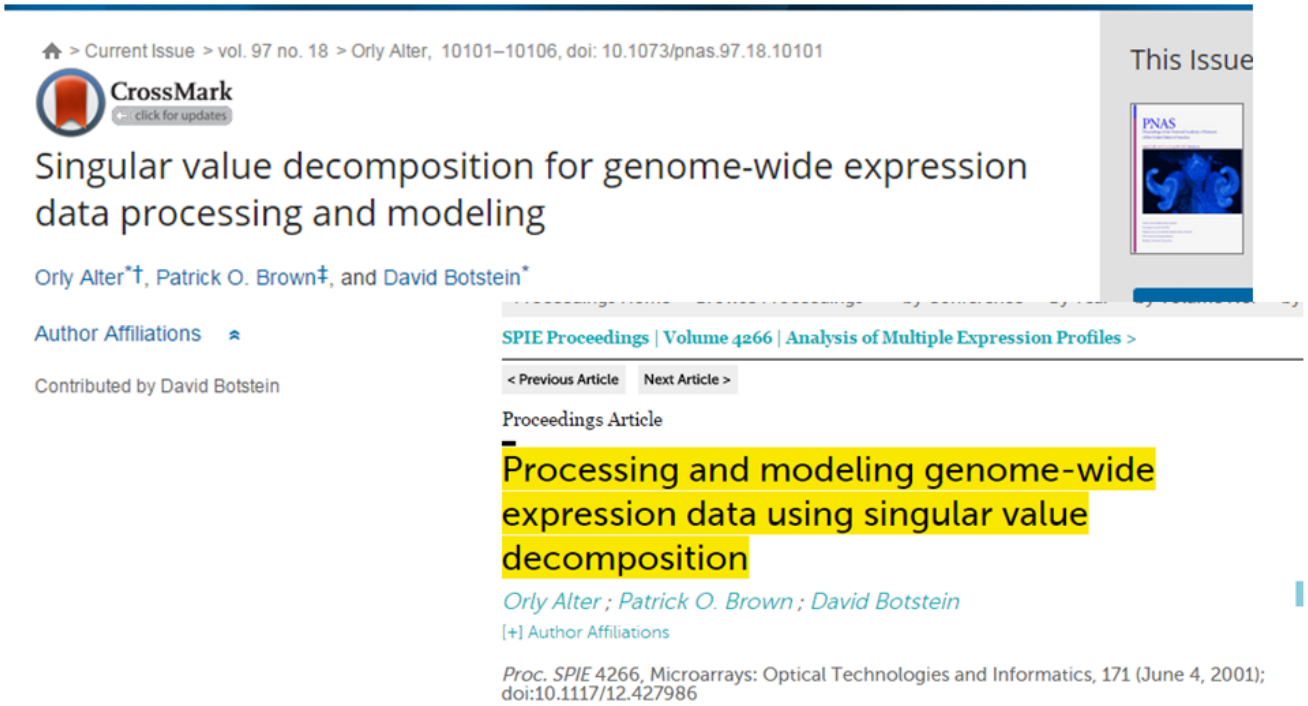
E.g., a rank 2 approximation to matrix X is obtaind by

$$U_{nx2} D_{(2x2)} V'_{px2}$$

PCA and SVD

- Informally speaking
  - PCA is a statistical method that can be solved using different approaches, one of which is SVD.
  - SVD is an algebraic method to decompose a rectangular matrix. Used in signal processing to yield approximation.

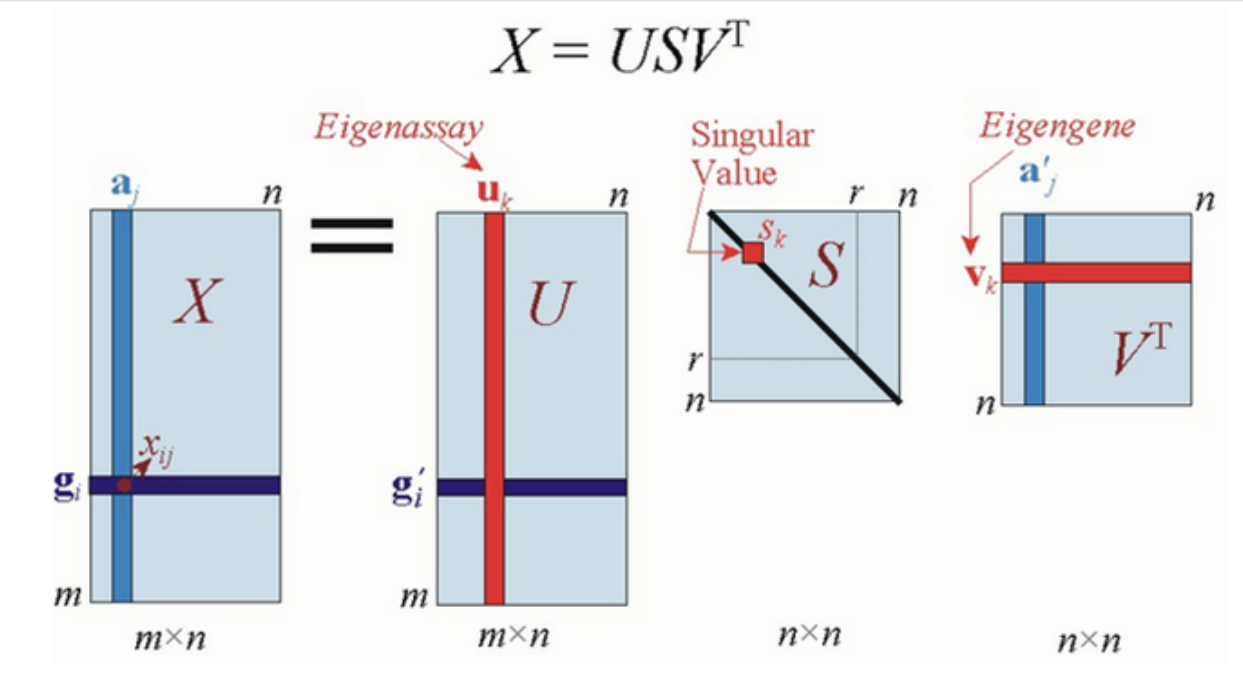
Case studies



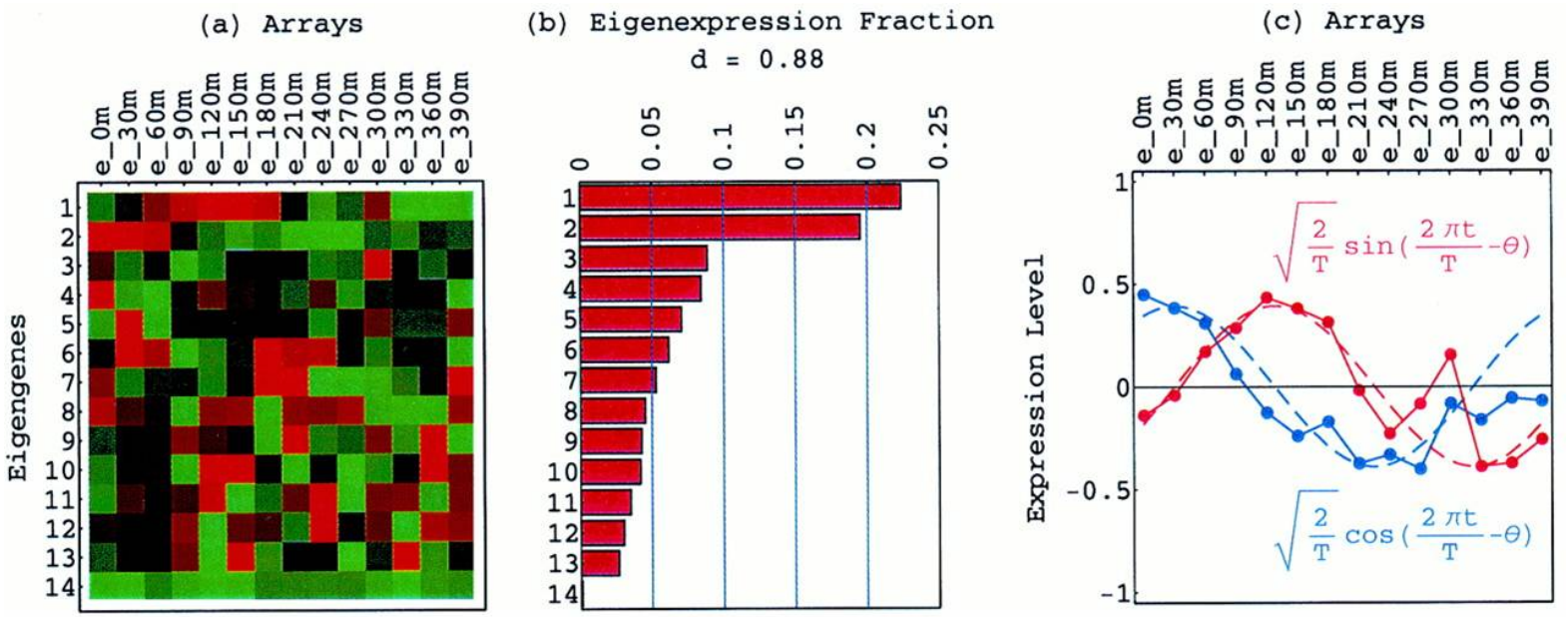
http://www.alterlab.org/publications/

Eigengenes, Eigenarrays and SVD

- Alter et alt. use the singular value decomposition (SVD) for transforming the dataset from the gene/array space to the eigengene/eigenarray space
- Each dimension is represented by an eigengene/eigenarray/eigenvalue triplet



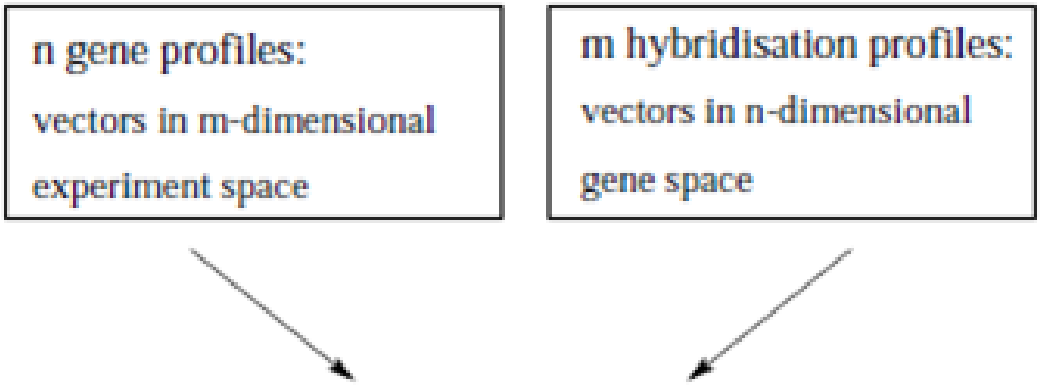
Example results



Correspondence Analysis

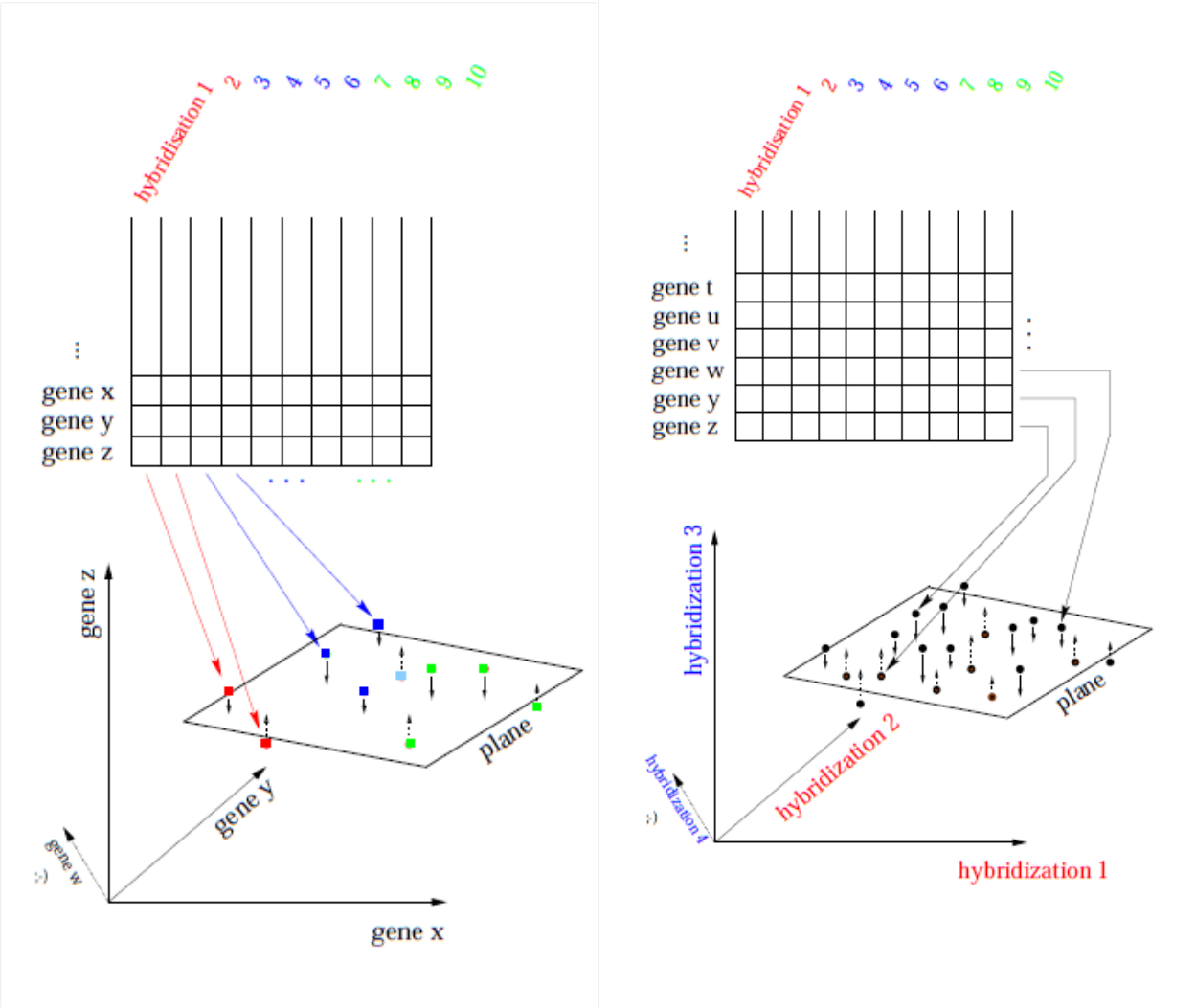
- Yet another dimension reduction technique?
- CA is a method for studying associations between variables
  - Like PCA displays low-dimensional projection
  - It is done for two variables simultaneously so it may help detect associations between them.
- CA is applied to reveal associations between genes and experiments

Brief methodology



- projection into a common subspace of low dimensionality for visualisation
- conserving point to point distances (total variance) as well as possible (-> explained variance)

Visualization



Genes and arrays are projected **on the same plane**

Mathematical formulation (1)

- Original data N (I x J matrix) with elements n_{ij}.

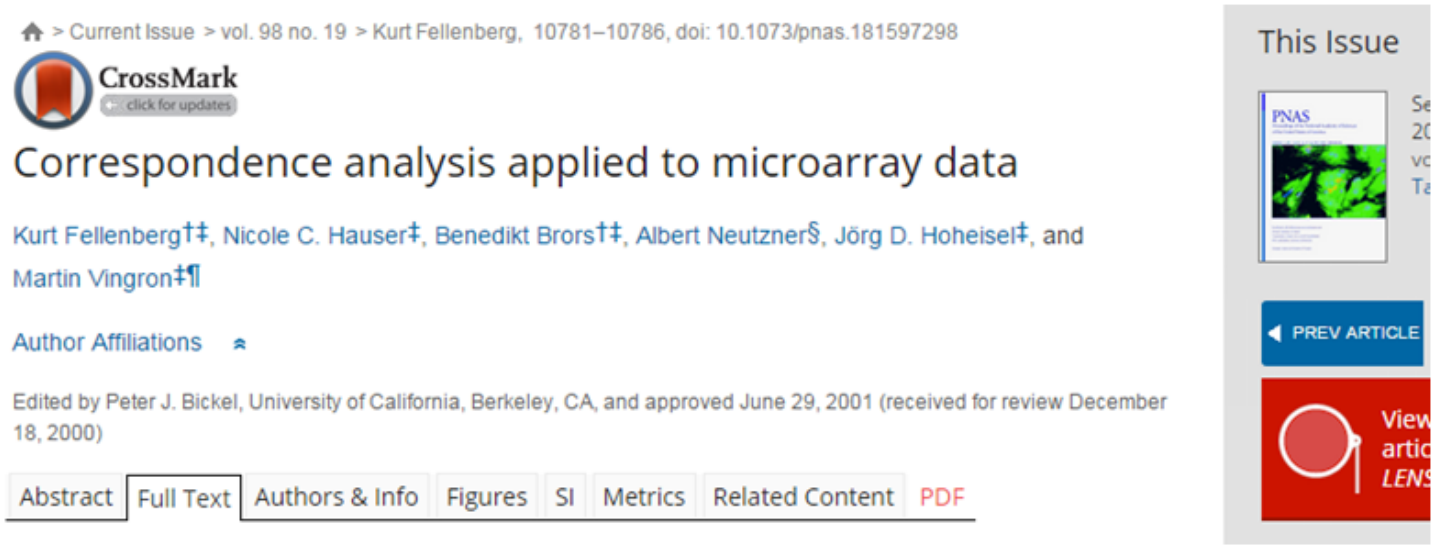
$$n_{+j} = \sum_{i=1}^{I} n_{ij}, \quad n_{i+} = \sum_{j=1}^{J} n_{ij}, \quad n_{++} = \sum i = 1 \sum j = 1^{J} n_{ij}$$

- row weights $r_i = n_{i+}/n_{++}$
- column weights: $c_j = n_{+j}/n_{++}$
- correspondence matrix P : $p_{ij} = n_{ij}/n_{++}$
- X^2 matrix S : $s_{ij} = \frac{(p_{ij} - r_i c_j)}{\sqrt{r_i c_j}}$
- singular value decomposition: $S = U \wedge V^T$, diag matrix $\wedge$ contains singular values λ_k of S

- principal coordinates:
  - for gene i (principle axis $k-1,...,J$) $f_{ik} = \frac{\lambda_k u_{jk}}{\sqrt{r_i}}$
  - for hybridazation j (in the same space) $g_{jk} = \frac{\lambda_k u_j k}{\sqrt{r_j}}$
- standard coordinates:
  - for gene i $f_{jk} = u_{ik}/\sqrt{r_i}$
  - for hybridization j $g_{jk} = v_{jk}/\sqrt{c_j}$
- HMS
  - Let N contain the hybridization medians of the original data matrix $N*$ if elements $n*_{ij'}$
  - N is submited to CA
  - Let P have elements $p*_{ij'} = n_{ij}/n^*_{++}$ then the principle coordinates for the supplementary hybridization from correspondence matrix P^* are

$$g^*_{j'k} = \frac{1}{\sum_i p^*_{ij'}} \sum_i \frac{p^*_{ij'} f_{ik}}{\lambda_k}$$

Case studies



> Current Issue > vol. 98 no. 19 > Kurt Fellenberg, 10781–10786, doi: 10.1073/pnas.181597298

CrossMark click for updates

Correspondence analysis applied to microarray data

Kurt Fellenberg†‡, Nicole C. Hauser‡, Benedikt Brors†‡, Albert Neutzner§, Jörg D. Hoheisel‡, and Martin Vingron‡¶

Author Affiliations

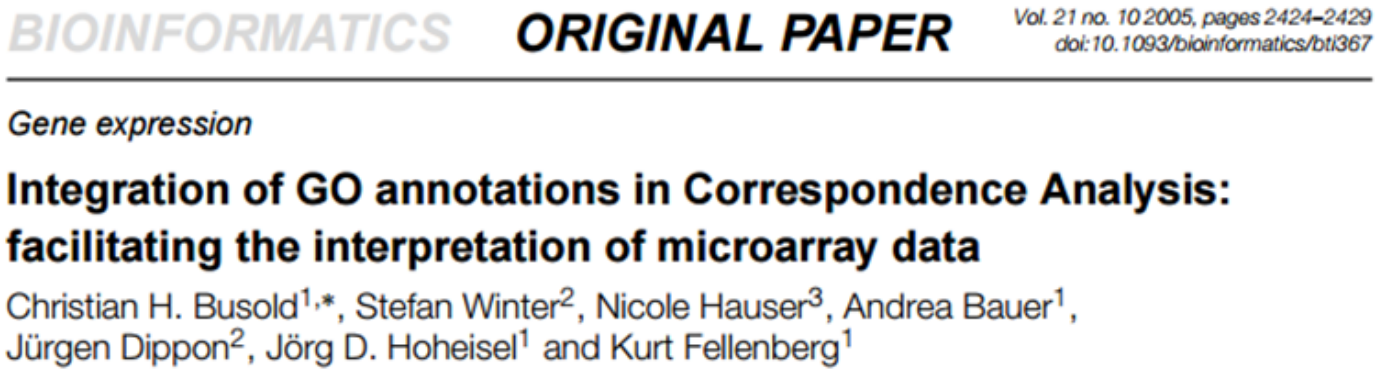
Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved June 29, 2001 (received for review December 18, 2000)

Abstract | Full Text | Authors & Info | Figures | SI | Metrics | Related Content | PDF

This Issue

PREV ARTICLE

View artic LENS



BIOINFORMATICS ORIGINAL PAPER

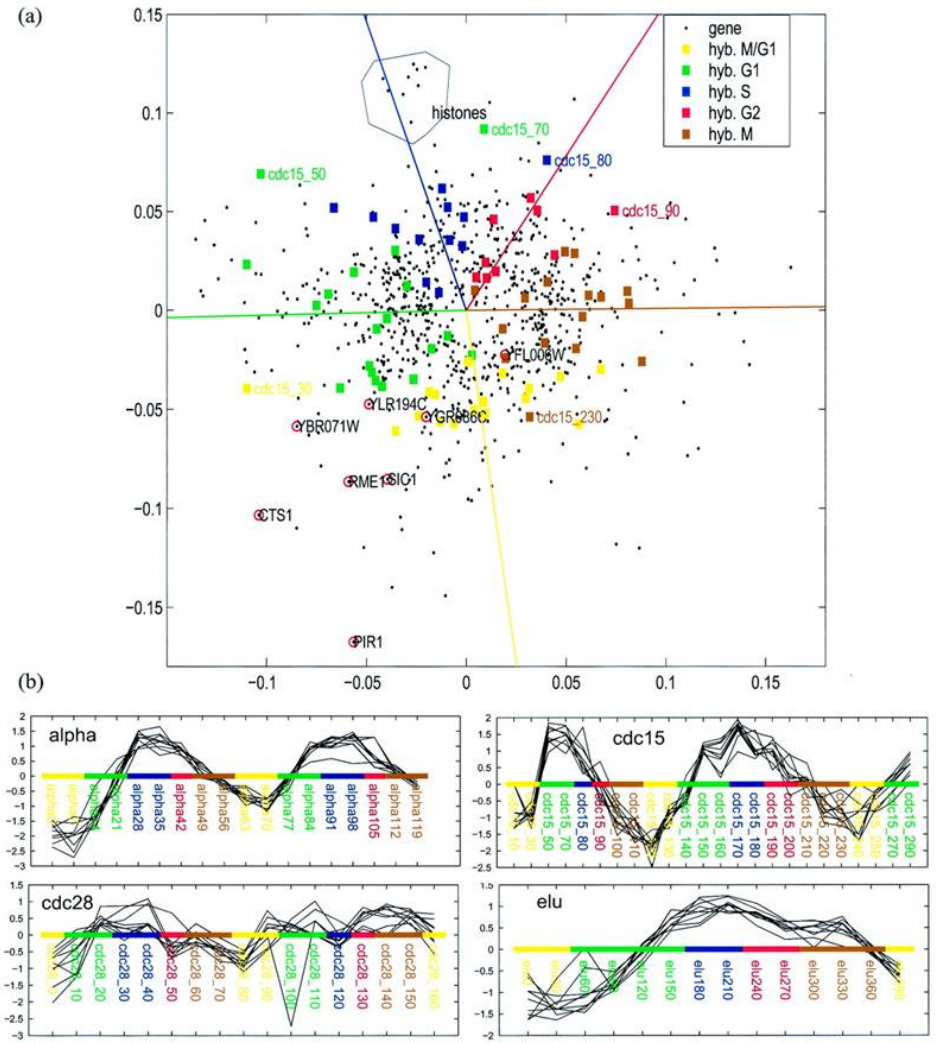
Vol. 21 no. 10 2005, pages 2424–2429
doi:10.1093/bioinformatics/bti367

Gene expression

Integration of GO annotations in Correspondence Analysis: facilitating the interpretation of microarray data

Christian H. Busold[1,*], Stefan Winter[2], Nicole Hauser[3], Andrea Bauer[1], Jürgen Dippon[2], Jörg D. Hoheisel[1] and Kurt Fellenberg[1]

CA Analysis of Cell-cycle synchronization data



Integration of GO annotations with CA

- CA can be enhanced by adding supplementary variables.
- These appear at indicative positions without contributing to scores computation.
- As an application of this feature, Gene Ontology or other annotations can be added so that *genes, experimental conditions and gene-annotations are viewed in a single plot*.

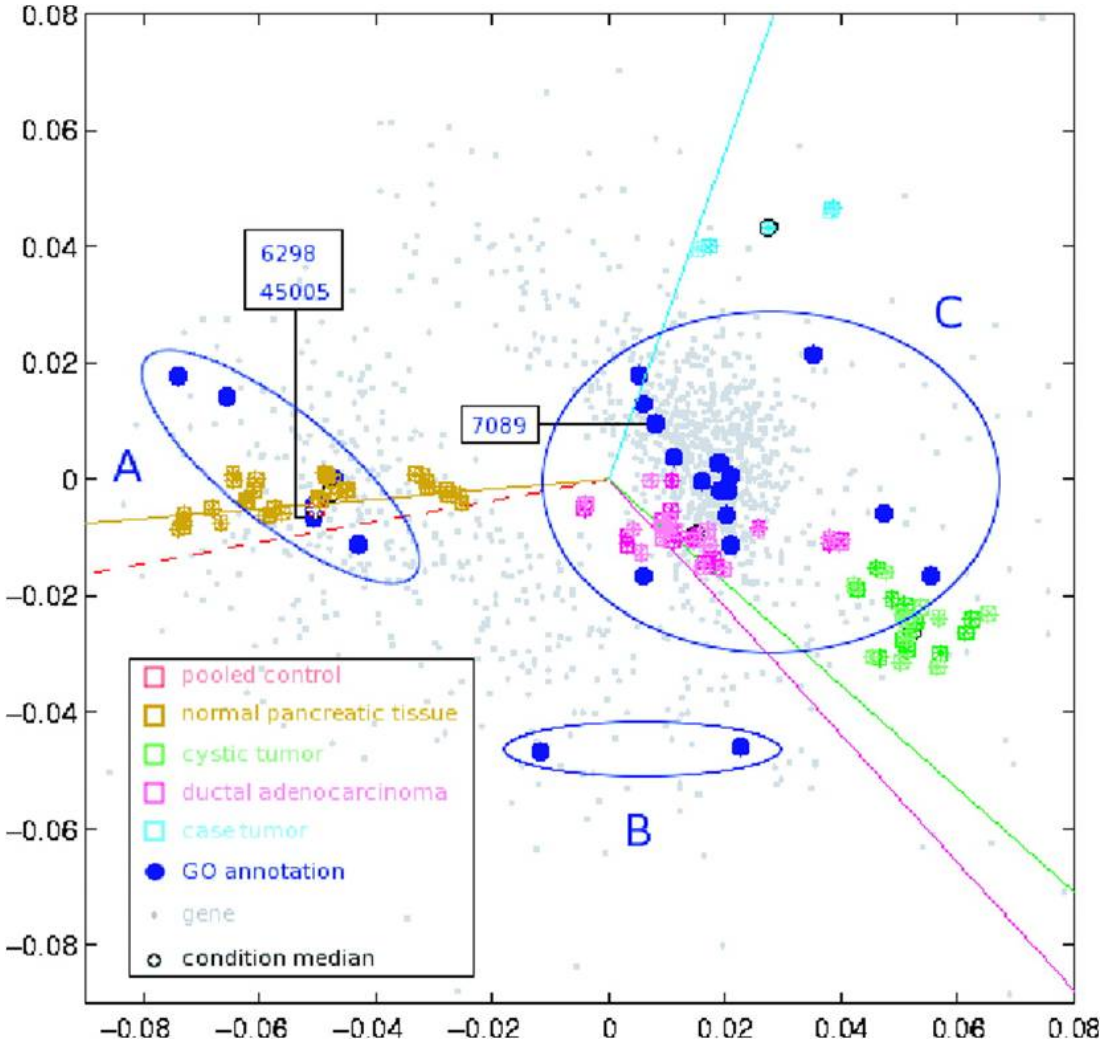
Adding annotations to an expression matrix

- GO Terms can be associated with genes or samples
  - Using a *Boolean implementation* as supplementary columns,
  - Using an *Intensity implementation* as supplementary rows,
- A filtering to select the most descriptive annotations can be applied to keep only annotations highly correlated with common transcription profiles.

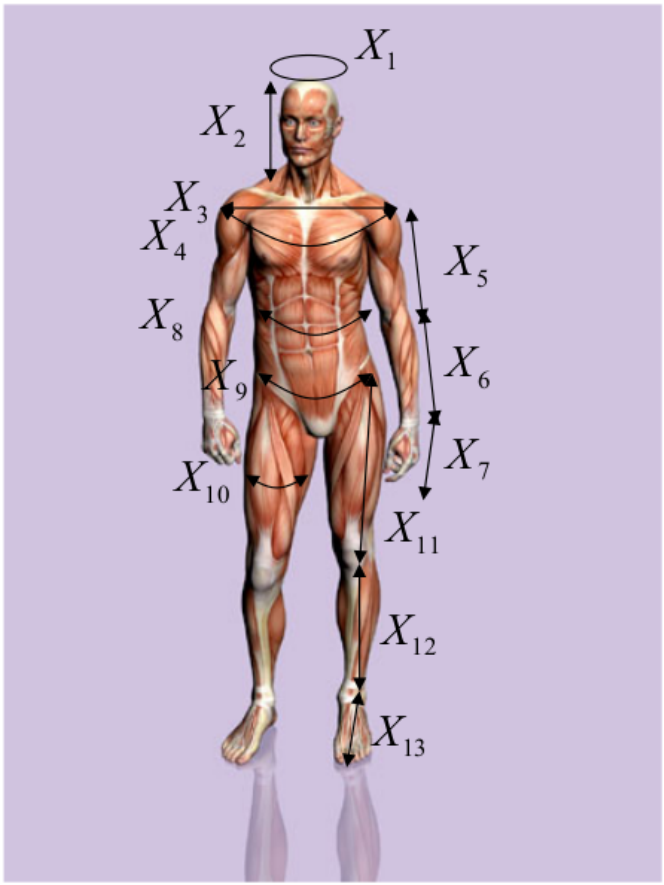
Adding annotations to an expression matrix

Gene	Exp. cond.1	Exp. cond. 2	Exp. cond. 3	Exp. cond. 4	Term 1	Term 2	.
A	13	300	23	432	1	0	.
B	457	398	355	932	0	1	.
C	24	458	44	364	1	1	.
D	324	245	98	34	0	0	.
E	478	928	293	99	0	1	.
F	38	485	21	375	1	0	.
.	.	.	.	.	.	.	.
Term 1	75	1243	88	1171			
Term 2	959	1784	692	1395			
.	.	.	.	.			

CA Map of Human Pancreatic Cancer Study



Factor Analysis



$$\underbrace{(X_1,\ldots,X_{13})}_{\text{Observable variables}} = f\underbrace{(gene_1, gene_2, food)}_{\text{Non-observable (latent) variables}}$$

$$\mathbf{X} = \boldsymbol{\mu}_X + \Lambda\mathbf{f} + \boldsymbol{\varepsilon}$$

Average — Load matrix — Noise

Mathematical formulation (1)

$$\mathbf{X} = \boldsymbol{\mu}_X + \Lambda \mathbf{f} + \boldsymbol{\varepsilon} \longrightarrow \begin{pmatrix} X_1 \\ X_2 \\ \dots \\ X_{13} \end{pmatrix}_{p} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_{13} \end{pmatrix} + \begin{pmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} \\ \dots & \dots & \dots \\ \lambda_{13,1} & \lambda_{13,2} & \lambda_{13,3} \end{pmatrix}_{p \times m} \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix}_{m} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{13} \end{pmatrix}_{p}$$

$N_p(\boldsymbol{\mu}_X, \Sigma_X)$ $N_m(\mathbf{0}, I)$ $N_p(\mathbf{0}, \Sigma_\varepsilon)$

- Properties:
  - Factors are uncorrelated / independent
  - Factors and noise are uncorrelated
  - The liad matrix is the covariance between the observed variables and the factors
  - The variance of the observed variables and the factors
  - The variance of the observed variables can be explained by the loading matrix and the variance of the noise

$$EFF' = I$$

$$E\epsilon F' = 0$$

$$\wedge = E(X - \mu_x)F'$$

$$\sum_x \wedge \wedge' \sum_\epsilon$$

$$\sigma^2_{x_i} = \sum_{j=1}^{m} \lambda^2_{ij} + \sigma^2_{\epsilon_i} = h^2_i + \sigma^2_{\epsilon_i}$$

Mathematical formulation (2)

$X = \mu_X + \wedge f + \epsilon$ Properties:

- The load matrix and factors are not uniquely specified: any rotation of the factors can be compensated by the load matrix

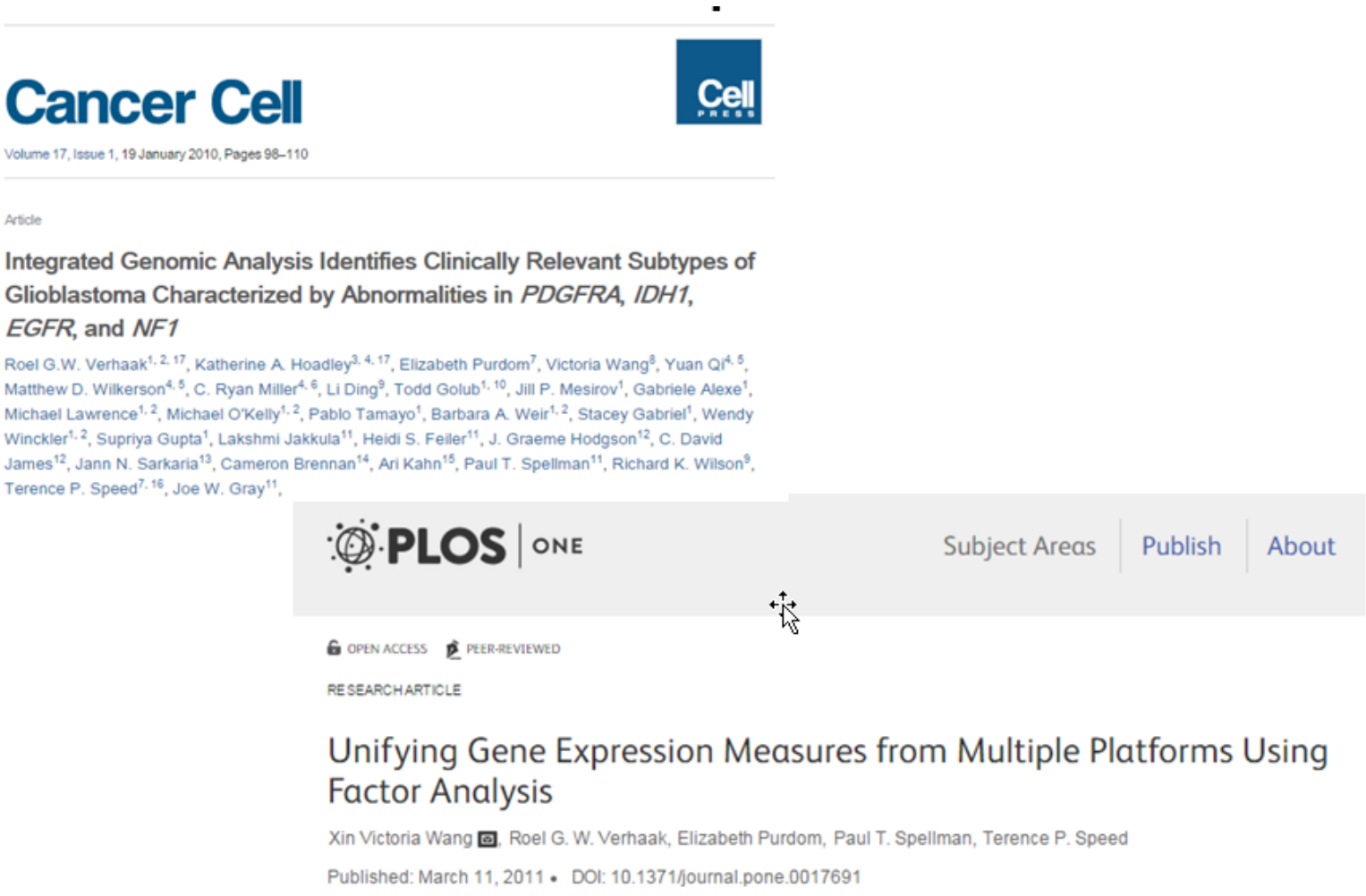
$$X = \mu_X + \wedge f + \epsilon = \mu_X + (\wedge H')(Hf) + \epsilon$$

Solution:

1. impose that $\wedge\wedge'$ is a diagonal matrix : Principal factor method
2. Impose that $\wedge' \sum_{\epsilon}^{-1} \wedge$ is a diagonal matrix: Maximum-Likelihoof method

$H' - >$ Any orthogonal matrix: This matrix provides the possibility of rotating the components so that we achieve a certain property (having the maximum number of zeros,...) that helps us to undestand the factors. There are several criteria offered by the programs: varimax, equamax, parsimax, quartimax, orthomax,...

Examples



Cancer Cell

Volume 17, Issue 1, 19 January 2010, Pages 98–110

Article

Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*

Roel G.W. Verhaak[1, 2, 17], Katherine A. Hoadley[3, 4, 17], Elizabeth Purdom[7], Victoria Wang[8], Yuan Qi[4, 5], Matthew D. Wilkerson[4, 5], C. Ryan Miller[4, 6], Li Ding[9], Todd Golub[1, 10], Jill P. Mesirov[1], Gabriele Alexe[1], Michael Lawrence[1, 2], Michael O'Kelly[1, 2], Pablo Tamayo[1], Barbara A. Weir[1, 2], Stacey Gabriel[1], Wendy Winckler[1, 2], Supriya Gupta[1], Lakshmi Jakkula[11], Heidi S. Feiler[11], J. Graeme Hodgson[12], C. David James[12], Jann N. Sarkaria[13], Cameron Brennan[14], Ari Kahn[15], Paul T. Spellman[11], Richard K. Wilson[9], Terence P. Speed[7, 16], Joe W. Gray[11],

PLOS ONE | Subject Areas | Publish | About

OPEN ACCESS PEER-REVIEWED

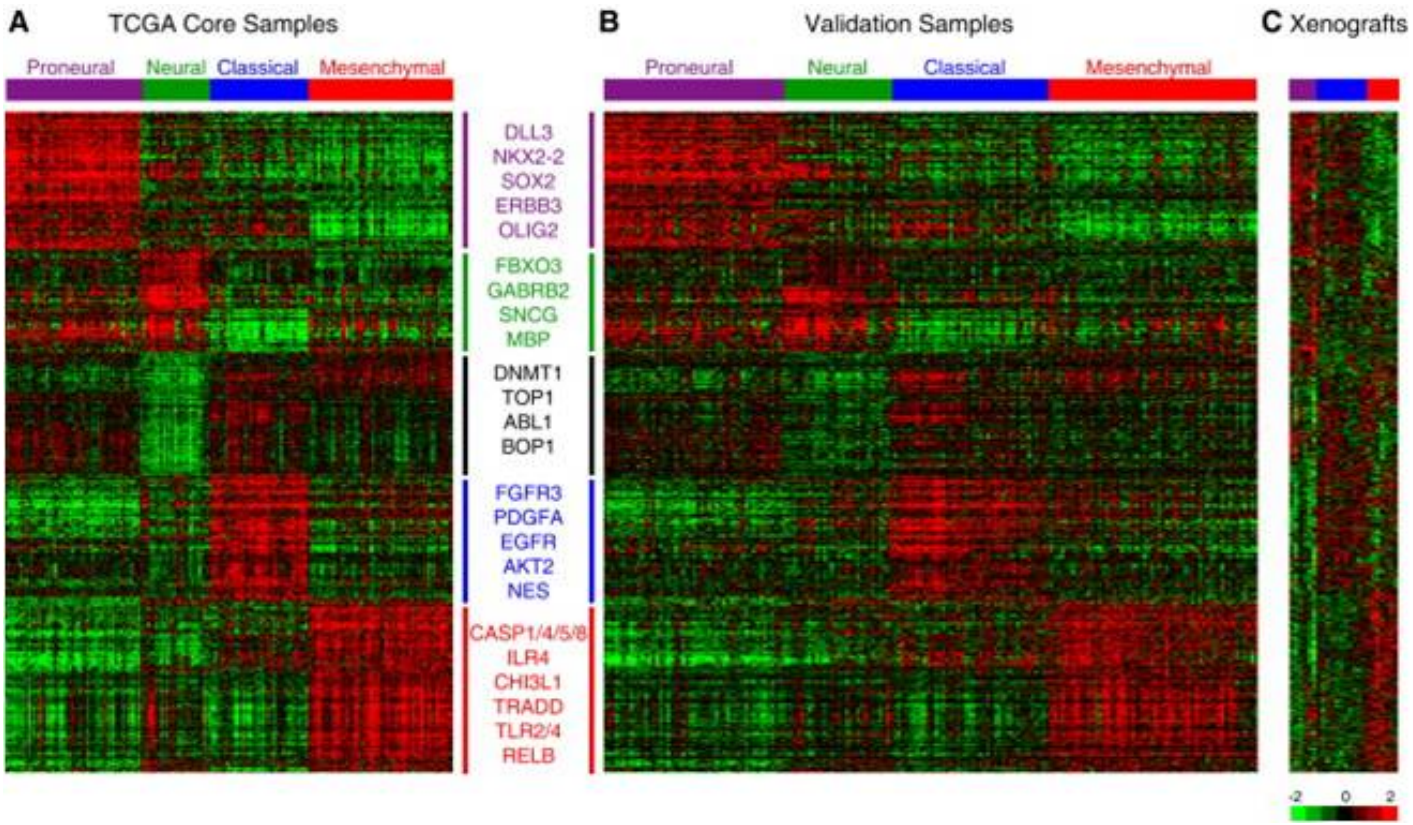
RESEARCH ARTICLE

Unifying Gene Expression Measures from Multiple Platforms Using Factor Analysis

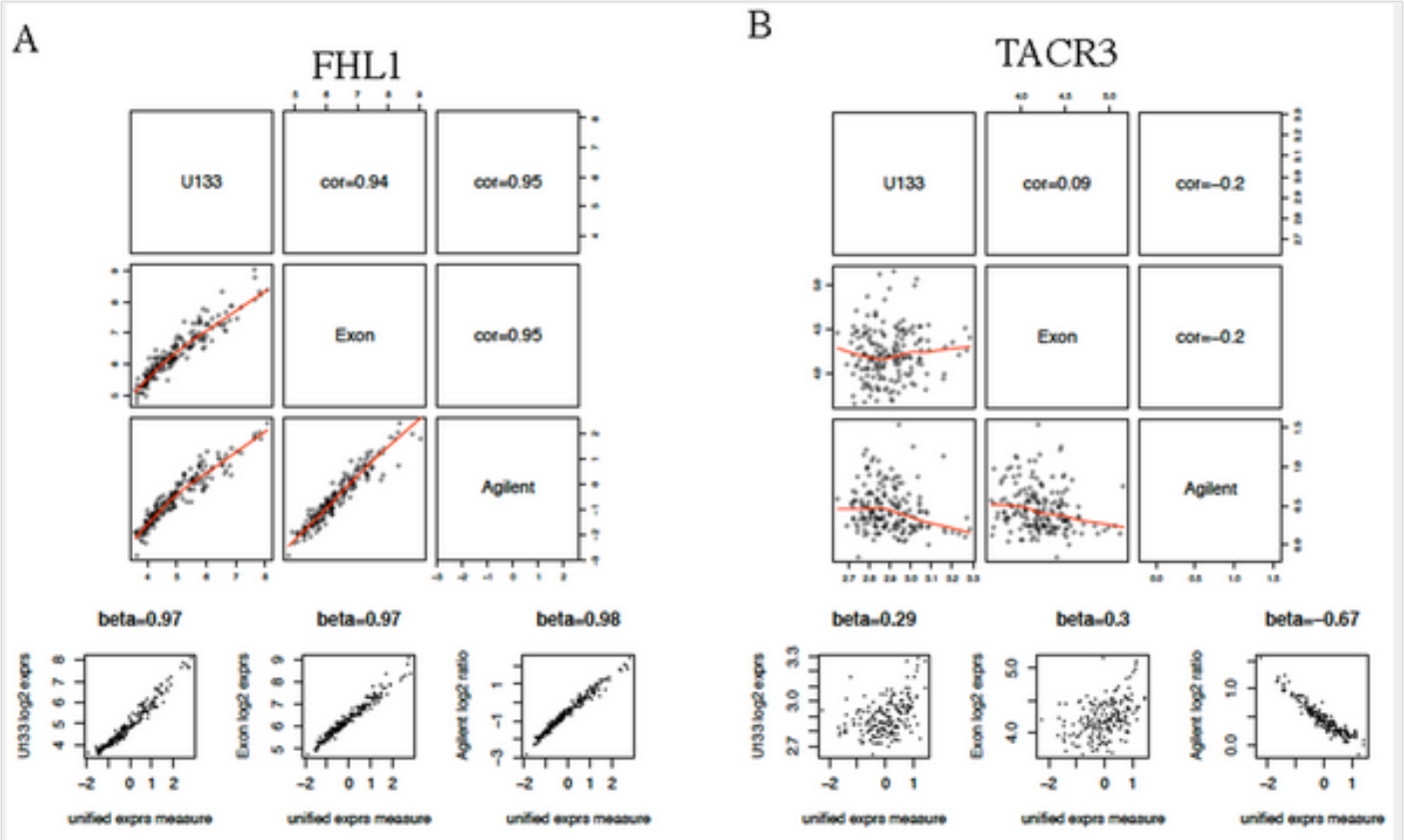
Xin Victoria Wang, Roel G. W. Verhaak, Elizabeth Purdom, Paul T. Spellman, Terence P. Speed

Published: March 11, 2011 • DOI: 10.1371/journal.pone.0017691

Gene Expression Data Identify Four Gene Expression Subtypes



Correlation between expression measures and unified FA-measure



In summary

- Applying basic ideas of multivariate statistics allows for powerful and informative analyses on omics data such as
  - Projections in reduced dimension to detect patterns on data
  - Joint visualization of genes, and biological annotations
  - Combination of values from distinct platforms into improved expression measures

However

- Increasing availability of omics technologies
- and projects such as The Cancer Genome Atlas
- generates multi-omics datasets that can/must be analyzed using many different methods
  - which extend classical approximations
  - or completely new approaches

(Modern) Multivariate methods for (omics) data integration

Multivariate methods for Omics data Integration

- Having multiple, possibly related, datasets is a common situation, and
- Many multivariate methods for simultaneousl analysis were already available when the multi-omics explossion arrived
  - Ecological data analysis, Survey analysis , Chemometrics
- The omics community has either
  - Adapted existing methods
  - Developed new ones
- Result: multi-populated list of methods that would probably benefit from a good re-organization

An arbitrary grouping

- Classical methods for analysis of two datasets
  - CCA, Colnertia Analysis, PLS
- Generalizations of classical methods
  - Generalized Singular Value Decomposition [omit'd]
  - Multiple Colnertia Analysis
- Simultaneous Component Analysis
  - Multiple Factor Analysis
- Sparse multivariate methods

Canonical Correlation Analysis

- Given two data tables Canonical Correlation Analysis searches successive pairs of axes (one per table) with maximum correlation.
- Analysis may lead to axes
  - with high correlation,
  - but possibly low percentages of explained variance.
- It may be difficult to give a biological interpretation to these axes.
- Besides this, CCA cannot be used when the number of variables is greater than that of samples!

Alternatives extensions to Canonical Correlation Analysis

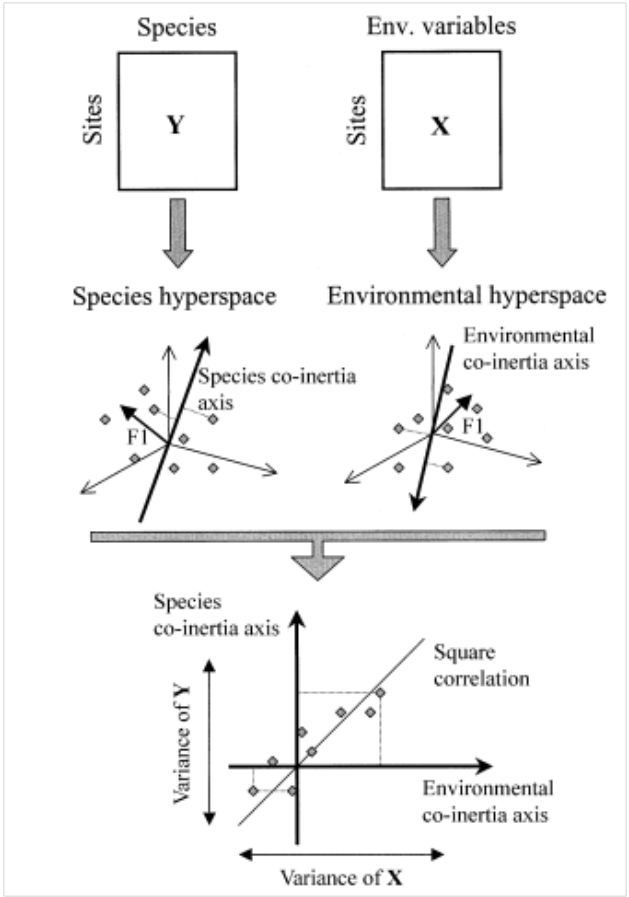
From CCA to CIA (Coinertia Analysis)

- CIA searches for pairs of axes with maximum covariance (instead of correlation).
- This ensures that CIA axes will have both
  - a high correlation and also
  - good % of explained variance for each table
- CIA can be applied to data where number of variables (genes) exceeds number of cases

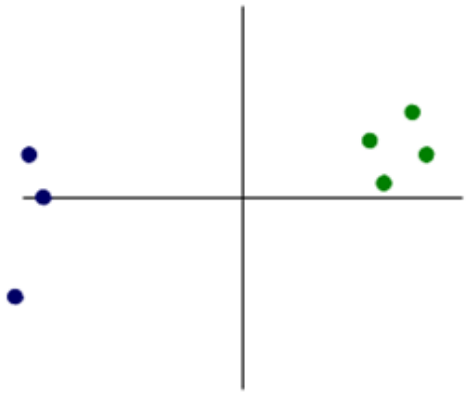
Principles of coinertia analysis (COIA).

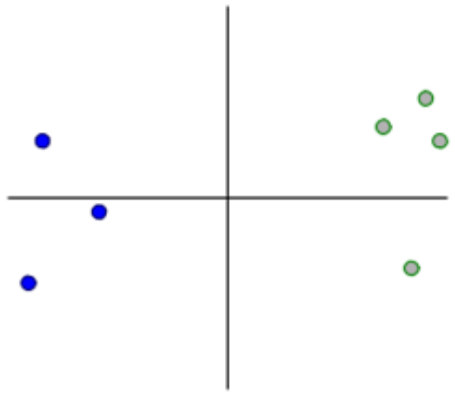
Properties:

- Given 2 datasets Y, X with
  - Same number of cases
  - # of vars >> # cases
  - Any type of vars
- Looks for 2 ordinations most similar
  - Find successive pairs of axes (a_i, b_i) such that
  - $Cov(a_i, b_i)$ is maximum



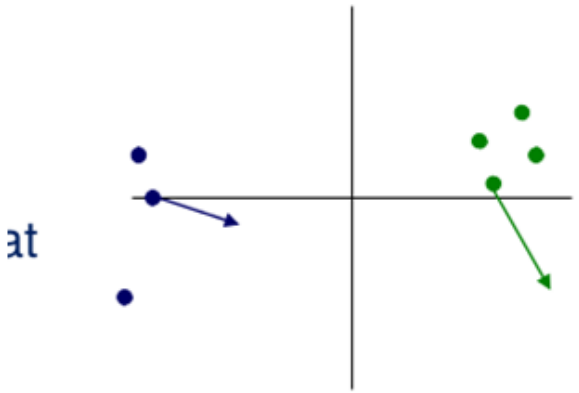
Coinertia plots



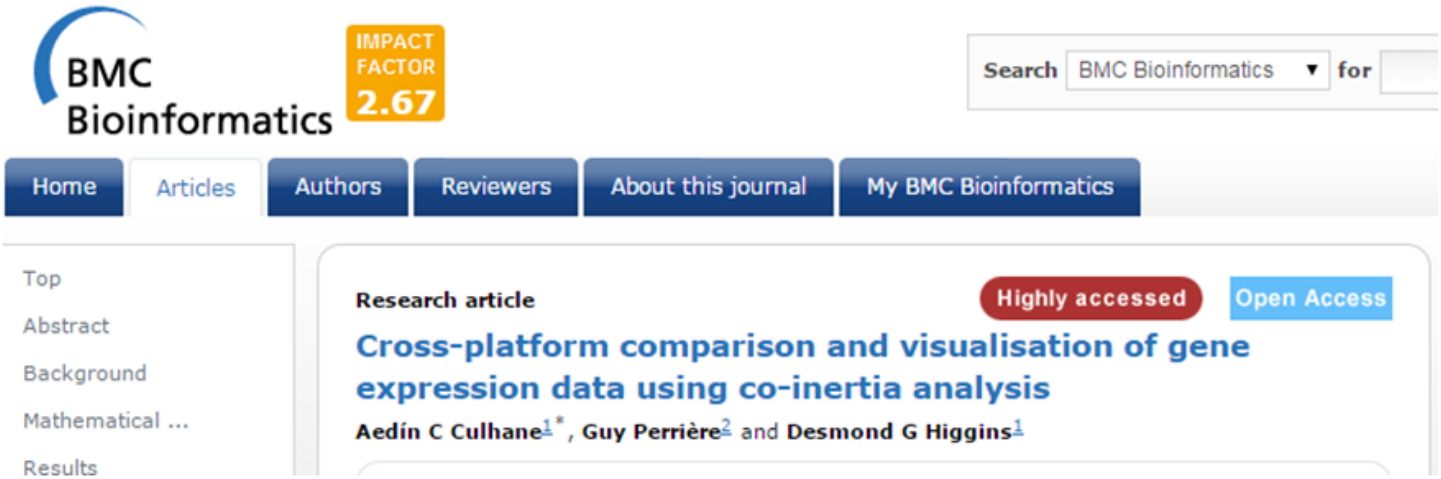


2 datasets A, B with same cases,
Cases ordination may differ between them
When are plotted together, use

- Points for position from ordination 1
- Arrows for position from orination 2
- Lines joininin points and arrows show divergent cases



Examples



2162 DOI 10.1002/pmic.200600898 *Proteomics* 2007, 7, 2162–2171

RESEARCH ARTICLE

A multivariate analysis approach to the integration of proteomic and gene expression data

Ailís Fagan[1], Aedín C. Culhane[2, 3] and Desmond G. Higgins[1]

[1] Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin, Ireland
[2] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
[3] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

Cross-platform comparison

Circles ●

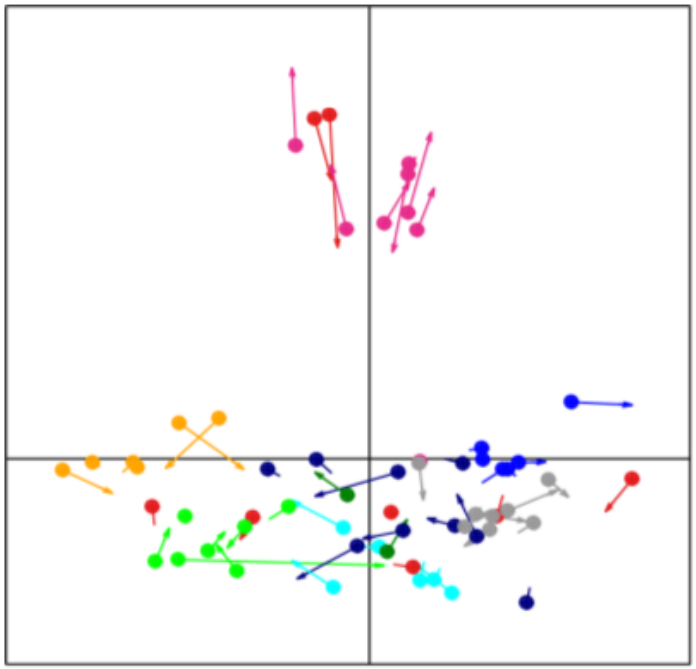
Spotted cDNA arrays

Arrows →

Affymetrix

Circles, arrows are joined by line.

Length of line is $\propto$ to divergence between gene expression profiles.



Integrative analysis of genomic and proteomic data

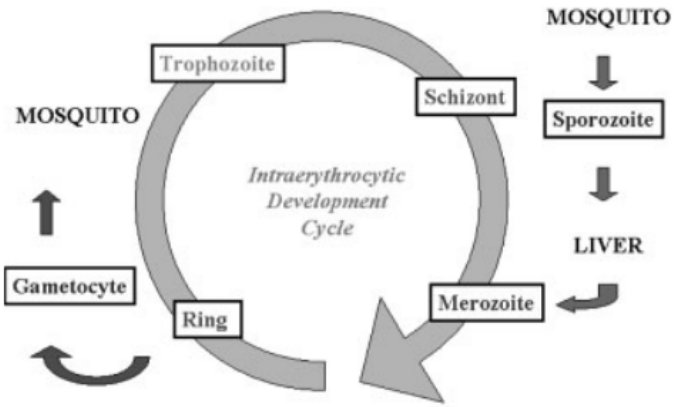
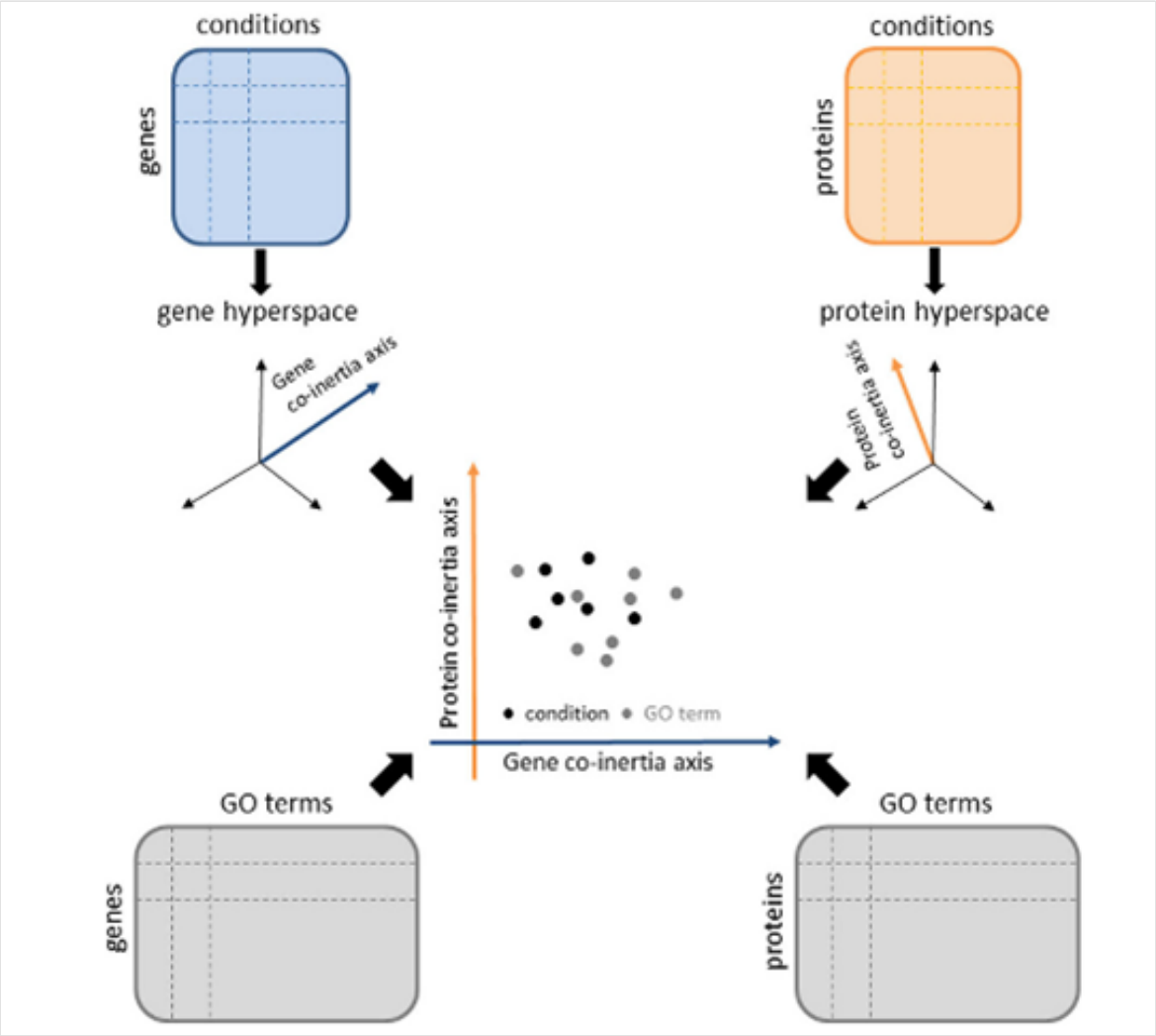


Figure 1. Summary of malaria life cycle. The six life cycle stages: merozoite, ring, trophozoite, schizont, gametocyte and sporozoite.

Properties:

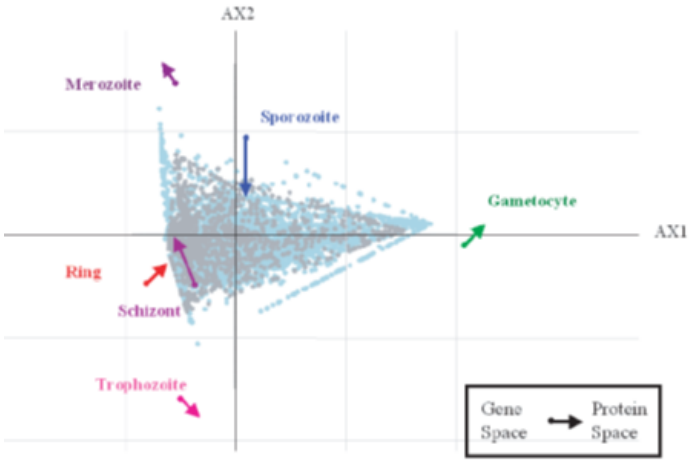
- Gene expression levels measured by 4294 probes across all six stages using Affymetrix custom arrays
- Multi-dimensional protein identification technology (MuDIPT) used to identify and measure abundances of 2904 proteins.

Multiple Colnertia Analysis

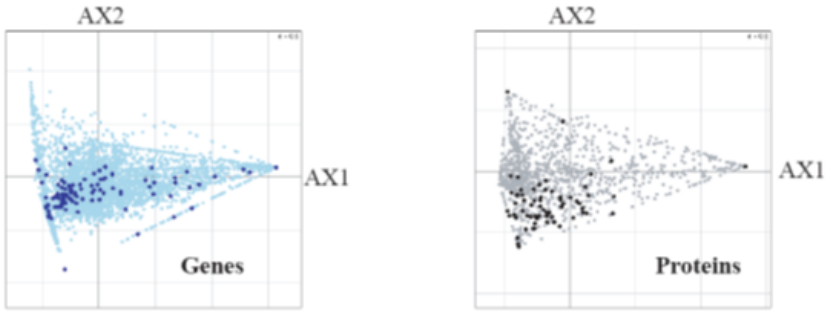


Project GO terms on Genes and Proteins space

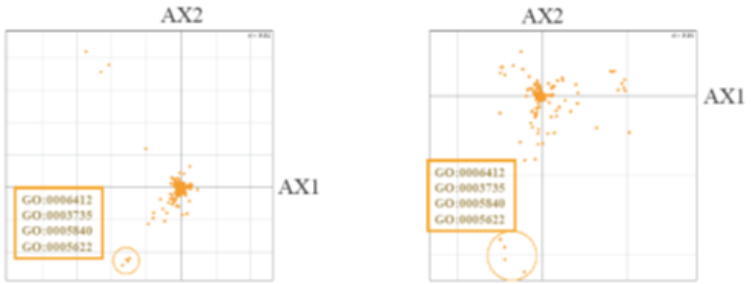
Sample with variables (tri-plot)



Variables



GO Terms



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

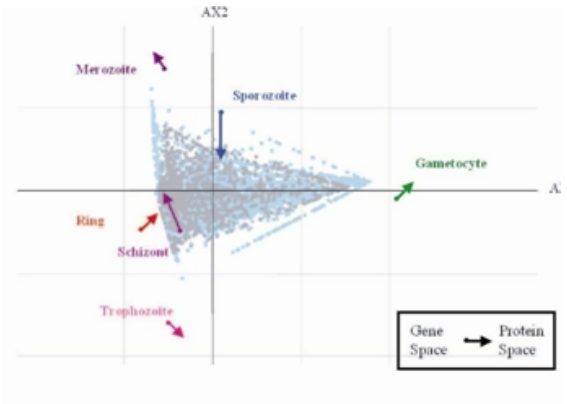
Axis 2 (vertical) 4.8% variance. Splits invasive stages (Merozoite and Sporozoite stages which invade red blood)

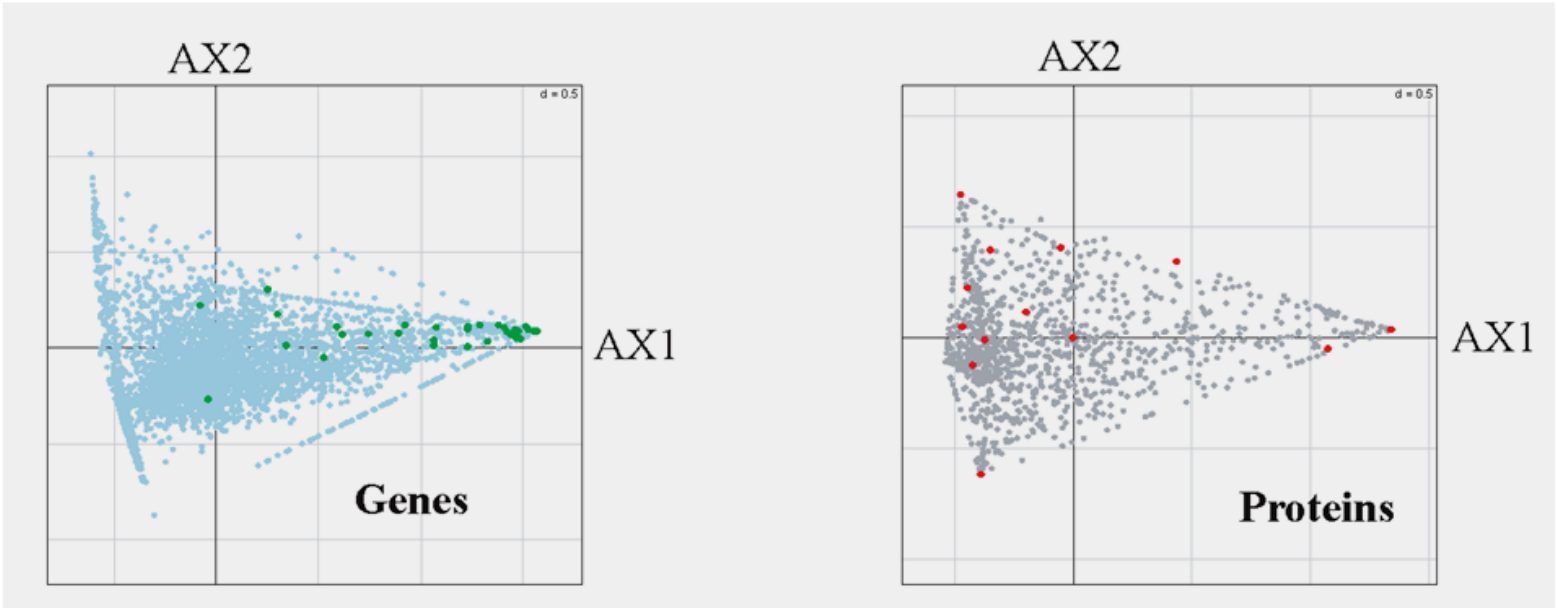
Detecting translationally repressed genes

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.





made4

Integrated pathway analysis of multiple 'omics datasets

Aedin Culhane

Friday August 1, 2014

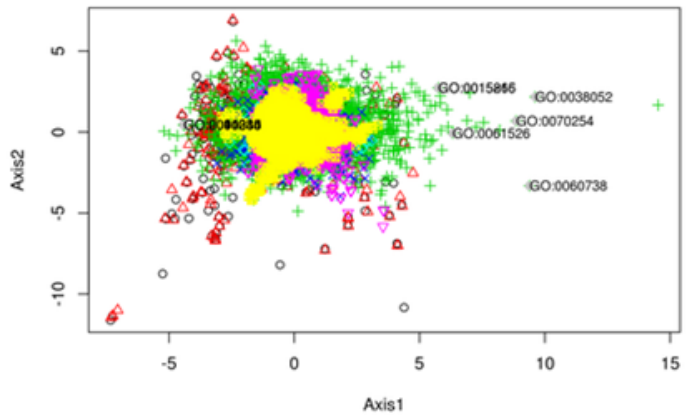
Introduction

Rapid advances and reducing costs have enabled laboratories to apply a multi-omic approach to biological systems. Studies frequently measure several biological molecules, including mRNA, proteins, metabolites, lipids and phosphoproteins using different technological platforms generating multiple datasets on the same set of biological samples. We have described the following different exploratory data analysis methods that enable one to identify co-relationships between high dimensional datasets.

- Multiple Co-Intertia Analysis (MCIA) simultaneously projects several datasets into the same dimensional space, transforming diverse sets of features onto the same scale and allows one identify the co-structure between datasets (Meng, Kuster, Culhane, and Moghaddas Gholami (2014)).
- iBBiG is a biclustering algorithm which we wrote for analysis of noisy binary data (Gusenleitner, Howe, E (2012)). We developed it to identify biclusters in results from GSEA/pathway analysis from >1,000 diff vectors which each contain p-values enrichment scores for many thousands of pathways, we discretize apply iBBiG to find pathways associated with covariates across many studies.

Today, I will provide an overview of the first MCIA.

TCGA Breast Cancer Dataset



Sparse multivariate methods

- High-dimensional data often contain many variables that are irrelevant for predicting a response or for an accurate group assignment.
- The inclusion of such variables in a predictive model leads to a loss in performance, even if the contribution of the variables to the model is small.
- Sparse methods are able to suppress these variables. This is possible by adding an appropriate penalty term to the objective function of the method

From Multiple Regression to Sparse PLS

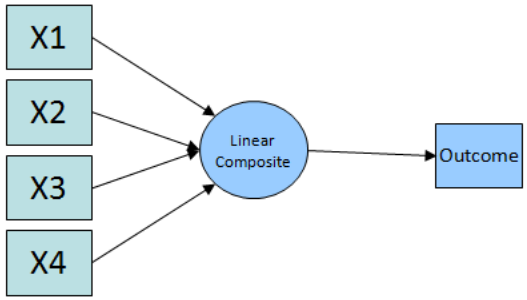
- The multiple regression approach creates a linear combination of the predictors that best correlates with the outcome.
- Standard multiple regression cannot be used in this context
- Instead Principal Components Regression can be used

Principal Components Regression

- With principal components regression, we
  - first create several linear combinations (equal to the number of predictors) and
  - then use those composites in predicting the outcome instead of the original predictors
- PCR has good properties
  - Components are independent $->$ deals with collinearity
  - Can use fewer of components while still retaining most of the predictor variance

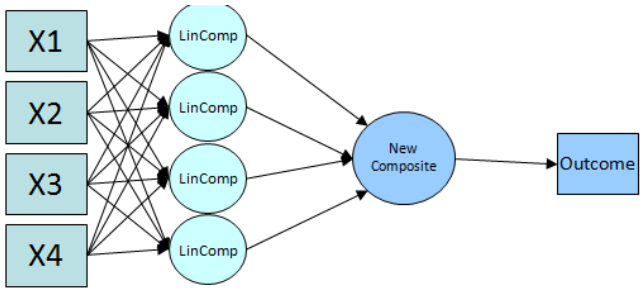
Multiple Regression

$Y = XB + E$, Note the bold, we are dealing with vectors and matrices



Principal Components Regression

$T = XW \quad Y = TQ + E$ Here T refers to our components, W and Q are coefficient vectors as B is above



$Y = XB + E$ where $B = BQ$

Partial Least Squares

- Partial Least Squares is just like PC Regression except in how the component scores are computed
  - PC regression = weights are calculated from the covariance matrix of the predictors only.
  - PLS = weights are calculated (reflect the covariance structure between) from the covariance matrix of both predictors and response.
- PLS, like regression, aims at finding a set of scores that has highest correlation with the response.

Why PLS?

- PLS (unlike PCR)
  - extends to multiple outcomes
  - allows for dimension reduction
  - creates components with an eye to the predictor-DV relationship
- PLS (unlike MR) is less restrictive in terms of assumptions.
  - Distribution free
  - No collinearity
  - Independence of observations not required
- Unlike Canonical Correlation, it maintains the predictive nature of the model

Mathematical formulation (PLS)

- Partial Least Squares regression maximizes the covariance between each linear combination (components) associated to each data set

$$max \quad cov(X_{h-1}u_h, Y_h v_h), \quad h = 1...H$$

$$||u_h|| = 1, ||v_h|| = 1$$

  where $X\,(n \times p)$ and $Y\,(n \times q)$ are two datasets.
- Similarly to PCA, the PLS components indicate the similarities between samples (useful plots!).
- Loading vectors indicate the contribution of the variables of the same type to the PLS component (useful for variable selection)!

From PLS to sparse PLS

- PLS regression maximises the covariance between each linear combination (components) associated to each data set
- sparse PLS has been developed to include variable selection from both data sets.
- Two modes ('regression' and 'canonical') available to model the relationship between the two sets

Mathematical formulation

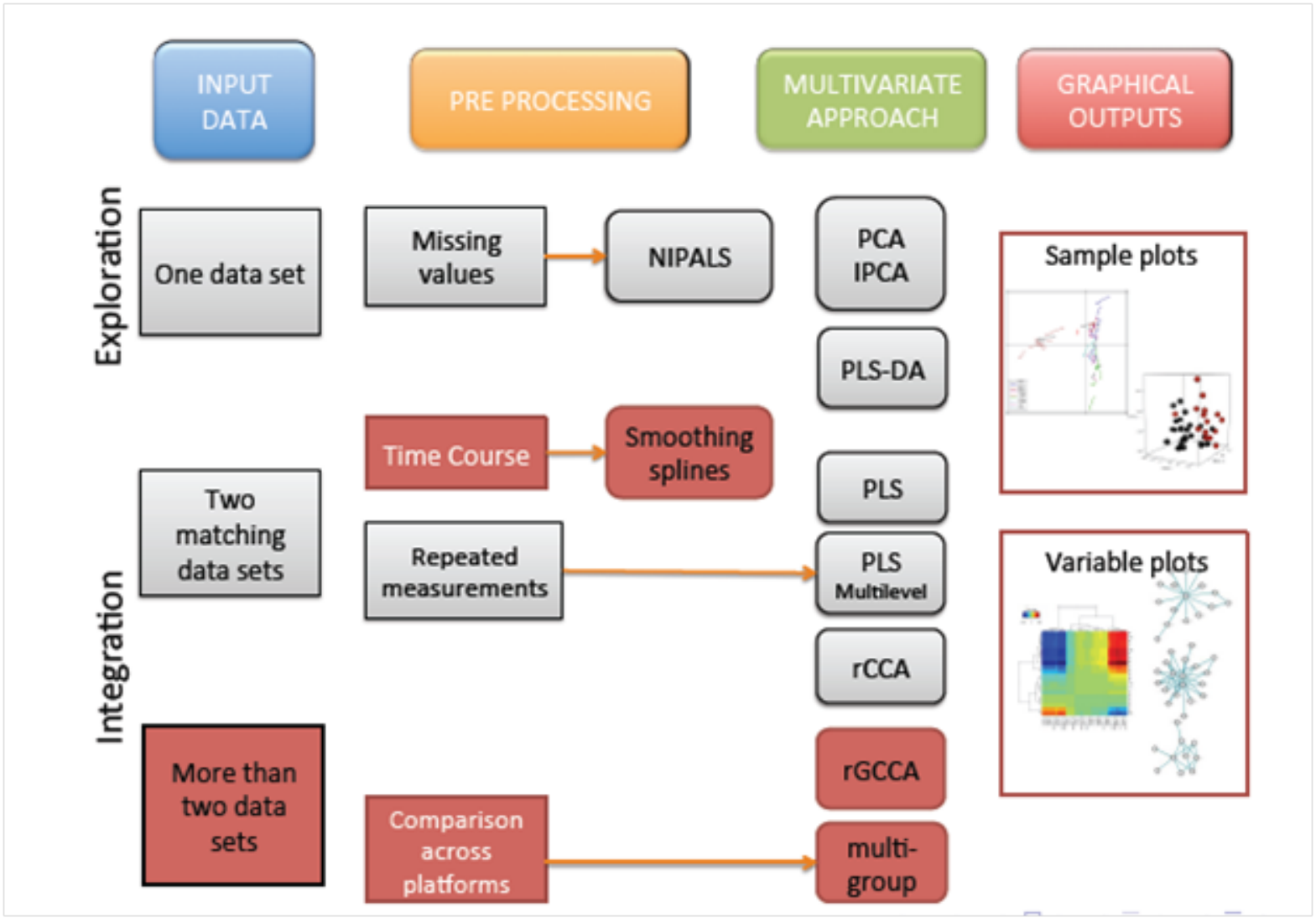
Use the PLS-SVD variant that directly gives the latent variables and loading vectors and low rank rank approximation. Let $M_h = X_h' Y_h$, sparse PLS solves the optimization problem:

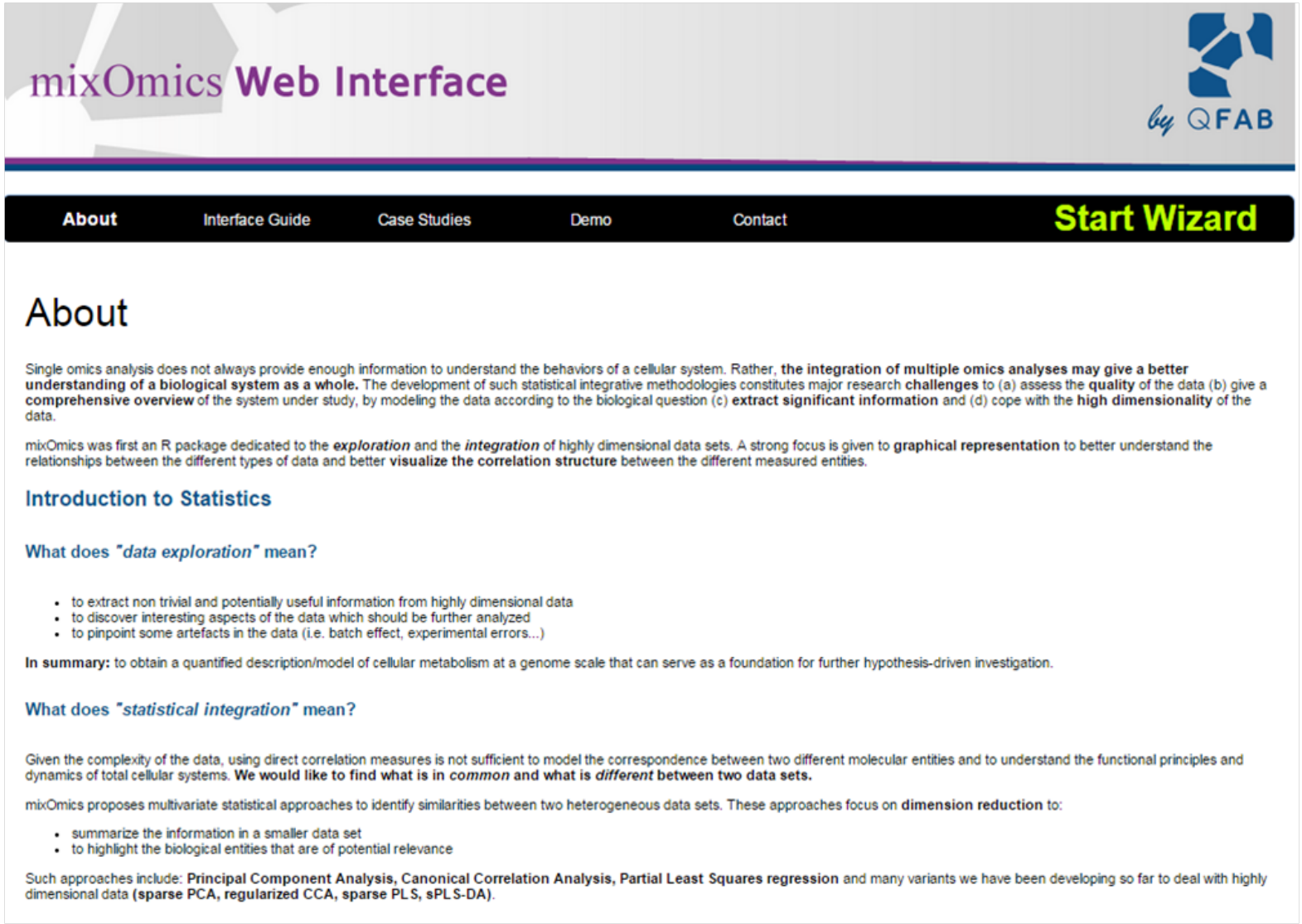
$$min||M_h - u_h v_h'||_F^2 + P_{\lambda_1}(u_h) + P_{\lambda_2}(v_h)$$

where P_{λ_1} is a penalty function

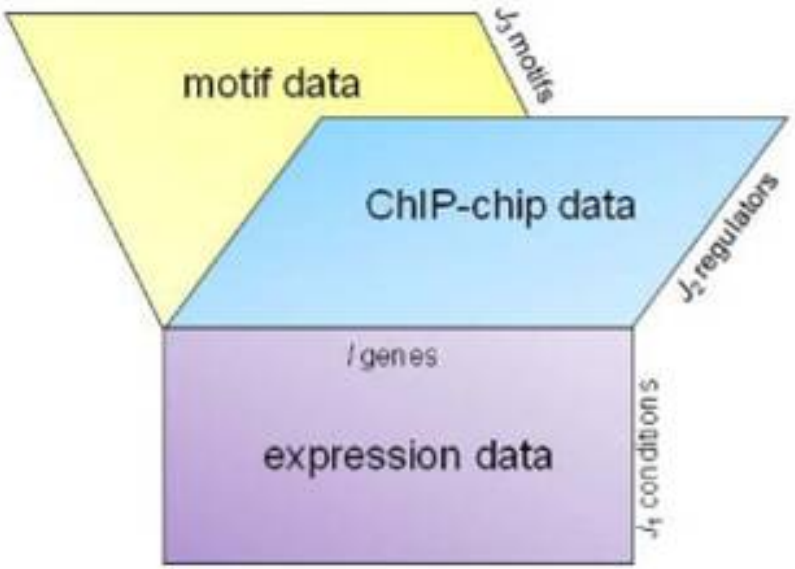
- obtain simultaneously sparse loadings u_h and v_h
- simultaneous select variables from both data sets which are correlated across samples

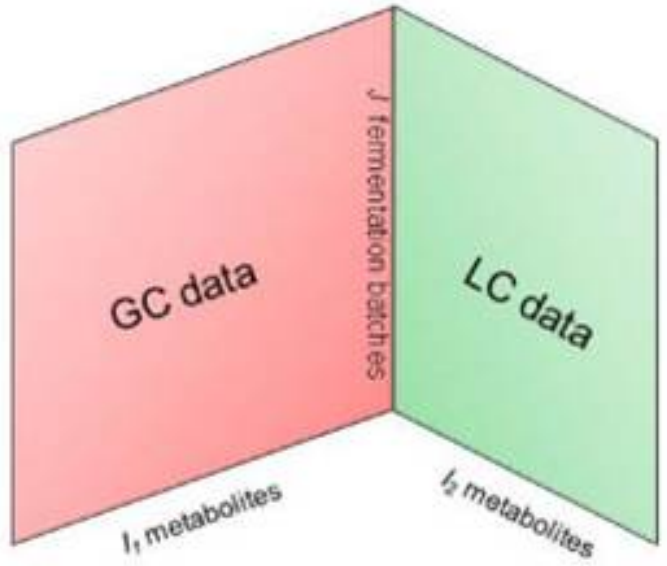
Mixomics


INPUT DATA
PRE PROCESSING
MULTIVARIATE APPROACH
GRAPHICAL OUTPUTS
Exploration
One data set
Missing values
NIPALS
PCA
IPCA
PLS-DA
Sample plots
Integration
Two matching data sets
Time Course
Smoothing splines
PLS
Repeated measurements
PLS
Multilevel
Variable plots
rCCA
More than two data sets
rGCCA
Comparison across platforms
multi-group

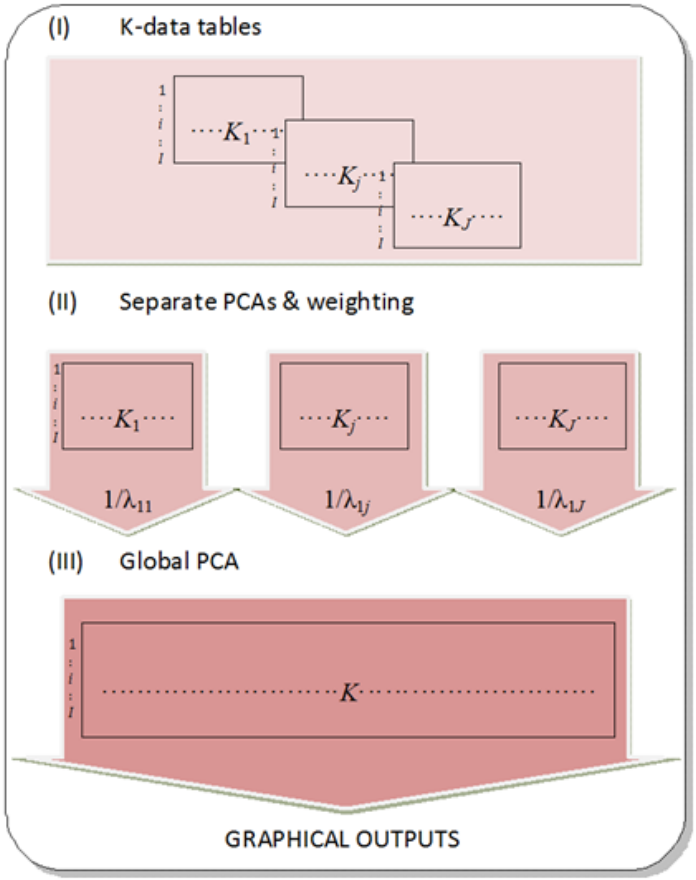

mixOmics Web Interface
by QFAB
About
Interface Guide
Case Studies
Demo
Contact
Start Wizard
About
Single omics analysis does not always provide enough information to understand the behaviors of a cellular system. Rather, the integration of multiple omics analyses may give a better understanding of a biological system as a whole. The development of such statistical integrative methodologies constitutes major research challenges to (a) assess the quality of the data (b) give a comprehensive overview of the system under study, by modeling the data according to the biological question (c) extract significant information and (d) cope with the high dimensionality of the data.
mixOmics was first an R package dedicated to the exploration and the integration of highly dimensional data sets. A strong focus is given to graphical representation to better understand the relationships between the different types of data and better visualize the correlation structure between the different measured entities.
Introduction to Statistics
What does "data exploration" mean?
to extract non trivial and potentially useful information from highly dimensional data
to discover interesting aspects of the data which should be further analyzed
to pinpoint some artefacts in the data (i.e. batch effect, experimental errors...)
In summary: to obtain a quantified description/model of cellular metabolism at a genome scale that can serve as a foundation for further hypothesis-driven investigation.
What does "statistical integration" mean?
Given the complexity of the data, using direct correlation measures is not sufficient to model the correspondence between two different molecular entities and to understand the functional principles and dynamics of total cellular systems. We would like to find what is in common and what is different between two data sets.
mixOmics proposes multivariate statistical approaches to identify similarities between two heterogeneous data sets. These approaches focus on dimension reduction to:
summarize the information in a smaller data set
to highlight the biological entities that are of potential relevance
Such approaches include: Principal Component Analysis, Canonical Correlation Analysis, Partial Least Squares regression and many variants we have been developing so far to deal with highly dimensional data (sparse PCA, regularized CCA, sparse PLS, sPLS-DA).

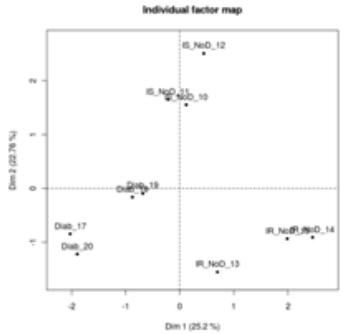
Simultaneous Component Analysis

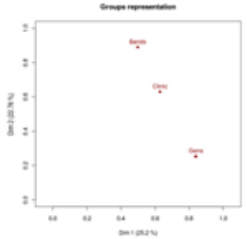


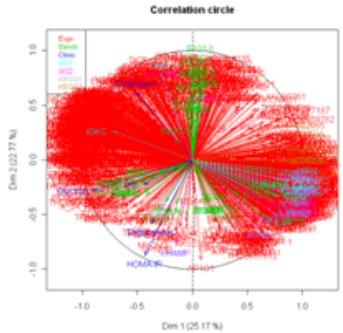


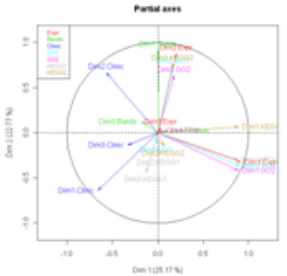
Multiple Factor Analysis



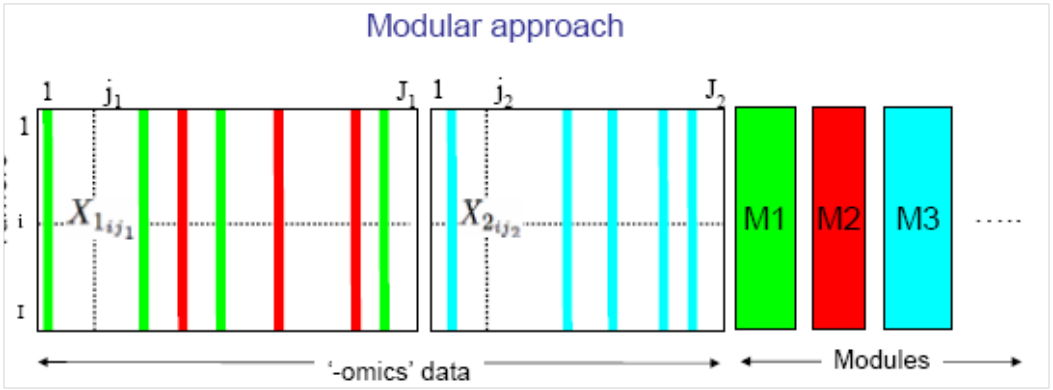








MFA admits supplementary info

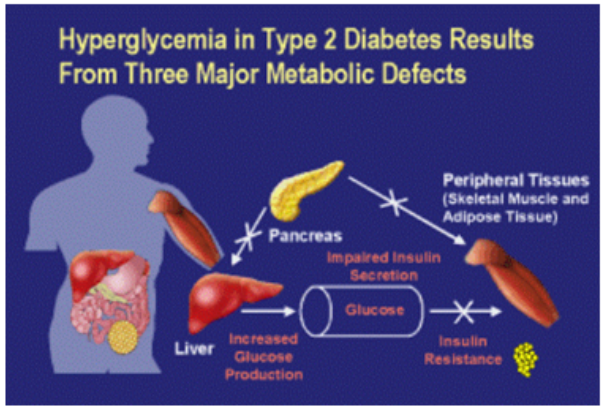


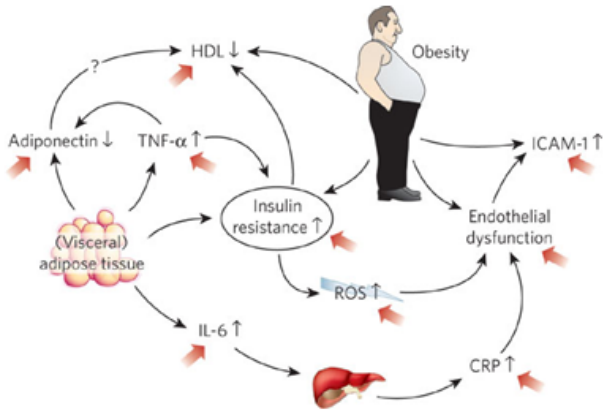
The assets of MFA appear when

- integrating both numerical and categorical groups of variables,
- and when supplementary groups of data need to be added in the analysis.

Insulin Resistance

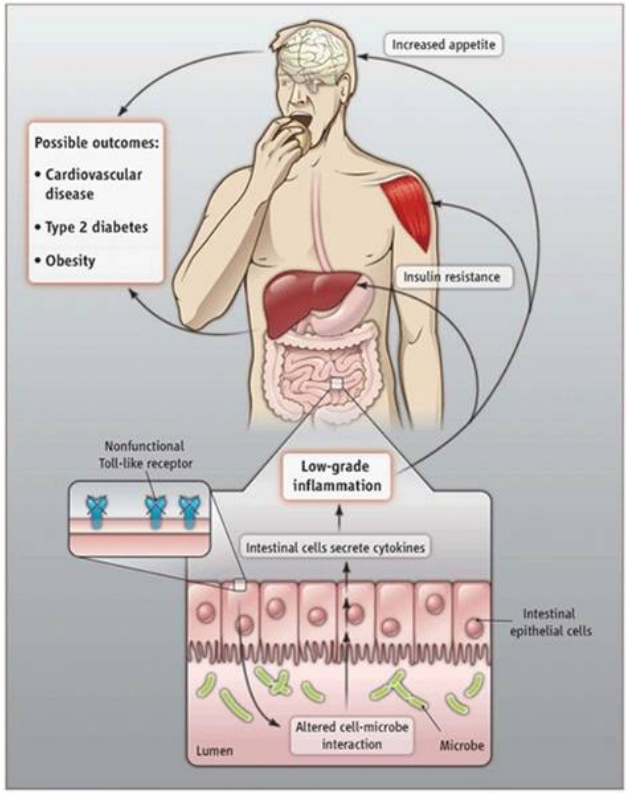
- Insulin resistance means cells become less sensitive to insulin,
- This provokes the pancreas to over-compensate by working harder and releasing even more insulin.
- Insulin-resistance + Insulin over-production leads to two common outcomes: diabetes, or obesity





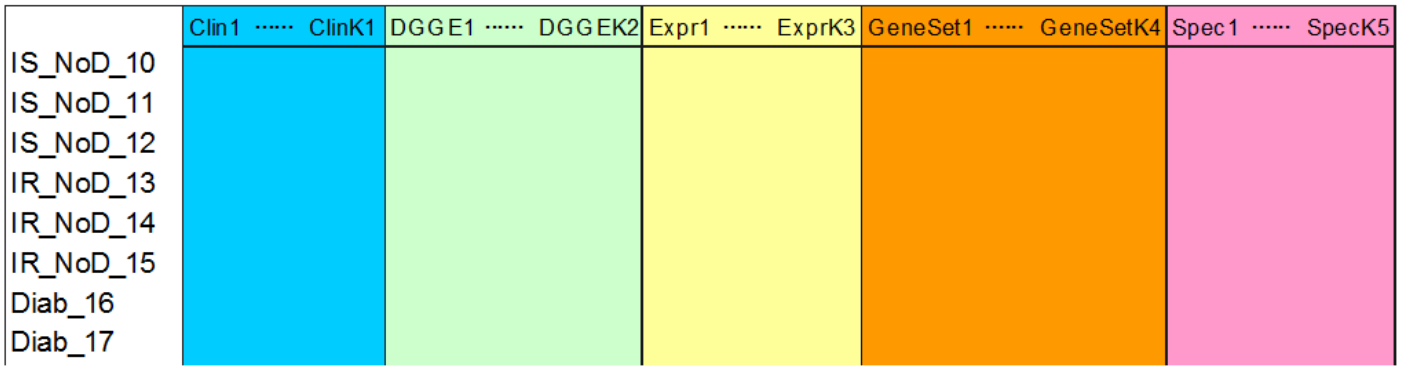
IS/IR and Gut Microbiota

- Human gut microbiome is related to health & weight
  - varies in healthy people
  - varies in lean and obese
- It is reasonable to postulate insuline sensitivity to be associated with changes in bacterial microflora.

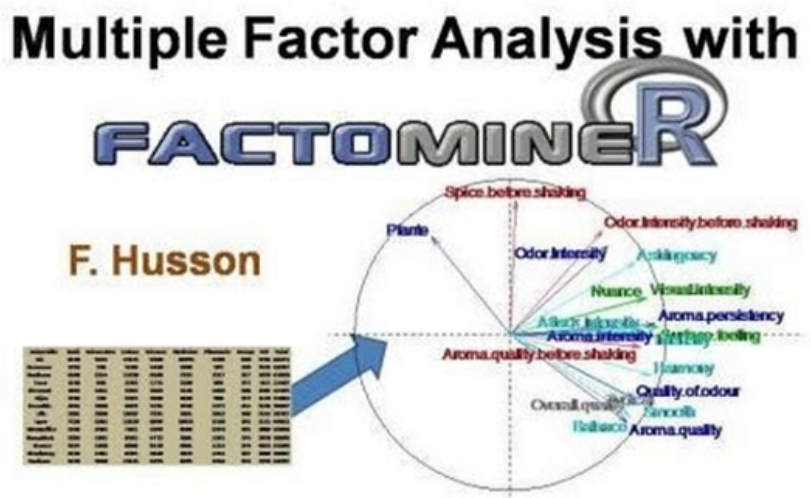


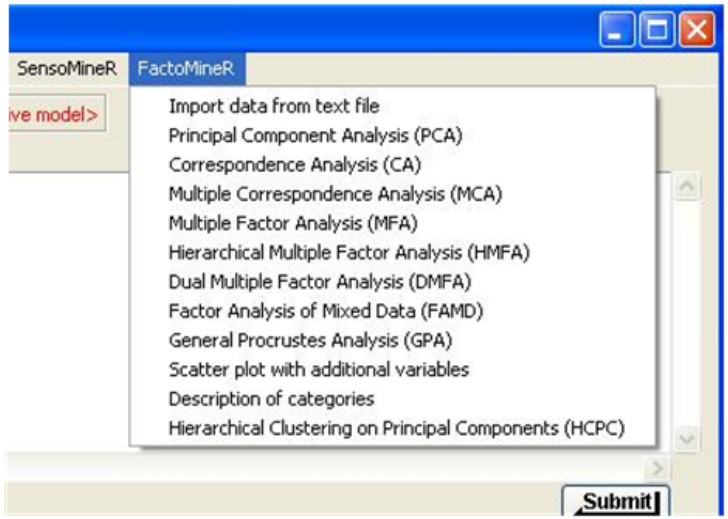
Data for relating IR/IS with Microbiome

- Clinical variables (BMI, Homa, Ins, HDL, ...)
- Microarrays
  - Expression matrix an related annotations (GO)
- Microbial flora diversity based on
  - Denaturing Gradient Gel Electrophoresis



FactoMineR


Multiple Factor Analysis with
FACTOMINER
F. Husson
Spice.before.shaking
Odor.Intensity.before.shaking
Plante
Odor.Intensity
Nuance
Aroma.persistency
Aroma.quality.before.shaking
Quality.of.odour
Aroma.quality


SensoMineR
FactoMineR
Import data from text file
Principal Component Analysis (PCA)
Correspondence Analysis (CA)
Multiple Correspondence Analysis (MCA)
Multiple Factor Analysis (MFA)
Hierarchical Multiple Factor Analysis (HMFA)
Dual Multiple Factor Analysis (DMFA)
Factor Analysis of Mixed Data (FAMD)
General Procrustes Analysis (GPA)
Scatter plot with additional variables
Description of categories
Hierarchical Clustering on Principal Components (HCPC)
Submit

In summary

- Many families of multivariate methods available
  - Need to be related, classified, filtered
- Few common ideas
  - Rely on spectral decomposition theory and low rank matrix approximation
  - Use weights to combine data
- Promising approaches are those that
  - Allow inclusion of biological information
  - Provide hints for interpretability
  - Implementations are available

Other topics

- Biomarker selection from integrative analysis
- Network inference/analysis
- Specific data integration problems
  - microRNA-mRNA; CGH-expression; transcriptomics-proteomics; methylation-expression;
- Sources of data for integrative analysis
- Non-R and Web tools
- `http://eib.stat.ub.edu/IODA`

Some lessons learned

- There is no universal "IODA" method
  - Biological question should be first
- All data are not equally informative
  - Gene expression and what else
- There are more mathematical/statistical tools than end-user bioinformatical solutions
  - Opportunity for developers

Thanks for your attention