# 2

# Principal Component Analysis in the Era of «Omics» Data

Louis Noel Gastinel
*University of Limoges, University Hospital of Limoges, Limoges*
*France*

## 1. Introduction

### 1.1 Definitions of major «omics» in molecular biology and their goals

The «omics» era, also called classically the post-genomic era, is described as the period of time which extends the first publication of the human genome sequence draft in 2001 (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). Ten years after that milestone, extensive use of high-throughput analytical technologies, high performance computing power and large advances in bioinformatics have been applied to solve fundamental molecular biology questions as well as to find clues concerning human diseases (cancers) and aging. Principal «omics», such as Gen-*omics*, Transcript-*omics*, Prote-*omics* and Metabol-*omics,* are biology disciplines whose main and extremely ambitious objective is to describe as extensively as possible the complete class-specific molecular components of the cell. In the «omics» sciences, the catalog of major cell molecular components, respectively, genes, messenger RNAs and small interfering and regulatory RNAs, proteins, and metabolites of living organisms, is recorded qualitatively as well as quantitatively in response to environmental changes or pathological situations. Various research communities, organized in institutions both at the academic and private levels and working in the «omics» fields, have spent large amounts of effort and money to reach. standardization in the different experimental and data processing steps. Some of these «omics» specific steps basically include the following: the optimal experimental workflow design, the technology-dependent data acquisition and storage, the pre-processing methods and the post-processing strategies in order to extract some level of relevant biological knowledge from usually large data sets. Just like Perl (Practical Extraction and Report Language) has been recognized to have saved the Human Genome project initiative (Stein, 1996), by using accurate rules to parse genomic sequence data, other web-driven. programming languages and file formats such as XML have also facilitated «omics» data dissemination among scientists and helped rationalize and integrate molecular biology data.

Data resulting from different «omics» have several characteristics in common, which are summarized in Figure 1: (a) the number of measured variables n (SNP, gene expression, proteins, peptides, metabolites) is quite large in size (from 100 to 10000), (b) the number of samples or experiments p where these variables are measured associated with factors such as the pathological status, environmental conditions, drug exposure or kinetic points

(temporal experiments) is rather large (10 to 1000) and (c) the measured variables are organized in a matrix of n x p dimensions. The cell contents of such a matrix usually record a metric (or numerical code) related to the abundance of the measured variables. The observed data are acquired keeping the lowest amount of possible technical and analytical variability. Exploring these «omics» data requires fast computers and state-of-the-art data visualization and statistical multivariate tools to extract relevant knowledge, and among these tools PCA is a tool of choice in order to perform initial exploratory data analysis (EDA).

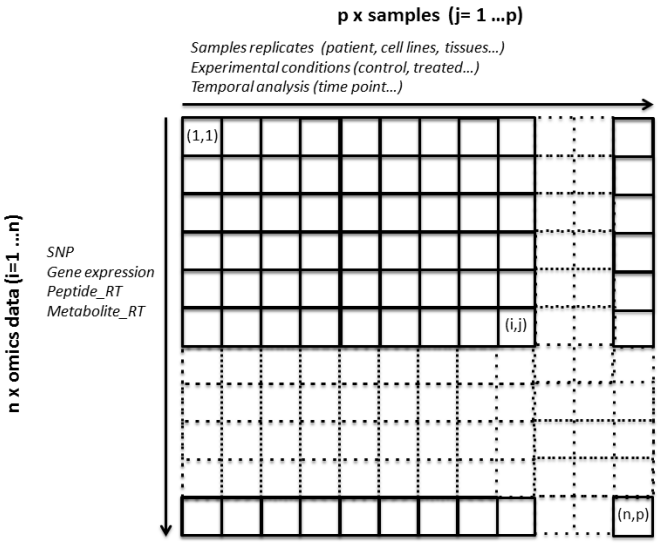



Fig. 1. General organization of raw «omics» data represented in a n x p matrix.

Rows contain the measured quantitative variables (n) and columns contain the samples or experimental conditions tested (p) from which variables n are measured and for which grouping information or factors is generally present. Each cell (i,j) of this matrix contains a measured quantitative information which is usually the abundance of the molecule under study.

### 1.1.1 Genomics and genetics data are different

Genomics and genetics data are of different types. Genomics data are related mainly to the collection of DNA sequences modeled as linear strings composed of the four nucleotides symbolized by the letters A, C, G and T (bases). These strings are usually obtained following large sequencing efforts under the supervision of academic and private consortia. NextGen sequencing technologies are used to acquire the data and, specialized softwares are used to assemble sequences in one piece in order to complete an entire genome of thousands of megabases long. The final result of these extensive and costly efforts is the establishment of the genome sequence of all living organisms and particularly the human genome. Genomics has been particularly successful these last few years in determining micro-organism genomes such as bacteria and viruses. Genomics is regularly used in academic research and even proposes on-demand service for the medical community to obtain emerging

pathological genomes (SRAS, toxic strains of *Escherichia coli* …) that allow a fast medical response. Genomics aims to attain the technical challenge of obtaining 99.99% accuracy at the sequenced nucleotide level, and completeness and redundancy in the genome sequence of interest. However, the understanding or interpretation of the genome sequence, which means finding genes and their regulatory signals as well as finding their properties collected under the name "annotations", are still the most challenging and expensive tasks.

Genetics data, or genotype data, are related to the sequencing efforts on the human genome, particularly at the individual level. Genetics data record that the status of some nucleotides found at a certain position in the genome are different from one person to another. These base- and position-specific person-to-person variations are known as SNP or Single Nucleotide Polymorphism. When the frequency of the variation in a population is greater than 1%, this variation is considered as a true polymorphism possibly associated with traits (phenotypes) and genetic diseases (mutations). Moreover this information is useful as a genetic biomarker for susceptibilities to multigenic diseases or ancestrality and migration studies.

### 1.1.2 Transcriptomics data

Transcriptomics data consist in the recording of the relative abundance of transcripts or mature messenger RNAs representing the level of gene expression in cells when submitted to a particular condition. Messenger RNAs are the gene blueprints or recipes for making the proteins which are the working force (enzymes, framework, hormones…) in a cell and allow the cell's adaptation to its fast changing environment. Transcriptomics give a snapshot of the activity of gene expression in response to a certain situation. Generally mRNA abundances are not measured on an absolute scale but on a relative quantitative scale by comparing the level of abundance to a particular reference situation or control. Raw transcriptomics data associated with a certain gene g consist in recording the ratio of the abundances of its specific gene transcript in two biological situations, the test and the control. This ratio reveals if a particular gene is over- or under- expressed in a certain condition relative to the control condition. Moreover, if a set of genes respond together to the environmental stress under study, this is a signature of a possible common regulation control (Figure 2). Furthermore, transcriptomics data are usually organized as for other «omics» data as large tables of n x p. cells with p samples in columns and n genes in rows (Figure 1). A data pre-processing step is necessary before analyzing transcriptomics data. It consists in $\log_2$ intensity ratios transformation, scaling the ratios across different experiments, eliminate outliers. Multivariate analysis tools, particularly PCA, are then used to find a few genes among the thousands that are significantly perturbed by the treatment. The signification level of the perturbation of a particular gene has purely statistical value and means that the level of measured variation in the ratio is not due to pure chance. It is up to the experimentalist to confirm that it is truly the biological factor under study, and not the unavoidable variation coming from technical or analytical origin inherent to the acquisition method, that is responsible for the observations. To estimate this significance level it is absolutely necessary to measure ratios on a certain replicative level, at least three replicates per gene and per situation. ANOVA and multiple testing False Discovering Rate (FDR) estimates are generally used. Further experimental studies are mandatory to confirm transcriptomics observations. Moreover, Pearson correlation coefficient and different linkage clustering methods are used for each gene in order to perform

their hierarchical clustering and to group genes with similar behavior or belonging to the same regulation network.

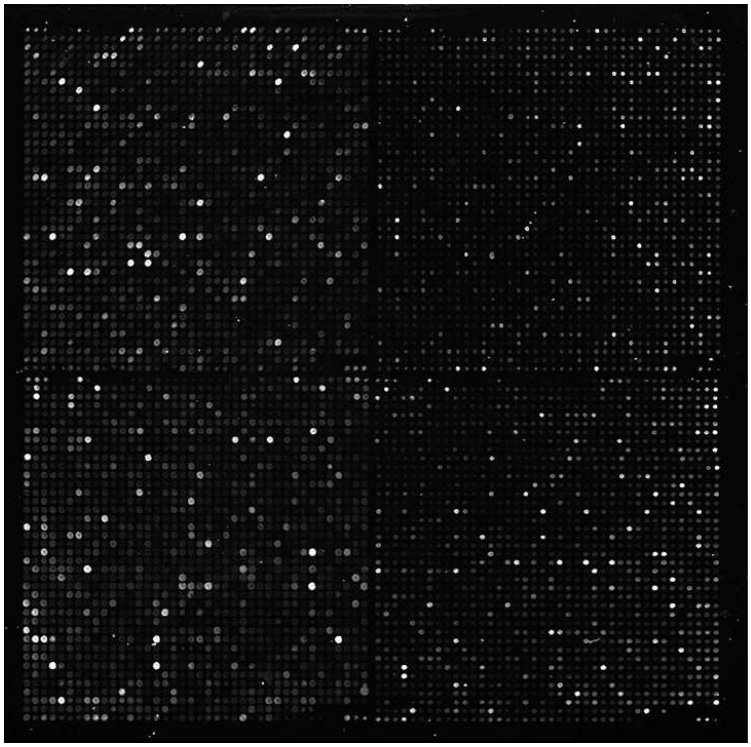


Fig. 2. A picture of a DNA microarray used in high-throughput transcriptomics.

DNA chip of 18 x 18 mm In size containing 6400 yeast gene Specific sequences organized as a matrix in which gene coordinates (x,y) are known. After hybridization with transcripts labeled respectively with green and red fluorochromes from two situations (treated versus untreated), 2 images in red and green fluorescence are recorded and superposed. Spot intensity seen on this image is then mathematically converted to a ratio of relative abundance of gene expression in the two situations under study (DeRisi et al., 1997).

### 1.1.3 Proteomics and metabolomics data

Proteomics and metabolomics data consist in measuring absolute or relative abundances of proteins and metabolites in the organism, tissue or cells after their proper biochemical extraction. These two fundamental and different classes of molecules are important for preserving cell integrity and reactivity to environment changes. These molecules are generally recognized and their abundances measured by mass spectrometry technologies after a liquid (HPLC) or gas (GC) chromatographic separation is performed to lower the high complexity level of analytes in the sample under study. Proteins have the large size of a few thousands of atoms and weigh a few thousands of Daltons (1 Dalton is the mass of a hydrogen atom) in mass, contrary to metabolites that are smaller molecules in size and mass

(less than 1000 Daltons). Mass spectrometers are the perfect analytical tool to separate physically ionized analytes by their mass-to-charge ratio (m/z) and are able to record their abundance (peak intensity). Mass spectrometry data are represented graphically by a spectrum containing abundances versus m/z ratios or by a table or a peak list with two columns containing m/z and abundances after performing a de-isotopic reduction step and a noise filtration step.

Because of the large size of protein molecules, entire proteins should be cut in small pieces, called peptides, of 10-15 amino acids by using a protease enzyme trypsin. These peptides then have the right size to be analyzed directly by mass spectrometers. Peptide abundances are recorded. and their sequences even identified by collision-induced fragmentation (CID) breaking their peptide bonds, which some mass spectrometers instruments can perform (Triple Quadrupole mass spectrometrer in tandem, MALDI TOF TOF, Ion traps).

Raw data from metabolomics and proteomics studies originating from mass spectrometry techniques have the same basic contents. However, contrary to previous «omics», analytes are first separated by a chromatographic step and one analyte is characterized by its unique retention time (rt) on the separation device, its mass-to-charge ratio (m/z) and its abundance (a). This triad (rt - m/z - a) is a characteristic of the analyte that is measured accurately and found in the final «omics» data matrix n x p. Because of the separation step, multiple chromatography experiments should be normalized on both the scale of abundance and the scale of retention time to be further compared. A relevant multiple alignment of the chromatographic separations of different p samples is necessary and is performed by using sophisticated methods and models (Listgarten & Emili, 2005). This alignment step consists in recognizing which analyte is recorded in a given retention time bin and in a given m/z bin. Analytes found in common in the chosen bin are by definition merged in intensity and considered to be the same analyte. The same m/z analyte is recorded across multiple chromatographic steps and should be eluted at the same rt with some degree of tolerance both on rt (a few minutes) and on m/z (a few 0.1 m/z). The rows in the prote- and metabol- «omics» final matrix n x p contain the. proxy "m/z_rt," or "feature" and on the columns are the samples where the analytes come from. The cell content of this matrix record the abundance. "m/z_rt" is a set of analytes which have the same m/z with the same retention time rt, hopefully only one. Data can also be visualized as a 3D matrix with 3 dimensions: rt, m/z and abundances (Figure 3). For convenience it is the "m/z_rt" versus the 2D sample matrix which is further used in EDA for sample comparisons. The absolute value of intensity of the m/z analyte with retention rt corresponds to the mass spectrometry response given by its detector (cps).

### 1.2 Technologies generating «omics» data, their sizes and their formats

#### 1.2.1 Genetics data format

Genome-wide studies using genetics data consist in recording the status of a particular DNA position or genotype in the genome called SNP or Single Nucleotide Polymorphism among few thousand of genes for a certain number of samples. The SNP status is obtained by accurately sequencing genomic DNA and recording its sequence in databases such as Genbank (www.ncbi.nlm.nih.gov/genbank). The SNP status is then coded by a simple number, 0, 1, 2, according to the nature of the nucleotide found at the genome's particular

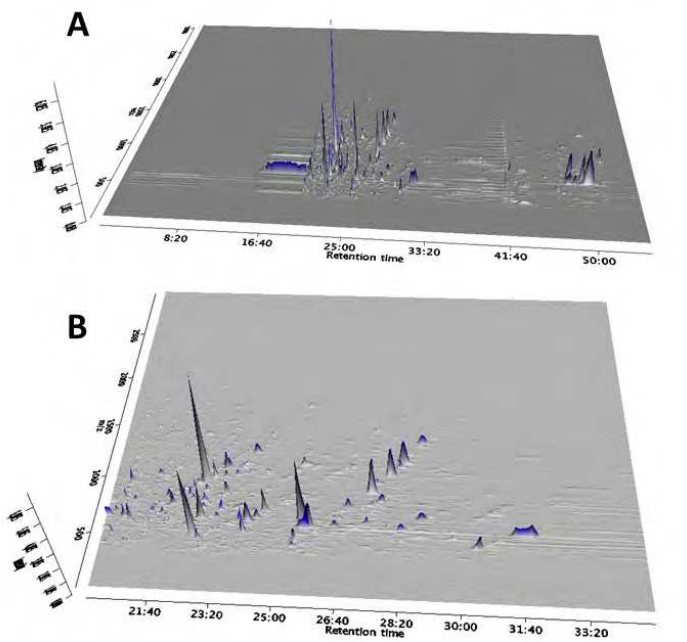



Fig. 3. A 3D representation of a mass spectrum of a liquid chromatographic separation in LC-MS typical analysis of proteomics and metabolomics data.

(A) Urinary native peptides without noise filtration and (B) with noise filtration are shown on a smaller time scale (20 to 34 minutes). These spectra were obtained using MZmine 2.1 with raw data converted first to the mzxml format (Pluskal et al., 2010).

position. It is not rare for the n x p matrix used in genetics data to have for dimension n=500000 SNP positions recorded for p=1000 individuals grouped according to ethnical, geographical or disease status. SNP positions, sequence, type and frequencies are maintained and accessible on different websites such as dbSNP (www.ncbi .nlm.nih.gov/projects/SNP), the International HapMap project (hapmap.ncbi.nlm.nih.gov), the SNP consortium (snp.cshl.org), the Human Gene Mutation Database or HGMD (www.hgmd.org), the 1000 Genomes project (www.1000genomes.org), the Pharmacogenomics database PharmGKB (www.pharmgkb.org) and the genotype-phenotype association database. GWAS Central (www.gwascentral.org). This information is particularly relevant in order to attempt SNP associations to disease status or health conditions. In recent human genetic studies, genotype data have been harvested, consisting in collecting for a few thousand human samples of different classes (ethnic groups, disease status groups, and so on) all the SNP profiles for particular genes (or even better all the genome). Algorithms such as EIGENSOFT suite is used to find statistically acceptable genotype-phenotype associations (Novembre & Stephens, 2008; Reich et al, 2008). The suite contains the EIGENSTRAT tool which is able to detect and correct for population bias of allele frequency, also called stratification, and suggests where the maximum variability resides among the population. PCA was demonstrated as a. valuable tool for detecting population substructure and correcting for stratification representing allele frequency

differences originating from ancestry between the considered population before associating SNPs profile and disease status (Price et al., 2006). These studies were recently published for making qualified inferences about human migration and history.

### 1.2.2 Transcriptomics data formats

In order to analyze in parallel the large population of mRNAs or transcriptomes that a cell is expressing, a high-throughput screening method called DNA microarrays is used today. These DNA chips, some of which are commercially available (ex: Affymetrix), contain imprinted on their glass surface, as individualized spots,. thousands of short nucleic acid sequences specific of genes and organized in matrices to facilitate their location. (Figure 2) Pangenomic DNA chips contain sequences representing ALL the genes known today for a certain species (a few tens of thousands). These chips are hybridized with equal quantity of mRNA or complementary DNA copies of the mRNA prepared from control and treated samples, including fluorescent red (treated) and green (control) nucleotide analogs, in order to keep track of the sample origin. After subsequent washing steps, green and red fluorescence signals present on the chip are measured and the red-to-green ratio is calculated for each gene. The colors of the spots are from red (treated) to green (control) indicating over- and under- abundance of gene expression in the treated condition. A yellow color indicates an equal abundance of gene expression (no effect of condition) and a black spot indicates absence of gene expression in both conditions. Major free-access transcriptomics databases are the Stanford microarray database (smd.stanford.edu) and the NCBI GEO omnibus (www.ncbi.nlm.nih.gov/geo). The size of these arrays depends on the gene population under study. It is not rare to study transcriptomics on n= 7000 genes (yeast) or more on pangenomic arrays n = 20000 – 30000 (Arabidopsis, humans, mice …). The number of DNA microarrays p is generally of the order of a few tens to a few hundreds, taking into account experimental replicates.

Alternative techniques exist to study gene expression, but they are not applied on a large- or genomic-wide scale as DNA microarrays,. and they are used in order to confirm hypotheses given by these later experiments. Among them, the technique using qRT-PCR or quantitative Reverse Transcription Polymerise Chain Reaction or its semi-high throughput variant called microfluidic cards (AppliedBiosystems) allow to quantify gene expression focused on 384 selected genes in one sample.

### 1.2.3 Proteomics and metabolomics data formats

Numerous mass spectrometry technologies are available today to perform proteomics and metabolomics analyses in specialized laboratories. These «omics» have not yet attained the mature status and the standardization level that transcriptomics has now attained, particularly at the level of data acquisition, data storage and sharing, as well as data analysis. However, some consortia, such as the human proteomic organization HUPO (www.hupo.org) and PeptidesAtlas (www.peptidesatlas.org), are spending a considerable amount of efforts and money to find standardization rules. One of the main difficulties in working with these «omics» data resides in maintaining intra- and inter-laboratory reproducibility. The second difficulty is that few mass spectrometers associated with the chromatographic separation devices are able to record a quantitative signal that is directly

proportional to analyte abundance. Semi quantitative data are generally obtained with matrix-assisted laser desorption ionization (MALDI) and quantitative signal is better obtained with electrospray ionization (ESI) methods. The use of relevant working strategies is necessary to lower technical and analytical variabilities, and this is also accomplished through the use of numerous replicates and internal standards with known or predictive mass spectrometry behaviors. The third difficulty is inherent to the commercially available instruments for which data acquisition and processing use computational tools and proprietary data formats. There are, however, a few format converters that are accessible, among them OpenMS (open-ms.sourceforge.net) and TransProteomicPipeline (tools.proteomecenter.org). These techniques are used extensively with the aim of detecting and quantifying biomarkers or molecular signals. specific to drug toxicity, disease status and progression, sample classification, and metabolite pathways analysis. The size of proteomics and metabolomics matrices depends on the accuracy level measured on the analytes mass and the range of mass under study. n varies from a. few hundreds to a few tens of thousands of m/z analytes and the p dimension of experiments or samples is largely dependent on the biological question. (a few tens).

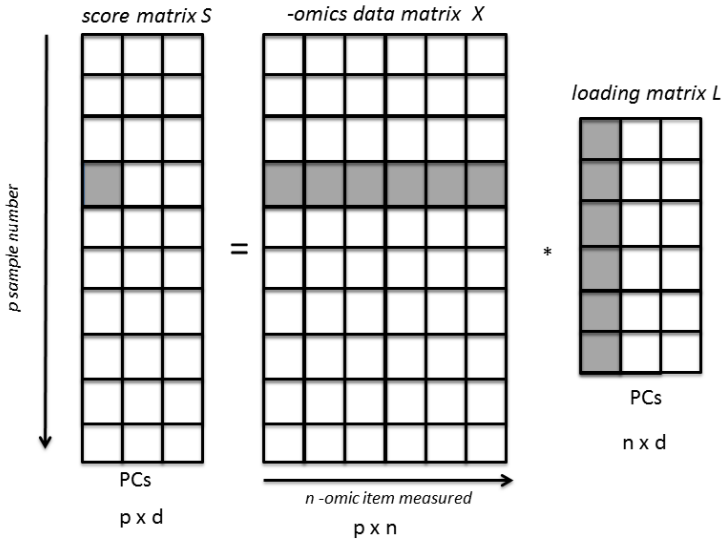
Alternative techniques to confirm proteomic expression use protein chips and immunoprecipitation techniques that are antibody dependent. Another mass spectrometry technique, which is Liquid Chromatography coupled to Selected Reaction Monitoring (LC-SRM), is also used to confirm the level of expression of a particular analyte focusing on its physico-chemical properties as well as its chemical structure. In this case a high specificity and sensibility are generally obtained because the mass spectrometer records the signal related to both occurrences of the presence of one analyte mass (precursor ion) and the presence of one of its specific fragments obtained by collision-induced dissociation (CID).

## 2. Exploratory data analysis with PCA of «omics» data

### 2.1 The principles of PCA in «omics»

The original. n x p matrix (Figure 1) or its transposed p x n (Figure 4) contains raw data with n generally much larger than p. The data should be preprocessed carefully. according to the nature and accuracy of the data. In the «omics», particularly proteomics and metabolomics, autoscaling and Pareto normalizations are the most used (Van der Berg et al, 2006). Autoscaling is the process of rendering each variable of the data (the «omics» item) on the same scale with a mean of 0 and a standard deviation of 1. PCA consists in reducing the normalized n x p matrix to two smaller matrices, an S score matrix and an L loading matrix. The product of scores S and the transposed loadings L′ matrix plus a residual matrix R gives the original n x p matrix X according to the formula X = S*L′ + R. PCs are the dimension (d) kept for S and L matrices and numbered $PC_1$, $PC_2$, $PC_3$…. according to the largest variance they capture. $PC_1$ captures most of the variability in the data followed by $PC_2$ and $PC_3$ (Figure 4). PCA helped originally to detect outliers. PCA axis capture the largest amount of variability in the data that scientists in the «omics» fields want to interpret and to relate to biological, environmental, demographic and technological factors (variability in replicates). Therefore variance in the higher PCs is often due to experimental noise, so plotting data on the first two to three PCs not only simplifies interpretation of the data but also reduces the noise. The scoring plot displays the relationship existing between samples, meaning that two similar samples will be distantly close in the PC space. Furthermore, the loading plot

displays the relationship between the «omics» items (genes, peptides, proteins, metabolites, SNPs). A strong correlation between items will be expressed by a linear arrangement of these points in the loading matrix. Moreover PCA biplots representing score and loading scatter plots superposed together are useful to detect the importance of particular loadings («omics» measured items) responsible for separating these sample clusters.




$$s_{4,1} = x_{4,1}.l_{1,1} + x_{4,2}.l_{2,1} + x_{4,3}.l_{3,1} + x_{4,4}.l_{4,1} + x_{4,5}.l_{5,1} + x_{4,6}.l_{6,1}$$

Fig. 4. Relationship between the X «omics» data matrix X, the S score matrix and the L loading matrix in principal component analysis.

Here the original «omics» matrix has been transposed for convenience, with n=6 being the number of omic items experimentally measured and p = 10 being the number of samples considered. d=3 is the number of PCs retained in the model, facilitating its graphical exploration. The highlighted row in X and column in L show what is required to generate a $PC_1$ score for sample 4.

PCA could reveal main patterns in the data but can detect some systematic non-biologically related or unwanted biologically related bias defined as batch effects. The existence of batch effects in «omics» data is being more and more recognized to frequently misguide biological interpretations. A large number of softwares can calculate these S and L matrices for large data sets. A PLS toolbox from Eigenvector research (www.eigenvector.com) running under MATLAB (www.matworks.com) contains representative 2D or 3D graphics of PCs space. Moreover, statistical indexes such as Q residues allow to estimate the disagreement behavior of some variables (samples) in the model, and Hotelling's T2 indexes measures the multivariate distance of each observation from the center of the dataset. R (cran.r-project.org) contains, in the statistical package included in the basic R installation, the prompt() function, which performs PCA on the command line. Particular «omics» -specific softwares containing sophisticated normalization, statistics and graphic options for proteomics and metabolomics data are available, such as DAnteR (Poplitiya et al., 2009) and MarkerView (www.absciex.com).

### 2.2 Interactive graphic exploratory data analysis with Ggobi and rggobi

Scatter plots are still the simplest and most effective forms of exploratory analyses of data but are limited to a pairwise comparison with just two samples in any one scatterplot diagram. Ggobi (www.ggobi.org) and rggobi, an alternative of ggobi with R GUI interface (www.ggobi.org/rggobi), are free tools that allow doing a scatter plot matrix with some limitation, as they graphically display a small number of explicative variables (less than 10). Ggobi and rggobi have an effective way of reducing a large multivariate data matrix into a simpler matrix with a much smaller number of variables called principal component or PCs without losing important information within the data. Moreover, this PC space is graphically displayed dynamically as a Grand Tour or 2D tour. Moreover, samples can be specifically colored or glyphed by using a "brushing" tool according to their belonging to some factors or categorical explicative variables (patient status, sample group, and so on…). Moreover, unavailable measurements (NA) are managed by Ggobi by using simple value replacements (mean, median, random) as well as sophisticated multivariate distribution modeling (Cook & Swayne, 2007).

### 2.3 PCA for «omics» data

#### 2.3.1 PCA for genetics data

PCA was used almost 30 years ago by Cavalli-Sforza L.L in population genetics studies to produce maps summarizing human genetic variation across geographic regions (Menozzi et al., 1978). PCA is used also in genotype-phenotype association studies in order to reveal language, ethnic or disease status patterns (Figure 5). Recently it has been shown that these studies are difficult to model with PCA alone because of the existence of numerous unmeasured variables having strong effects on the observed patterns (Reich et al., 2008; Novembre & Stephens, 2008). When analyzing spatial data in particular, PCA produces highly structured results relating to sinusoidal functions of increasing frequency with PC numbers and are sensitive to population structure, including distribution of sampling locations . This observation has also been seen in climatology. However PCA can reveal

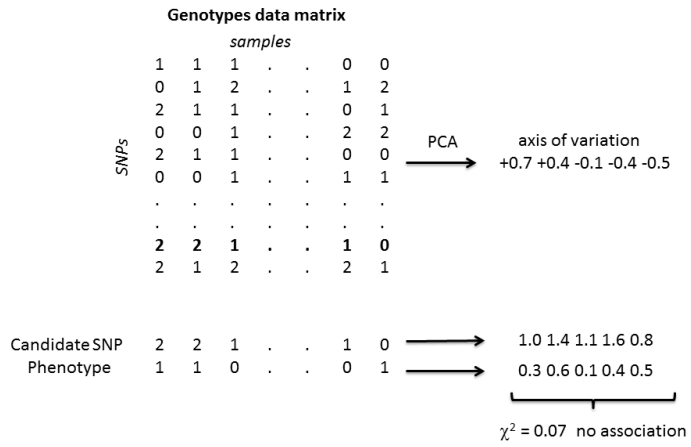



Fig. 5. The EIGENSTRAT algorithm from the EIGENSOFT suite.

some patterns on the data, and reliable predictions require further genetic analysis and integration with other sources of information from archeology, anthropology, epidemiology, linguistic and geography (François et al., 2010).

Genotype data consists of a n x p matrix where p individuals are recorded for their n SNPs. in their genomes. PCA is applied to infer continuous axes of genetic variation. A single axis of variation is indicated here. A genotype at a candidate SNP and phenotype are continuously adjusted by amounts attributable to ancestry along each axis. A $\chi^2$ show here no significant association for this particular SNP (Price et al., 2006).

### 2.3.2 PCA for transcriptomics data

Gene expression array technology has reached the stage of being routinely used to study clinical samples in search of diagnostic and prognostic biomarkers. Due to the nature of array experiments, the number of "null-hypotheses" to test, one for each gene, can be huge (a few tens of thousands). Multiple testing corrections are often necessary in order to. screen non-informative genes and reduce the number of null-hypotheses. One of the commonly used methods for multiple testing control is to calculate the false discovery rate (FDR) which is the ratio of the number of false rejections among the total number of rejections. FDR adjustment on raw p-values is effective in controlling false positives but is known to reduce the ability to detect true differentially expressed genes.

In transcriptomics studies, PCA is often used for the location of genes relative to each other in a reduced experiment space. Genes are plotted with respect to the two orthogonal linear combinations of experiments that contain the most variance (Lu et al., 2011). Transcriptomics also use other multivariate tools for classification and clustering (Tibshirani et al., 2002). A very fast and effective classification strategy is linear discriminant analysis. In classification problems there are positive training examples that are known members of the class under. study and negative training examples that are examples known not to be members of the class. The test examples are compared to both sets of training examples, and the determination of which set is most similar to the test case is established. In this process the test example is "classified" based on training examples. Clustering is a commonly used categorizing technique in many scientific areas using K-means grouping technique. Using this approach the user can cluster data based on some specified metric into a given number of clusters. Users can cluster arrays or genes as desired into a pre-specified number of clusters. The algorithm has a randomized starting point so results may vary from run to run.

### 2.3.3 PCA for proteomic and peptidomic data

#### 2.3.3.1 Urinary peptides and biomarker discovery study

PCA was used in order to distinguish urine samples containing or not pseudo or artificial spiked-in analytes or pseudo biomarkers (Benkali et al., 2008). The objectives were to analyze variations in the data and distinguish their sources. These variations could arise from (a) experimental variations due to changes in the instrument or experimental conditions, (b) real variations but of no interest in the primary objective, such as male versus female subjects, drug treatments, metabolites of a therapeutic agent… and (c) relevant

differences that reflect changes in the system under study (spiked-in or not spiked-in). The experiment consisted in using human urines from 20 healthy volunteers splitted in two groups of ten, one which was spiked-in with few synthetic peptides at a certain variable concentration and the other without. Urines were processed using the same peptide extraction solid phase extraction (SPE) protocol, by the same experimentalist, and peptide compositions were recorded by off-line nanoLC-MS MALDI TOF/TOF. Data were processed with MarkerView software version 1.2 (www.absciex.com). PCA preprocessing consisted in using Pareto scaling without weighing and no autoscaling because Pareto scaling is known to reduce but not completely eliminate the significance of intensity, which is appropriate for MS because larger peaks are generally more reliable and all variables are equivalent. Different scaling methods are worth trying because they can reveal different features in the data with peak finding options and Pareto normalization (Van der Berg et al., 2006).

More than 5000 features (or m/z analytes) were retained from which respective abundances were observed. The n x p matrix contains n= 5000 and p = 20 samples. Scores and loading on PCs were calculated with 3 PCs capturing 80.4% of total data variability. Figure 6 shows $PC_1$ (70.6%) versus $PC_2$ (7.4%) (Figure 6A), as well as. $PC_1$ (70.6%) versus $PC_3$ (2.4%) (not shown). Sample points in the scoring scatterplot were colored according to. their group assignment before analysis (unsupervised). PCs scores on the $PC_1$-$PC_2$ projection axis allowed us to define the A9 sample as an outlier behaving as an unspiked B group sample (labeling tube error perhaps). We had to discard this sample for the rest of the analysis. This analysis was carried out on samples blinded to categorical label (spiked and unspiked) and the coloring of samples on the graphic was only carried out after the PCA. Spiked samples (A samples) are in red color and unspiked samples in blue color (B samples). The high positive value of loadings (green points) in the $PC_1$ and $PC_2$ axes are associated with features (or m/z analytes) most responsible to discriminate the two sample groups. The relative abundance of the spiked analyte of m/z = 1296.69 and its two $^{13}C$. isotopically stable labeled variants, 1297.69 and 1298.69, is shown in the spiked group (A samples, red points) and in the unspiked group (B samples, blue points). Moreover they tend to lie close to straight lines that pass through the origin in the loading plots (Figure 6A). These points (green points) are correlated because they are all the isotopic forms of the same spiked compound. The same is observed for other spiked analytes (904.46, 1570.67, 2098.09 and 2465.21). Finally superposed mass spectra from 20 samples of both groups show the relative abundance of analytes (panel B and insert focused on m/z = 1296.69 analyte and its natural $^{13}C$ isotopes).

Variables with large positive $PC_1$ loadings are mainly in group A (spiked samples) and absent or at lower intensities in group B. $PC_1$ separates both groups but $PC_2$ seems to separate both groups A and B in two half-groups (Figure 6A). What is the nature of the variation captured by $PC_2$ where some loadings (900.40, 1083.54 and 1299.65) give high positive $PC_2$ values and negative $PC_1$ values ?. The examination of Figure 7 shows that these analytes show a progessive increase in their intensity with a gradient following the order of their analysis in the instrument. The data were acquired in the order they are displayed (from left to right) and group A members were acquired before members of group B, which introduces a bias or a batch effect in these data. To avoid this effect, the samples should be acquired in a random order,. with group members mixed.

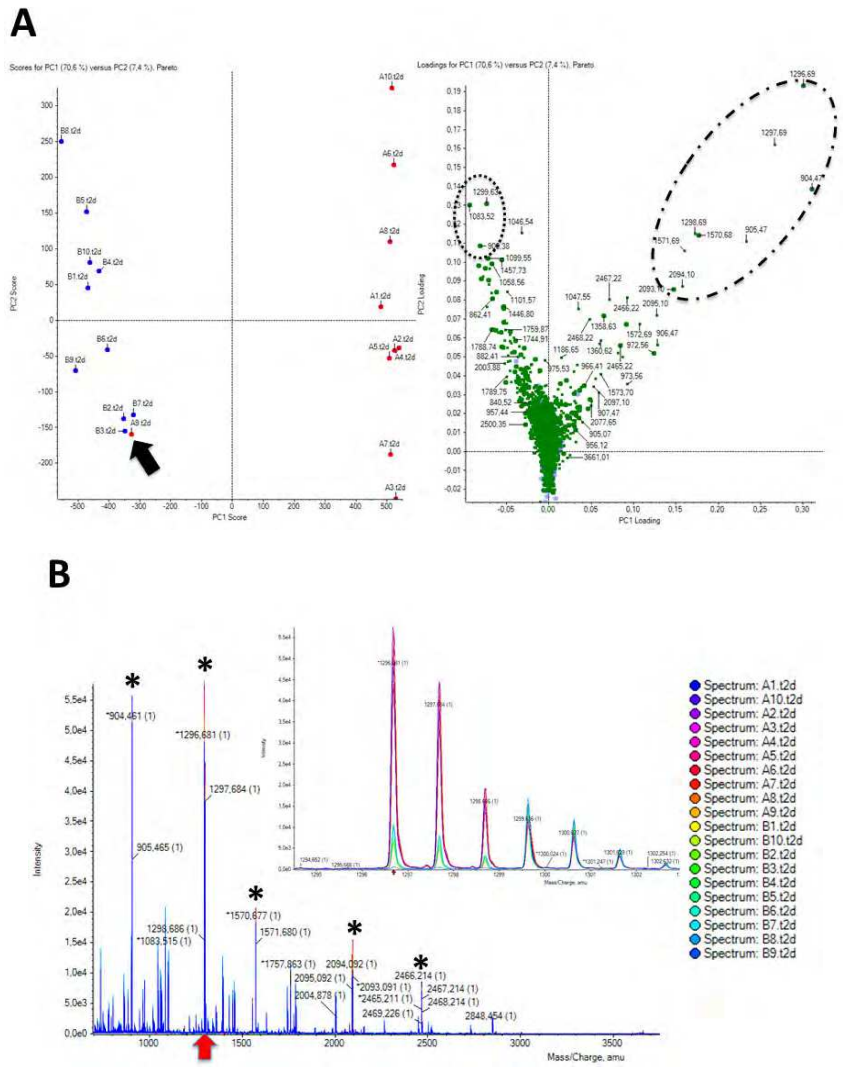



Fig. 6. PCA of 20 urine samples spiked or not spiked with synthetic peptides (pseudo biomarkers).

(A) Scores and loadings plots of $PC_1$ and $PC_2$ axes show a good separation of group A (spiked,. colored in red) from group B (not spiked, colored in blue). The A9 sample (black arrow) is an outlier and behaves as a B member group. It should be removed from the analysis. Loading plots show the 5000 analytes (green points) from which the majority are not contributing to the variability (0,0). Some analytes contribute to the large positive variation in the $PC_1$ axis (spiked peptides) and to the positive $PC_2$ (bias effect). (B) Superposition of the 20 spectra of urine samples after their alignment. with. a symbol (*) indicating the m/z of spiked peptides. The insert corresponds to the enlargement of the spectra located at the red arrow in the spectra, showing the abundance of. the 1296.69. and their $^{13}C$ isotopes among the 20 samples, particularly in the A group.

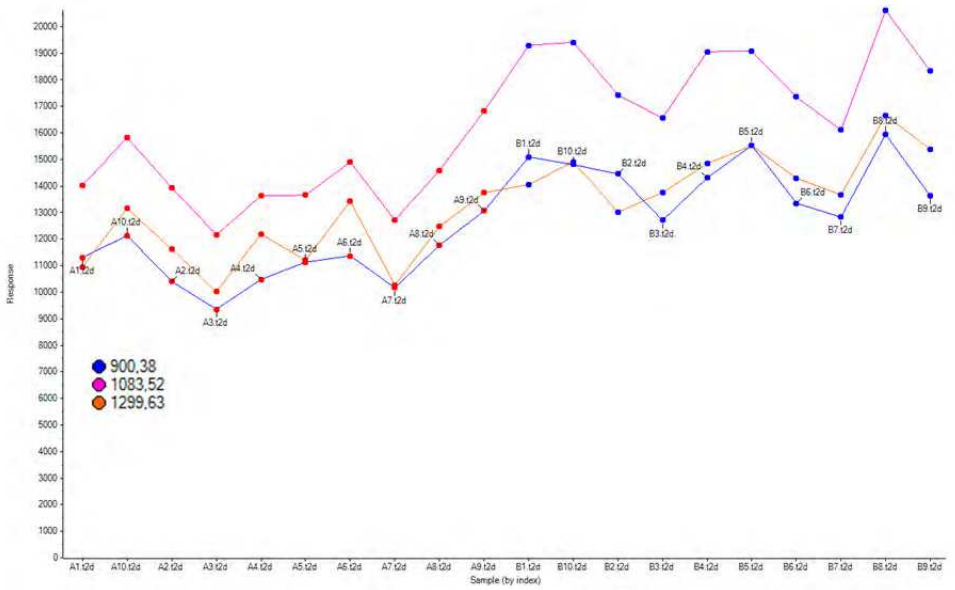



Fig. 7. A variation. recognized in the $PC_2$ axis probably due to a batch effect in the sample processing. Analytes. 900.38, 1083.52 and 1299.63. responsible for the positive value of scores in the $PC_2$ axis (Figure 6, panel A) see their intensity increase slightly during their acquiring time (from left to right), signaling a probable batch effect.

#### 2.3.3.2 Microbial identification and classification by MALDI-TOF MS and PCA

MALDI-TOF MS mass spectrometry has recently been used as a technique to record. abundance of proteins extracted from different phyla of bacteria with the aim of finding phylum-specific patterns and use them to classify or recognize these bacteria in a minimum culture time. Sauer, S. has pioneered the technique of rapid extraction of proteins from alcohol, strong acid treatment or direct transfer from single colonies of bacteria, with or without the need to cultivate them (Freiwald & Sauer, 2009). Ethanol/Formic acid extraction of proteins of two clones of each. 6 bacteria strains, *Klebsiela pneumonia* (KP), *Acinetobacter baumanii*, (AB), *Lactobacillus plantarum* (LP), *Pseudomonas aeruginosa* (PA), *Escherichia coli* MG (MG), *Bacillus subtillis* (BS) were prepared. Mass spectra of proteins were recorded in five analytical replicates in the range 4000 to 12000 Daltons (Figure 8A). Major extracted proteins come from abundant ribosomal proteins. The natural variants in their amino acid sequence are responsible for the differences of masses in the peaks observed in the spectra, and their abundance is characteristic of the bacteria. Moreover, 6 bacteria clones (X1 to X6) were blindly analyzed in triplicate. PCA was used in order to distinguish the axes of greater variability in the data.

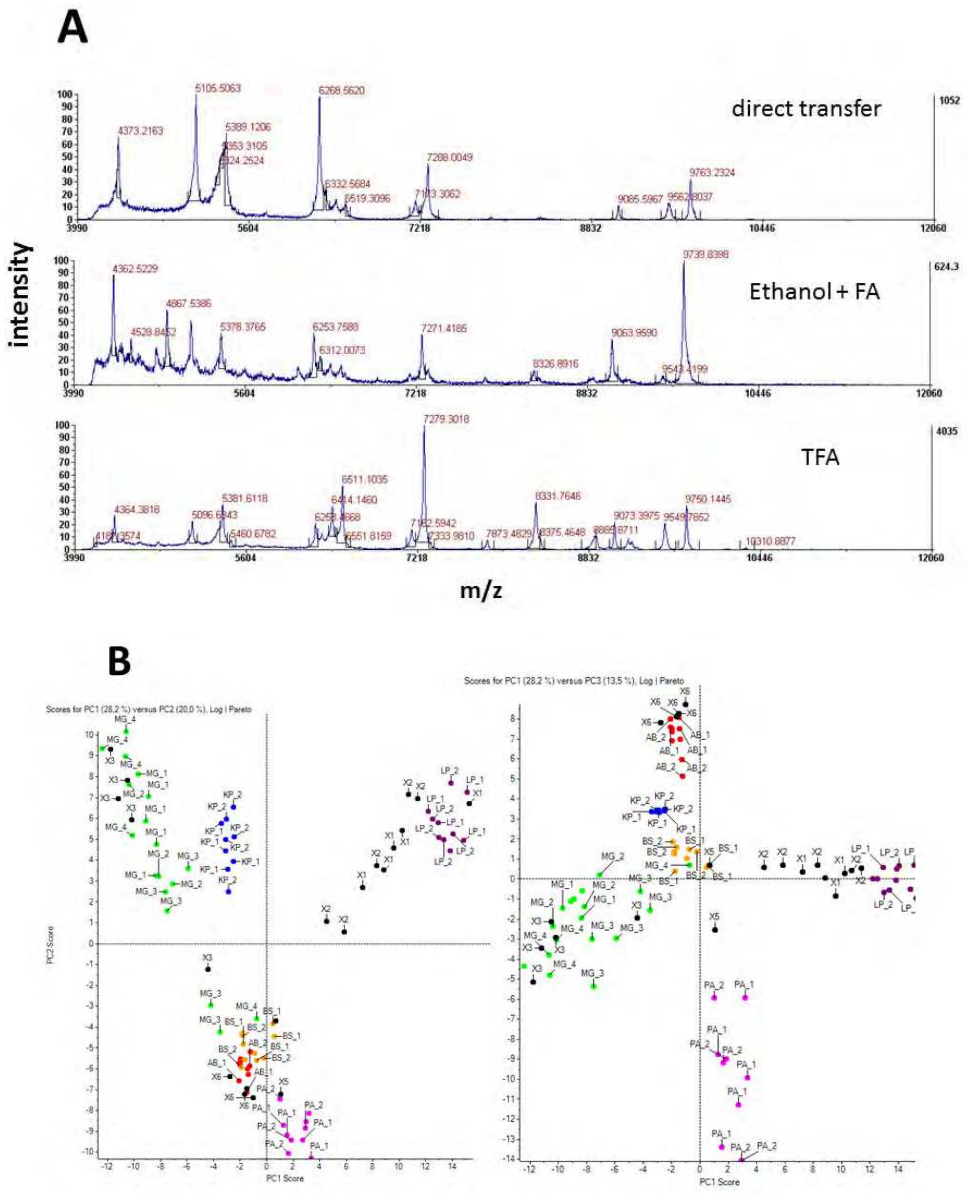



Fig. 8. MALDI TOF MS spectra of protein extracts of 6 bacteria strains and PCA.

(A) MS spectra of *Escherichia coli* MG clone proteins extracted with 3 different methods, direct transfer, ethanol/Formic acid (FA). and Trifluoroacetic acid(TFA). Peak abundances. may vary according to the extraction methods, from bacteria strains and from analytical variability. (B) Scores of $PC_1$ versus $PC_2$ and $PC_1$ versus $PC_3$ of the full 6 bacteria dataset,. including unknown X1 to X6 samples. (black points). (unpublished Maynard & Gastinel, 2009).

For this analysis, $PC_1$ takes 28.2% of variability, $PC_2$ 20% and $PC_3$ 15%, for a total of 63.2% captured in the model (Figure 8B). The $PC_1$ axis separates LP and PA bacteria strain. from. KP, MG, BS and AB strains. The $PC_2$ and $PC_3$ axes separate. KP and LP from BS , AB and PA. but not so well for the MG strain. Moreover, X samples are 100% separated in their correct respective clusters. The broad distribution of samples in the score plot is probably due to the relatively poor analytical reproducibility in the ionization of. MALDI TOF MS analysis. From the loading score plots (not shown) few proteins of particular m/z responsible for this bacteria strain separation have been recognized in protein databases by their annotation. Among them, the 5169 Daltons protein is attributed to the 50S ribosomal protein annotated as L34 of *Acinetobater baumanii*.

### 2.3.4 PCA for metabolomics data

Humic acids are one of the major chemical components of humic substances, which are the major organic constituents of soil (humus), peat, coal, many upland streams, dystrophic lakes, and ocean water. They are produced by biodegradation of dead organic matter. They are not a single acid; rather they are a complex mixture of many different acids containing carboxyl and phenolate groups so that the mixture behaves functionally as a dibasic acid or, occasionally, as a tribasic acid. Humic acids can form complexes with ions that are commonly found in the environment, creating humic colloids. Humic and fulvic acids (fulvic acids are humic acids of lower molecular weight and higher oxygen content than other humic acids) are commonly used as a soil supplement in agriculture, and less commonly as a human nutritional supplement. Humic and fulvic acids are considered as soil bioindicators and reflect an equilibrium between living organic and non-organic matters.

Mass spectrometry has been used to estimate signature analytes and patterns specific to some soils (Mugo & Bottaro, 2004). Fulvic acids were prepared from a soil using different extraction protocols resulting in 5 samples,. H1, H1H2, EVM1, EVM2 and EAA. Are these extraction protocols similar and which analytes are they extracting more efficiently? MALDI MS spectra from 150 to 1500 m/z range were recorded in the presence of the MALDI matrix alpha-cyano-4-hydroxycinammic acid (CHCA). Normalization of intensities were done with the 379 m/z analyte in common to these samples, and Pareto scaling was chosen during the alignment process performed by MarkerView. Figure 9A shows that the PCA analysis reveals poor separation of samples with $PC_1$ explaining 25.8%, $PC_2$ 18.7% and $PC_3$ 14.8%. of variability (a total of 59.3% captured). Samples are not so well separated by the first PC axis, demonstrating the large influence of factors other than soil extraction differences (chemical precipitation, physical precipitation, filtration). Discriminant Analysis associated with PCA (supervised PCA-DA) was attempted to further separate these known 5 groups (Figure 9B). This supervised technique means that it uses class information based on the assigned sample group to improve their separation. Figure 9B shows a dramatic improved separation but this may be based on noise. Peaks which are randomly more intense in one group as compared to another can possibly influence the results, and careful examination of loading plots as well as analyte profiles across the samples is necessary to avoid batch effects. This analysis is also affected by samples incorrectly assigned to wrong group and outliers.

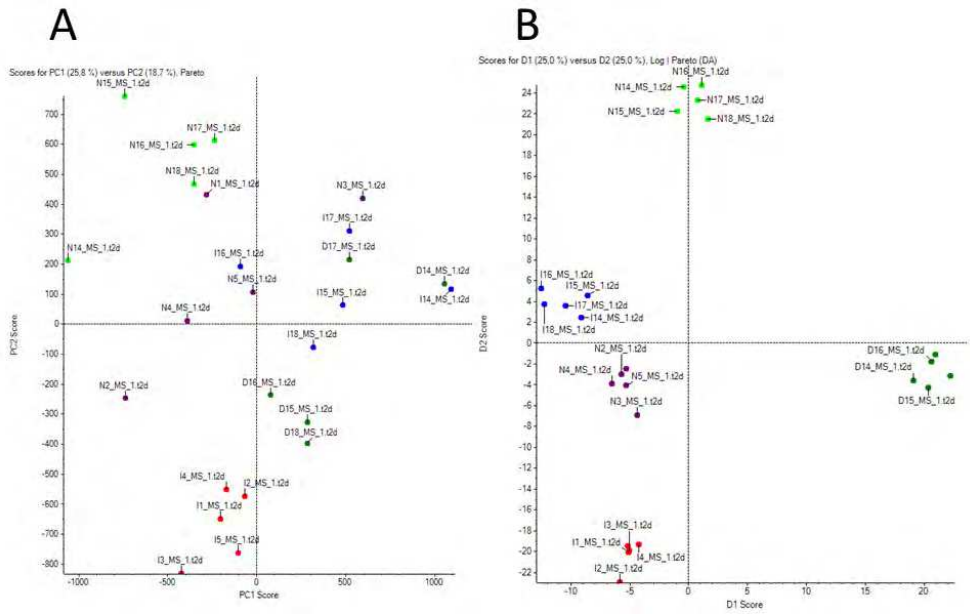



Fig. 9. Mass spectrometric study of fulvic acids profiles in the range 150 to 1000 m/z present in 5 sample preparations and analyzed in five analytical replicates by PCA.

(A) Unsupervised PCA and (B) PCA-DA or supervised PCA with group information included before reducing data. Only the first and second PC axes are shown for the score plots. (unpublished Basly & Gastinel, 2009).

In Figure 9, PCA (A) and PCA-DA (B) show a different pattern of the variability in the dataset. $PC_1$. axis does not discriminate H1H2 (green) , EVM1 (blue) and EAA (violet) as for. the $PC_2$ axis. The $PC_2$ axis discriminates however EVM2 (pale green). and H1 (red). Keeping these 5 different protocols as 5 different classes, PCA-DA (Figure 9B) however discriminates quite well these these 5 preparation analyses in quintuplicates. EVM1 and EAA are still the closest group. Loading score plots reveal which analytes are the most favored in a particular extraction method relative to another.

## 3. How can PCA help to reveal batch effects in «omics» data?

### 3.1 What are batch effects?

Batch effect is one overlooked complication with «omics» studies and occurs because high-throughput measurements are affected by multiple factors other than the primary tested biological conditions (Leek et al, 2010; Leek & Storey, 2008). These factors are included in a comprehensive list among which are laboratory conditions, reagents batches, highly trained personnel differences, and hardware maintenance. Batch effect becomes a problem when these conditions vary during the course of an experiment, and it becomes a major problem when the various batch effects are possibly correlated with an outcome of interest and lead to incorrect conclusions (Ransohoff, 2005; Baggerly, et al., 2004). Batch effects are defined as

a sub-group of measurements that have qualitatively different behaviors across conditions and are primarily unrelated to the biological or scientific variables under study. Typical batch effect is seen when all samples of a certain group are measured first, and when all samples of a second group are measured next. Batch effect occurs too when a particular batch of reagent (ex: Taq polymerase enzyme for PCR experiments) is used with all samples of the first group, and another reagent batch is used with all samples of the second group. Typical batch effects are also seen when an experimentalist/technician acquires all samples from the first group and a different experimentalist/technician works with the other group or when the instrument's characteristics (example for MALDI mass spectrometry: laser or detector replacements) used to acquire the data have been deeply modified. Data normalization generally does not remove batch effect unless normalization takes into account the study design or takes into account the existence of a batch problem.

### 3.2 How to find evidence of batch effects

The first step in addressing batch and other technical effects is to develop a thorough and meticulous study plan. Studies with experiments that run over long periods of time, and large-scale, inter-laboratory experiments, are highly susceptible to batch effects. Intra-laboratory experiments spanning several days and several personnel changes are also susceptible to batch effects. Steps necessary to analyze batch effects require different. levels of analysis, according to the recent review of Leek J.T (Leek et al., 2010). What follows are some of the recommended actions:. Performing a hierarchical clustering of samples that assigns a label to the biological variables and to the batch surrogates estimates, such as laboratory and processing time; plotting individual features (gene expression, peptides or metabolites abundances). versus biological variables and batch surrogates using ggobi for example; calculating principal components of the high-throughput data and identifying components that correlate with batch surrogates. If some batch effects are present in the data, artifacts must be estimated directly, using surrogate variable analysis (SVA) (Leek et al., 2007). Recently, the EigenMS algorithm has been developed and. implemented within a pipeline of bioinformatic tools of DanteR in order to correct for technical batch effects in MS proteomics data analysis (Polpitya et al., 2008; Karpievitch et al., 2009). The algorithm uses an SVA approach to estimate systematic residual errors using singular value decomposition taking account primary biological factors and substracting those estimates from raw data in the pre-processing data analysis. The estimated/surrogate variables should be treated as standard covariates in subsequent analyses or adjusted for use with tools such as Combat (Johnson & Li, 2007). After adjustments that. include surrogate variables (at least processing time and date), the data must be reclustered to ensure that the clusters are not still driven by batch effects.

### 3.3 How to avoid batch effects

Measures and steps must be taken to minimize the probability of confusion between biological and batch effects. High-throughput experiments should be designed to distribute batches and other potential sources of experimental variation across biological groups. PCA of the high-throughput data allows the identification of components that correlate with batch surrogate variables.

Another approach to avoid and prevent batch effects is to record all parameters that are important for the acquisition of the measures and the relevant information related to demographic and grouping factors. The structure of a database under MySQL with attractive web graphic user interface (GUI) should be conceived at the same time as the study design is defined. Such a database was constructed for a mass spectrometry based biomarker discovery project in kidney transplantation in a French national multicenter project. The BiomarkerMSdb database structure contains 6 linked tables: Demographic data, Peptide Extraction Step, Liquid Chromatography Separation, Probot Fractionation, Spectrometry Acquisition and Data Processing. Figure 10 shows the details of the form that the user must complete to record demographic data of patients enrolled in this project. This approach, with an internet interface used to facilitate data exchange between laboratories enrolled in the project, allows to keep track of essential parameters that could interfere with future interpretations of the results. At minimum, analyses should report the processing group, the analysis time of all samples in the study, the personnel involved, along with the biological variables of interest, so that the results can be verified independently. This is called data traceability.

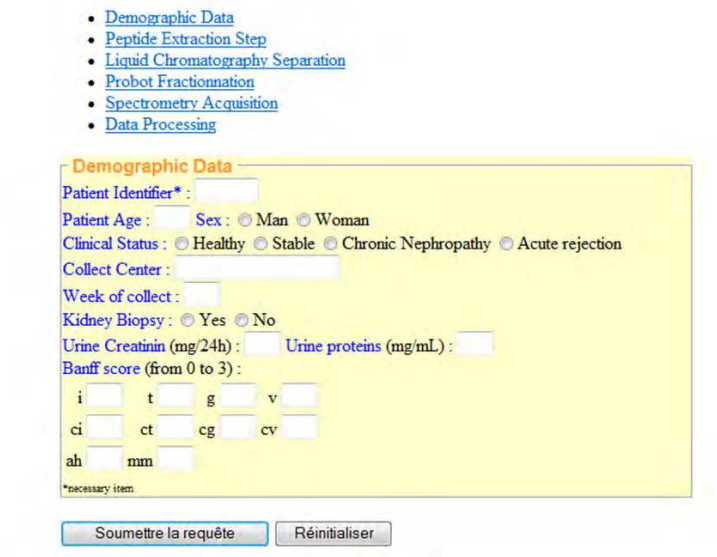



Fig. 10. Extract of the internet form used to interact with BiomarkerMSdb, a relational database under MySQL constructed to record essential parameters involved in a biomarker discovery project using LC-MS strategies. This form is used to fill one of the 6 tables of the database called. "Demographic Data"(unpublished Moulinas & Gastinel, 2011).

## 4. Conclusion and perspectives

Observational and experimental biology is confronted today with a huge stream of data or "omics" data acquired by sophisticated and largely automated machines. This data is recorded under a digital format that has to be stored safely and shared within the scientific community. One of the challenges of modern biologists is to extract relevant knowledge

from this large data. For that purpose, biologists not only should be acquainted with the use of existing multivariate and statistical data mining tools. generally used by meteorologists, economists, publicists and physicists but also should conceive their own specific tools. Among the tools available for multivariate analysis of «omics» data, Principal Component Analysis (PCA) as well as PLS and PCA-DA derivatives have demonstrated their utility in visualizing patterns in the data. These patterns consist sometimes in detecting outliers that spoiled data and that could be removed from them. PCA quantify major sources of variability in the data and allows to show which variables are most responsible for the relationship detected between the samples under study. Moreover, unwanted sources of variability can be revealed as batch effects and partially corrected by surrogate variable analysis (SVA) and EigenMS approaches. However, there are some limitations to using PCA in "omics" data. These limitations result from the large choice of methods of the data pre- and post-processing and the technical difficulty in displaying graphically all the data. Ggobi and rggobi allow to display quite large data using Grand tour and 2D tour, showing dynamic projections with the ability to color and glyph points according to factors (brushing). PCA is an invaluable tool in the preliminary exploration of the data and in filtering or screening them according to noise, outliers and batch effects before using other multivariate tools such as classification and clustering. An appropriate educational program should be pursued in universities in order to expose the theory and praticability of these tools to future biologists.

## 5. References

Baggerly, K.A., Edmonson, S.R., Morris, J.S. & Coombes, K.R.(2004) High-resolution serum proteomic patterns for ovarian cancer detection. *Endocrine-Related Cancer*, 11, 583-584.

Basly, J.C & Gastinel, L. (2009) Influences of extraction protocols in fulvic acids preparations from soils. *Master 2 Training Report*, GREASE Department, University of Limoges, France.

Benkali, K., Marquet,P.,. Rerolle, J., LeMeur,Y. & Gastinel, L. (2008) A new strategy for faster urine biomarker identification by nano-LC-MALDI-TOF/TOF mass spectrometry. *BMC Genomics*, 14, 9, 541-549.

Cavalli-Sforza,L.L., Menozzi,P. & Piazza,A. (1994) The History of Geography of Human Genes, Princeton University Press, ISBN: 9780691087504, NJ, USA.

Cook, D & Swayne, DF (2007) Interactive and dynamic graphics for data analysis with R and ggobi. Springer, ISBN: 978-0-387-71761-6, NJ, USA.

DeRisi,J.L., Vishwanath, R.I & Brown, P.O. (1997) Exploring the Metabolic and Genetic Control of the Gene Expression on a Genomic Scale. *Science*, 278, 681 – 686.

François, O., Currat,M., Ray, N., Han, E., Excoffier, L.& Novembre, J. (2010) Principal component analysis under population genetic models of range expansion and admixture. *Molecular Biological Evolution*, 27, 6, 1257-1268.

Freiwald, A. & Sauer, S (2009) Phylogenetic classification and identification of bacteria bymass spectrometry", *Nature protocols*, 4, 5, 732-742.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome, *Nature*, 409,860-921.

Johnson, W.E. & Li, C (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8,1,118-127.

Karpievitch, Y.V., Taverner, T., Adkins, J.N., Callister, S.J., Anderson, G.A., Smith, R.D. & Dabney, A.R. (2009) Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. *Bioinformatics*, 25, 9, 2573-2580.

Leek,J.T., Scharpf, R.B., Corrada –Bravo, H., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K. & Irizarry,R.A (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Genetics*, 11, 733-739.

Leek,J.T. & Storey, D.J (2008) A general framework for multiple testing dependence. *Proceeding of the National Academy of Sciences*, 105, 48, 18718-18723.

Listgarten, J. & Emili, A. (2005) Statistical and Computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular and Cellular Proteomics*, 4.4, 419-434.

Lu, J., Kerns, R.T., Peddada, S.D. & Bushel, P.R. (2011) Principal component analysis-based filtering improves detection for Affymetrix gene expression. *Nucleic Acids Research,* 2011, 1-8.

Maynard,C. & Gastinel, L. (2010) Phylogenetic relationship between bacteria revealed by MALDI TOF mass spectrometry. *Life Science Undergraduate Training Report*, Microbiology Department, University of Limoges, France.

Menozzi, P., Piazza, A & Cavalli-Sforza,L (1978) Synthetic maps of human gene frequencies in Europeans. *Science*, 201,4358, 786-792.

Mugo, S.M. & Bottaro, C.S. (2004) Characterization of humic substances by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communication in. Mass Spectrometry*. 18,20, 2375-2382.

Moulinas, R & Gastinel,L (2011) How to detect batch effect in mass spectrometry analysis – Constitution of BiomarkerMSdb. *Master 1 SVT Report* , University of Limoges, France.

Novembre, J. and Stephens, M. (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40, 5, 646-649.

Pluskal, T., Castillo, S., Villar-Briones, A., & Oresic, M. (2010) MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC *Bioinformatics* 11:395-405.

Poplitiya,A.D., Qian, W-J., Jaity, N., Petyuk, V.A., Adkins, J.N., Camp II, D.G., Anderson, G.A. and Smith, R.D (2008) DAnTE: A statistical tool for quantitative analysis of –"omics" data. *Bioinformatics*, 24,13, 1556-1558.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38,8,904-908.

Ransohoff, D.F. (2005) Lessons from controversy: ovarian cancer screening and serum proteomics. *Journal of The National Cancer Institute*,97, 4, 315-317.

Rao,P.K. & Li, Q (2009) Principal component analysis of proteome dynamics in iron-starved mycobacterium tuberculosis. *J. Proteomics Bioinformatics*, 2,1, 19-31.

Reich, D., Price, A.L. and Patterson (2008) Principal component anaysis of genetic data. *Nature Genetics*, 40, 5, 491-492.

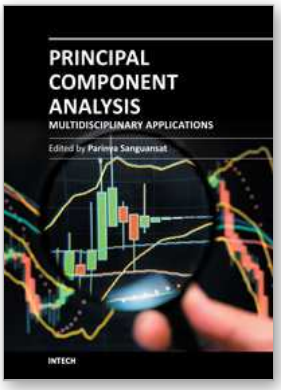
Sah, H.N. & Gharbia, S.E. (2010) *Mass Spectrometry for Microbial Proteomics*, Wiley,ISBN: 978-0-470-68199-2,UK.

Stein, L. (1996) How Perl saved the Human Genome Project, in *The Perl Journal,* 1,2, available from: www.foo.be/docs/tpj/issues/vol1_2/tpj0102-0001.html.

Tibshirani, R., Hastie, T., Narasimhan, B & Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proceedings of the National Academy of Sciences*, USA. 99, 10, 6567-6572.

Van der Berg, R., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K. & Ven der Werf, M.J(2006) Centering, scaling and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7, 142-156.

Venter, J.G. et al, (2001) The sequence of the Human Genome, *Science*, 291, 1305-1351.




**Principal Component Analysis - Multidisciplinary Applications**

Edited by Dr. Parinya Sanguansat

ISBN 978-953-51-0129-1
Hard cover, 212 pages
**Publisher** InTech
**Published online** 29, February, 2012
**Published in print edition** February, 2012

This book is aimed at raising awareness of researchers, scientists and engineers on the benefits of Principal Component Analysis (PCA) in data analysis. In this book, the reader will find the applications of PCA in fields such as taxonomy, biology, pharmacy,finance, agriculture, ecology, health and architecture.

**How to reference**

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Louis Noel Gastinel (2012). Principal Component Analysis in the Era of «Omics» Data, Principal Component Analysis - Multidisciplinary Applications, Dr. Parinya Sanguansat (Ed.), ISBN: 978-953-51-0129-1, InTech, Available from: http://www.intechopen.com/books/principal-component-analysis-multidisciplinary-applications/principal-component-analysis-in-the-era-of-omics-data

INTECH
open science | open minds

**InTech Europe**
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

**InTech China**
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2012 The Author(s). Licensee IntechOpen. This is an open access article distributed under the terms of the Creative Commons Attribution 3.0 License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.