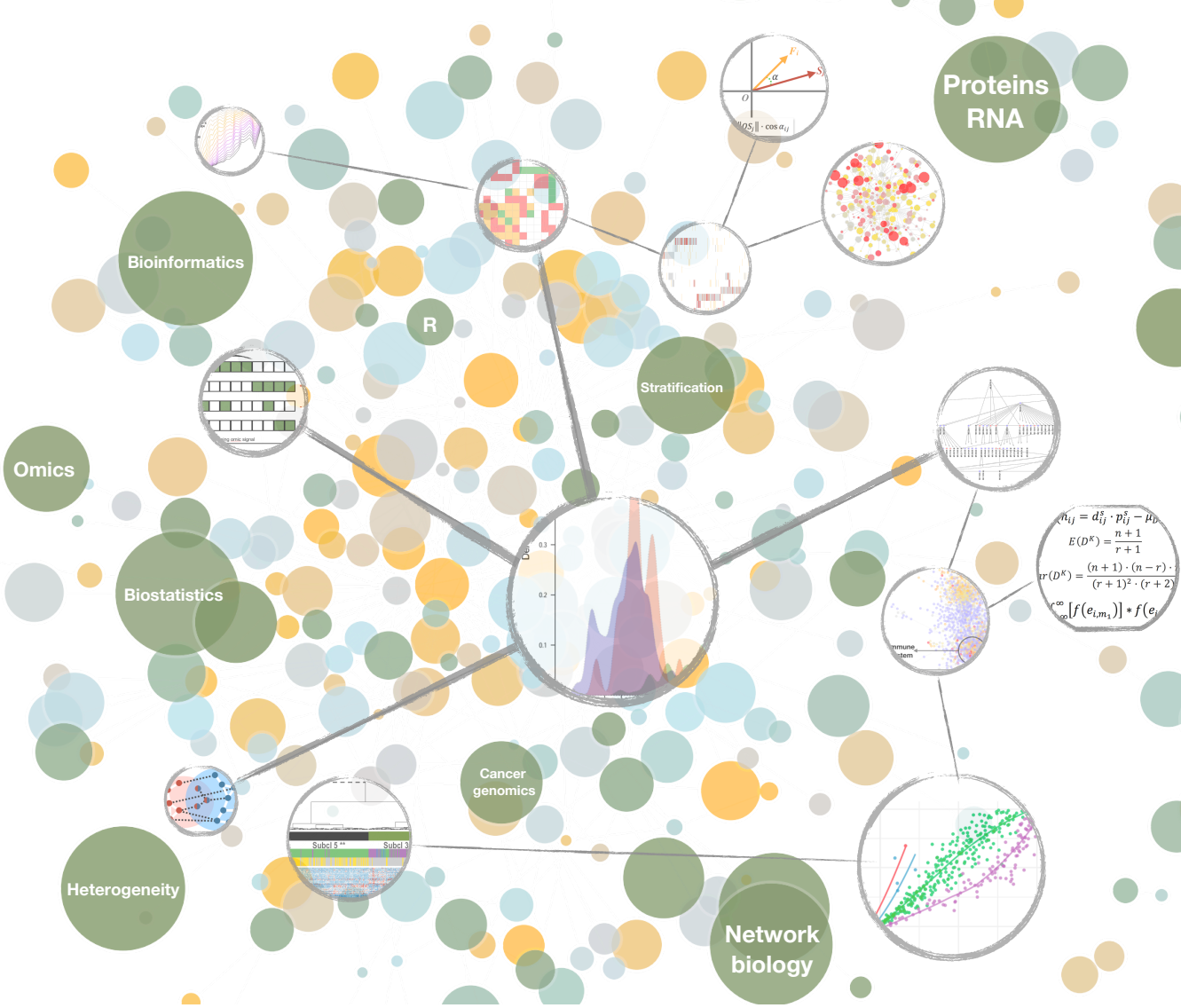


Bioinformatics applied to human genomics and proteomics:

development of algorithms and methods for the discovery of molecular signatures derived from omic data and for the construction of co-expression and interaction networks.

Francisco José Campos Laborie

VNiVERSiDAD D SALAMANCA

CAMPUS OF INTERNATIONAL EXCELLENCE

Ph.D. Supervisors

Javier De Las Rivas Sanz, Ph.D.
José Manuel Sánchez Santos, Ph.D.

Salamanca, Spain

2018

DOCTORAL THESIS

VNiVERSiDAD
D SALAMANCA

CAMPUS OF INTERNATIONAL EXCELLENCE

Bioinformatics applied to human genomics and proteomics:

development of algorithms and methods
for the discovery of molecular signatures derived from omic data and
for the construction of co-expression and interaction networks

Francisco José Campos Laborie

Ph.D. SUPERVISORS

Javier De Las Rivas Sanz, Ph.D.

José Manuel Sánchez Santos, Ph.D.

Salamanca, Spain.

2018

Dr. Javier De Las Rivas Sanz, con D.N.I. 15949000H, Investigador Científico del Consejo Superior de Investigaciones Científicas (CSIC), director del grupo de Bioinformática y Genómica Funcional en el Centro de Investigación del Cáncer (CiC-IBMCC), y profesor del Programa de Doctorado y del Máster de Biología y Clínica del Cáncer de dicho Centro y la Universidad de Salamanca (USAL).

Y el Dr. José Manuel Sánchez Santos, con D.N.I. 07870414K, Profesor Titular de Universidad del Departamento de Estadística, Facultad de Ciencias de la Universidad de Salamanca (USAL).

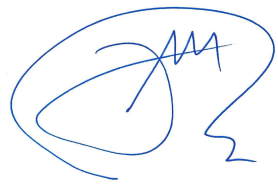
CERTIFICAN

que han dirigido la Tesis Doctoral titulada "Bioinformatics applied to human genomics and proteomics: development of algorithms and methods for the discovery of molecular signatures derived from omic data and for the construction of co-expression and interaction networks" realizada por D. Francisco José Campos Laborie, dentro del programa de doctorado Biociencias: Biología y Clínica del Cáncer y Medicina Traslacional del Centro de Investigación del Cáncer (CiC-IBMCC, CSIC/USAL).

Y AUTORIZAN

la presentación de la misma, considerando que reúne las condiciones de originalidad y contenidos requeridos para optar al grado de Doctor por la Universidad de Salamanca.

En Salamanca, a 20 de Julio de 2018



Dr. Javier De Las Rivas Sanz

Dr. José Manuel Sánchez Santos

Para la realización de esta Tesis Doctoral, el doctorando Francisco José Campos Laborie obtuvo en concurso público una *Ayuda destinada a financiar la Contratación Predoctoral de Personal Investigador*, cofinanciadas por el Fondo Social Europeo (FSE) y convocadas por la Junta de Castilla y León (ORDEN EDU/828/2014, de 29 de Septiembre de 2014). Anteriormente, pudo iniciar su investigación gracias a la contratación a través de un Proyecto de Investigación con *Celgene Institute for Translational Research Europe* (CITRE) asociado a la Fundación de Investigación del Cáncer.

La investigación de esta Tesis Doctoral ha sido realizada gracias a los fondos proporcionados por los siguientes Proyectos de Investigación, concedidos al laboratorio del Dr. Javier De Las Rivas: "*Data Representations and Similarity Measures for High-throughput Clinical Sample Profiles*", CITRE y FICUS (2013-2015); "*Biología molecular integrativa de hemopatías malignas: análisis bioinformáticos de datos transcriptómicos y proteómicos para identificar genes marcadores, genes causales y redes reguladoras asociadas a subclases patológicas de dos síndromes proliferativos*", ISCIII y Ministerio de Economía y Competetividad (2013-2015); "*Análisis genómico integrativo y búsqueda de marcadores específicos de Células Stem Mesenquimales (MSC) normales y alteradas en hemopatías malignas*", Junta de Castilla y León, Consejería de Sanidad (BIO/SA08/14); "*Genómica funcional de células stem mesenquimales (MSC) de individuos normales y pacientes con mieloma múltiple*", Junta de Castilla y León, Consejería de Sanidad (FIC335U14); "*Genómica y proteómica integrativa de hemopatías malignas mieloides y mieloma múltiple: estudio bioinformático de datos ómicos de muestras clínicas para identificar marcadores de pronóstico, respuesta y supervivencia, y estratificación de pacientes*", Ministerio de Economía y Hacienda (2016-2018); "*Plataforma de Bioinformática: bioinformatics and funcional genomics in cancer*", ISCIII (2017-2019).

Una parte de esta Tesis Doctoral se realizó, durante tres meses, en el laboratorio del Dr. Marc Vidal del *Center for Cancer Systems Biology* (CCSB), en el *Dana-Farber Cancer Institute* (DFCI) de la Universidad de Harvard, en Boston (Estados Unidos), gracias a la concesión de una *Travel Grant* por *Boehringer Ingelheim Fonds*. Esta colaboración seguirá vigente hasta finalizar el proyecto conjunto.

Esta Tesis Doctoral opta a la Mención de Doctorado Internacional por parte de la Universidad de Salamanca, y por ello se ha optado por la escritura en inglés en su totalidad, adjuntando un resumen en castellano.

A mi familia,
a mis abuelos,
a Ángela.

“Se equivocó la paloma, se equivocaba.”
Rafael Alberti

INDEX

ABBREVIATIONS

APID	Agile Protein Interaction Data Analyzer database.
AUC	Area Under Curve
BCC	Basal Cell Carcinoma – Breast Cancer
BLAST	Basic Local Alignment Search Tool
BM	Bone marrow
CA	Correspondence Analysis
CD	Compact disc
cDNA	Complementary DNA
CPM	Counts Per Million
DE	Differential expression
DEg	Differential expressed gene
DEV	Differential expression event
DECO	Decomposing heterogeneous Cohorts by Omic data profiling, method.
DLBCL	Diffuse large B-cell lymphoma
DNA	Deoxyribonucleic acid
EST	Expressed sequence tag
FDR	False discovery rate
FPR	False positive rate
FPKM	Fragments Per Kilobase per Million mapped reads
FS	Feature selection
GEO	Gene Expression Omnibus database
GEP	Genome-wide expression profiling
GO	Gene Ontology database
GTEx	Gene Tissue Expression consortium
HK	Housekeeping gene
HPA	Human Protein Atlas consortium
ID-BCC	Invasive ductal breast cancer
IL-BCC	Invasive lobular breast cancer
IQR	Interquartile range
iRNA	Interference RNA
LCA	Lowest Common Ancestor
lncRNA	Long non-coding RNA
MAD	Median average difference

MDS	Myelodysplastic syndrome
MI	Molecular interaction
miRNA	Micro RNA
mRNA	Messenger RNA
MNC	Mono-nucleated cells
MYA	Million years ago
NSCA	Non-Symmetrical Correspondence Analysis
OLS	Ontology Lookup Service
OMA	Orthologous MAtrix database
OR	Odds ratio
OSC	Osteosarcoma
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
pCom	Complete change (differential analysis)
pMaj	Majority change (differential analysis)
pMin	Minority change (differential analysis)
pMix	Mixed change (differential analysis)
pNULL	Null change (differential analysis)
PPI	Protein-protein interaction
RDA	Recursive Differential Analysis
RFE	Recursive Feature Elimination
rRNA	Ribosomal RNA
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RPKM	Read Per Kilobase per Million mapped reads
ROC	Receiving operative curve
SCC	Spearman Correlation Coefficient
sncRNA	Small non-coding RNA
SVM	Supper Vector Machine
SVD	Singular value decomposition
TE	Tissue-enriched gene
TPM	Transcripts Per Million
TPR	True positive rate
tRNA	Transference RNA
WPCC	Weighted Pearson Correlation Coefficient
Y2H	Yeast Two-Hybrid

LIST OF FIGURES

Letter code

I: Introduction, M: Material and Methods and R: Results.

LIST OF TABLES

LIST OF APPENDIXES

GENERAL OBJECTIVES

The present Ph.D. dissertation, entitled "**Bioinformatics applied to human genomics and proteomics: development of algorithms and methods for discovery of molecular signatures derived from omic data and for the construction of co-expression and interaction networks**", develops and applies Bioinformatics methods and tools to address current critical problems in the analysis of human omic data. As a main scope of the work, we approached two main issues in Bioinformatics: analysis of heterogeneous omic data from clinical samples and integration and analysis of different human omic biomolecular information in relational networks. As a general comment, all the work presented in this Ph.D. used and developed a wide variety of bioinformatic and statistical tools for the analysis, integration, and elucidation of molecular signatures and biological networks. Most of this data corresponds to sample cohorts generated in recent biomedical studies on specific human diseases.

This dissertation has been organised by main objectives into four different chapters focused on: **(i)** development of an algorithm for the analysis of changes and heterogeneity in large-scale omic data; **(ii)** development of a method for non-parametric feature selection; **(iii)** integration and analysis of human protein-protein interaction networks with subcellular location and **(iv)** integration and analysis of human co-expression networks derived from tissue expression data and evolutionary profiles of proteins.

Primary specific objectives of this Ph.D.

1a. Design and develop a **new bioinformatics method** to tackle the problems of **samples variability and heterogeneity**, detecting possible **sample mislabelling** and improving **outlier's identification**. To create such method, we designed a bioinformatic approach for in-depth analyses of large-scale omics data from biomedical samples to find and reveal all dependence relationships among omic features (i.e., genes, miRNAs, etc.) and samples (i.e., individuals of studied cohorts including their phenotypic and clinical characteristics). We also aimed to produce a

new statistic enclosing these properties that may improve current statistical approaches.

1b. **Write and implement a complete R package** corresponding to our **new method** to facilitate the use, accessibility, and interpretation of the results provided by the algorithm, accompanied by a detailed vignette and user guide.

2. Develop a **simple non-parametric statistic to measure the cohesiveness of categorical variables along a quantitative variable**, applicable to **feature selection** in different types of big data. Consequently, we will compare this cohesiveness statistic to the current state of the art of feature selection methods, either flat, wrappers or embedded approaches.

3. Integrate and analyse two high-throughput and systematic approaches from **high-quality proteomics** technologies: **HuRI** (Human Reference Interactome produced by Yeast-Two Hybrid technology) and **Cell Atlas** (comprehensive map of subcellular localization of all proteins of the human proteome, generated by antibody imaging). Explore the **inference of protein subcellular localization** given the integration of protein-protein interactome data and subcellular localization information, within the developed framework.

4a. Generation of a **robust human co-expression network** across multiple tissues based on the analysis of a large RNA-sequencing dataset from the Human Protein Atlas. Disclose and identification the **housekeeping (HK)** and **tissue-enriched (TE)** genes (or gene-products) based on this data, placing them in a relational human gene expression context.

4b. Integration the **human co-expression network** with **protein orthologous families** (derived from multiple sequence alignments, using OMA) plus an **evolutionary timeline of the human proteome** (using TimeTree), to investigate how old in evolution and how correlated are different human protein-coding genes. Generation of a relational network integrating: gene expression correlation (separating housekeeping or tissue-enriched) and protein evolutionary location in the timeline.

CHAPTER I

Development of a bioinformatic method for decomposing heterogeneous cohorts of samples using robust omic data profiling

BRIEF SUMMARY

Current approaches for differential analysis or supervised learning in the study of samples from patients with complex diseases have to deal with patient and individual diversity, disease heterogeneity and technical variability. Here, we present the development and use of a new computational method, called DECO (*DEcomposing heterogeneous Cohorts by Omic data profiling*), intended to analyse and understand heterogeneous omic data avoiding classical normalization approaches of reducing or removing uninformative variation.

Throughout **Chapter I**, DECO algorithm, statistical design and bioinformatic development are presented, including a detailed comparison to other current and well-established methods and the application to experimental transcriptomic data from several cohorts of cancer patients.

INTRODUCTION

The areas of precision medicine and big data analysis have grown almost exponentially in the last decade due to the hopes of improving patient diagnosis based on omic information (i.e., data produced by genomic, transcriptomic, proteomic or other omic global techniques). Collecting information from large sample populations was expected to enable the identification and precise treatment of every single patient:

"Precision medicine describes the definition of disease at a higher resolution by genomic and other technologies to enable more precise targeting of subgroups of disease with new therapies" (Ashley 2016).

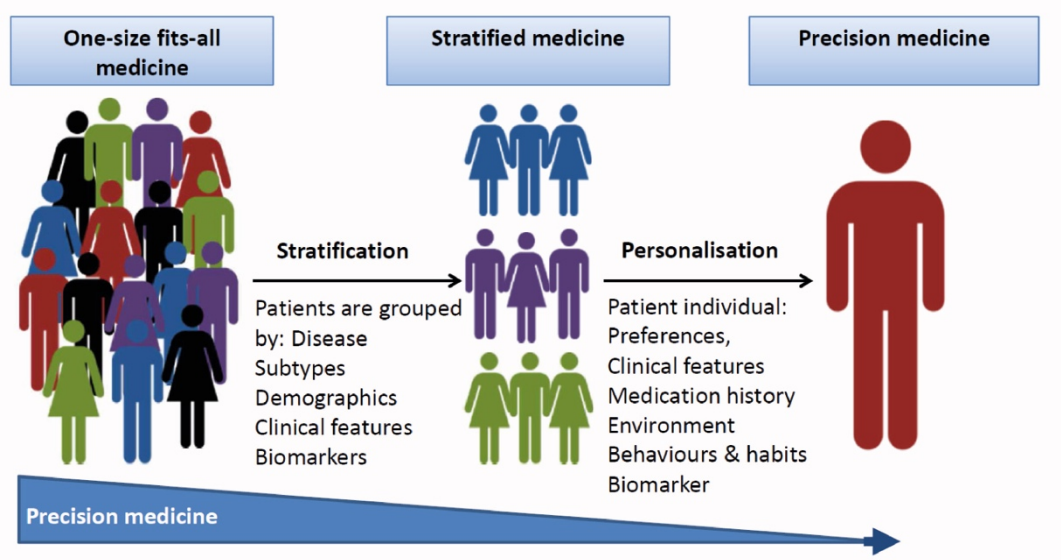


Figure 1-I-1. Hypothetical workflow of precision medicine from the raw population and big data to precision medicine for any patient, through stratification of patients into similar categories depending on omic biomarkers and phenotypical features.

Source: Manchester Precision Medicine Institute (http://www.mpmi.manchester.ac.uk/aboutprecisionmedicine/).

Theoretically, if genomic-phenotypical information of one patient is available, we would be able of classifying such patient into a specific category where diagnosis and treatment have been already studied and tested (**Fig. 1-I-1**). Most of the times, however,

collecting a huge amount of data make analyses noisy, and bring the need to delve inside data applying different filtering techniques or attempting to remove irrelevant information: trying to change from incomprehensive *big data* to informative *smart data*. Such accumulation of large-scale data creates a complexity that, combined with sample variability, gives rise to a difficult scenario, where it is very easy to make mistakes when searching for novel specific disease markers. Currently, there is an increasing trend for multi-omics integration (Huang et al. 2017) and single-cell omics platforms to directly approach this issue from the root (Qian et al. 2017, Levitin et al. 2018). Regardless, individual variability is one of the most intricate issues to deal with in biomedical studies of large patient cohorts even if several omic datasets are available (De Palma and Hanahan 2012, Rodriguez-Gonzalez et al. 2013).

Currently, large-scale omic techniques applied to clinical and biomedical studies are generating deep molecular profiles from patients. One of the omic techniques that have provided best and broader results is genome-wide expression profiling (also known as GEP) that can be achieved using multiple high-throughput platforms. Consistent changes in gene signals among disease subtypes are detectable using differential expression methods like SAM (Tusher et al. 2001) and LIMMA (Smyth 2004), which have been applied successfully in the last decade, mostly focused on control-case binary comparisons. However, clinical data from patients and human samples exhibit considerable variability unrelated to the property of interest. This problem is larger when comparing closely related pathological disease subtypes, where subtle differences can mark dramatic changes in diagnosis and prognosis. Apart from the patient heterogeneity mentioned above, clinical samples in the case of cancer studies can also show intra-tumour variability corresponding to the alteration of tumour cells related to microenvironment, evolving mutations or longitudinal changes along the progression of the disease (Bedard et al. 2013). In summary, the big impact of individual heterogeneity and genetic dynamics on biomedical omic studies makes finding specific and reproducible gene markers highly challenging (Beckman et al. 2012, Gillies et al. 2012, Cyll et al. 2017).

Along the Introduction of this Chapter I, we will revisit current statistic methods, introduce new theoretical concepts and detail the statistical background related to our method, called DECO. Particularly, we will focus on the current state of art of those approaches to heterogeneity issues in cancer omic analyses.

1. Gene expression and transcriptomics

From the classical point of view, gene expression refers to the biological process by which information contained in a DNA gene sequence is transcribed to mRNA for a posterior translation to protein (functional gene product). Classically, this molecular process from DNA to mRNA is called *transcription*, which carries out the expression of any gene. Later, this mRNA will be translated into a protein. **Figure 1-I-2** briefly describes the evolution of the central dogma of molecular biology. Interestingly, we know now that *transcription* not only generates mRNA molecules to be translated into proteins but also other RNA molecules are derived from non-coding genes: microRNA (miRNA), small non-coding RNA (sncRNA), long non-coding RNA (lncRNA), transference RNA (tRNA), ribosomal RNA (rRNA), etc. Apart from helping in the regulation of RNA processing, these other RNA molecules perform additional diverse functions not necessarily related to protein production (Huang et al. 2013).

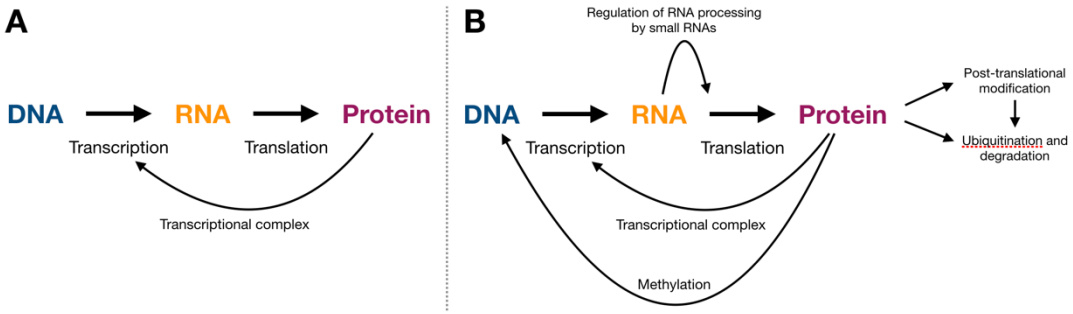


Figure 1-I-2. The central dogma of molecular biology: evolution from 1965 **(A)** to current general concept **(B)**. Adapted figure (Jafari et al. 2017).

The use of transcriptomic techniques has been arisen according to the development and improvement of these technologies. The first attempt was published in 1991 and included 609 mRNA sequences (expressed sequence tags - ESTs) from the human brain (Adams et al. 1991) obtained through an automated partial DNA sequencing procedure. However, current well-established transcriptomic methodologies, **microarrays** and **RNA-sequencing** (RNA-seq), were developed in mid-1990s and 2000s. For example, the first microarray analysis was published in 1995 (Schena et al. 1995) while first RNA-sequencing study was released in 2006 (Bainbridge et al. 2006). In 2008, there was a boom of RNA-seq analysis due to increased capacity of recording sequences (up to 10^9 sequences in a single experiment) reached by Illumina technologies. **Figure 1-I-3**

summarizes the methodological differences between microarray and RNA-seq technologies and how much different transcriptomic techniques have been used for analysis in terms of publications along the last 25 years (Lowe et al. 2017). As we can see, there was a great growth in the number of original research using RNA-seq due to its ability to read different RNA molecules and depth of coverage of gene expression signal.

1.1 Microarrays technology

Current microarray technology applied to gene expression allows measuring the abundance of thousands of transcripts via hybridisation of cDNA (complementary DNA obtained after reverse transcription of mRNA) to short oligonucleotides (probes of 25 - 50 oligomers) of specific sequence placed on a physical support (array). The probe generation by manufacturer needs a prior knowledge about the DNA sequence of the organism (i.e. genome reference). Once the hybridisation is done, the intensity of fluorescence of the probes corresponding to each biological entity (i.e. the set of probes for each gene) allows the quantification corresponding to the amount of cDNA hybridized (Barbulovic-Nad et al. 2006). Once processed, the microarray provides a global image that allows the quantification of thousands of genes at the same time.

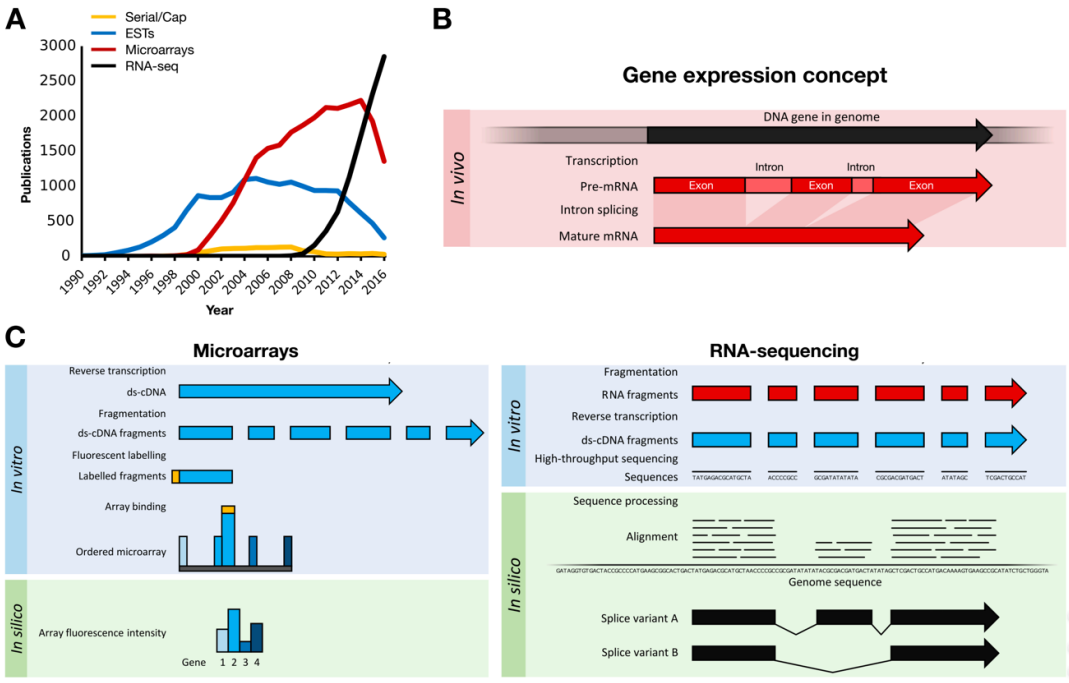
The microarray platforms most widely used and successful in transcriptomic studies have been produced by *Affymetrix*. After image scan, providing the raw signal of the microarrays, several data normalization processes (MAS5, Li-Wong, RMA, SVN or GCRMA) have been developed depending on the manufacturer and scope of the analysis (Li and Wong 2001, Hubbell et al. 2002, Huber et al. 2002, Wu et al. 2004). RMA method for *Affymetrix* platforms is the most extended and used microarray normalization method (Irizarry et al. 2003). All these different methods were intended to approach three main parts of any microarray normalization protocol: **(i)** background correction to remove noisy signal from fluorescence lecture; **(ii)** normalization to equal lectures from different samples within the same batch or experiment; and **(iii)** summarization of all probes composing a *probeset* (group of probe pairs that interrogate a sequence of a particular gene).

Additionally, since a probeset match to a specific sequence of a gene, some methods, usually called *gene-mappers* or *custom CDFs*, have been developed for *Affymetrix* platform to integrate all probesets lectures mapping a gene combining

duplicated probesets and removing no unique unspecific probes (Sandberg and Larsson 2007, Risueno et al. 2010). As a result of the normalization procedure and application (or not) of a custom CDF, we will obtain a matrix-table indicating the level of mRNA per gene per sample.

1.2 RNA-sequencing technology

Contrarily to microarray technology, RNA-sequencing (RNA-seq) technology is not biased by previous knowledge because it captures and sequences the transcripts contained at any RNA sample. Usually, the nucleotide sequences captured by this technique are around 100 bp (base pairs) but it may range between 30 and 100000 bp (100Kb). The depth of any RNA-seq experiment allows setting up different experiments depending on which kind of RNA molecules or biological process we are interested in. Usually, the background signal is very low for 100 bp reads in non-repetitive regions, providing a clear transcriptomic landscape for most of the organisms (Ozsolak and Milos 2011).



Modified from Lowe, Shirley et al. 2017. *PLOS Comput Biol*

Figure 1-I-3. Summary figure showing **(A)** trends of use of different transcriptomic techniques, **(B)** the classical gene expression concept and **(C)** two schematic figures about microarrays and RNA-seq technologies. Figure adapted (Lowe et al. 2017).

Typically, the data analysis provided by a RNA-seq experiment consists of 6 different steps: quality control, trimming, alignment, counting, normalization (for global adjustment of genome-wide and per sample) and differential expression. Although the insights about quality control or alignment procedures are out of the focus of this chapter, it is well-known they are crucial steps for a correct interpretation of posterior results. Regarding the absolute **quantification** step (includes counting and normalization), it provides the counts or number of reads per locus per sample and can be performed at gene, transcript or exon level. Along the last decade, different units have been proposed by experts for quantify amount of mRNA per sample to solve normalization issues related to gene length or sample background:

- **RPKM (Reads Per Kilobase per Million mapped reads):** Normalization method designed for single-end RNA-seq experiments that normalize first for differences in *sequencing depth* and second for differences in *gene size*. The procedure is: (i) add the total reads in each sample and divide by 10^6 (*per million scaling factor*); (ii) divide the read counts of each gene by the *per million* scaling factor, giving *reads per million* or RPMs; (iii) divide the RPM values by the length (in kilobases, Kb) of each gene, returning final RPKM values.
- **FPKM (Fragments Per Kilobase per Million mapped reads):** The same procedure that RPKM but designed for paired-end RNA-seq experiments.
- **TPM (Transcripts Per Million):** Normalization method similar to RPKM and FPKM with a change in the order of the operations: first normalizes for differences in *gene size* and second for differences in *sequencing depth*: (i) divide the read counts by the length (in kilobases, Kb) of each gene, giving the reads per kilobase or RPKs; (ii) count all the RPK values in each sample and divide by 10^6 (*per million scaling factor*); (ii) divide the RPK of each gene by the *per million* scaling factor, providing final *transcripts per million* or TPMs.
- **CPM (Counts Per Million):** Counts divided by the total number of reads (sequencing depth) and multiplied by 10^6. This normalization method does not take into account the length of genes.

Nowadays, TPMs are highly recommended instead RPKMs or FPKMs because the sum of TPMs per sample would give us the same number, while RPKMs or FPKMs can

vary along samples. However, most of the initial RNA-seq analyses described in literature were based on RPKM or FPKM calculations. All these units require a posterior transformation, where a change to the logarithmic scale log2(signal+1) should be performed because the distributions of expression signal are not linear. Previously, it is necessary adding 1 to remove those RPKM/FPKM/TPM/CPM values equal to 0, which would be transformed into (negative) infinite otherwise. We could analyse the expression signal by absolute quantification (estimation of the expression level of each gene) or by relative quantification (comparison of the expression between different types of samples).

Here, it is important to mention that these approaches tend to perform poorly if heterogeneous transcript distributions are present due to highly and differentially expressed genes which skew the count distribution (Bullard et al. 2010). For this reason, some newer normalization methods, like TMM, DESeq, UpperQuartile or PoissonSeq (Conesa et al. 2016) try to avoid or ignore these highly variable features.

1.3 Differential expression analysis

Since both microarray and RNA-seq technologies provide a matrix containing normalized levels of mRNA per gene/exon/transcript per sample, a differential expression analysis will aim to compare these levels among groups of samples if categories are suitable of comparison. In this way, we would test if an observed difference in these mRNA level is greater than what would be expected by chance.

Since the microarray technology was the first transcriptomic method to be widely accepted, the earliest methods for differential expression analysis were developed taking into consideration the particular statistical distribution and background derived from microarray normalization. As mentioned above, RMA method and Affymetrix platform have been the most used normalization method and microarray platform, respectively, so classical methods for differential expression on microarrays, like SAM or LIMMA, were accordingly designed. Due to microarray value distribution is not following any known statistical distribution, some investigators prefer applying a non-parametrical rank statistical hypothesis test, like the Mann-Whitney-Wilcoxon test or Rank Product test (Breitling et al. 2004), or assuming normal distributions to apply the classical t-test. It is important to mention that the results of any differential expression analysis may vary a lot depending on

which method was chosen. In fact, it has been found a very low-level concordance among results on real array sets (Chrominski and Tkacz 2015), what makes even necessary to know the statistical assumptions behind each method. A brief description of classical approaches is following:

- **SAM**: non-parametric statistical method to determine statistical significance between groups. It rebuilds the t-test to make it non-parametric (Tusher et al. 2001). SAM uses the FDR and q-value method presented in Storey (Storey 2002).
- **LIMMA**: parametric method based on linear models to discover significant changes between groups of samples. Depending on the experimental design (two groups or more than two groups), it will be based on the β-statistic or F-statistic for assessing the significance of expression changes (Smyth 2004).
- **Rank Product**: non-parametric method based on rankings of fold changes between groups of samples compared (Breitling et al. 2004).
- **Mann-Whitney-Wilcoxon**: non-parametric method used to test conformity between two populations. One of the most used test if the distribution is unknown.
- **T-test**: a statistical test based on the average and variance of the population to determine whether the two groups differ from one another. One of the most used test to assess significant changes if the distribution is normal.

Attending to RNA-sequencing differential expression analysis, there are different methods specifically designed for RNA-seq distributions. For example, **edgeR** (Robinson et al. 2010) or **DESeq** (also **DESeq2**) (Anders and Huber 2010) are based on negative binomial distributions, which has been associated with RNA-seq value distributions. Alternatively, **EBSeq** (Leng et al. 2013) or **baySeq** (Hardcastle and Kelly 2010) are Bayesian approaches to the negative binomial model. Additionally, **LIMMA** could be also applied to RNA-seq data and their authors developed a normalization method based on CPMs, called **voom** (Law et al. 2014), while SAM's author also developed a new version of SAM method, called **SAMSeq**, to deal with RNA-seq data (Li and Tibshirani 2013).

Along this Chapter I, some comparisons will be focused on classical and well-established methods like SAM, LIMMA and t-test and their performance on real and artificial transcriptomic datasets, not only from microarray technologies due to its larger historical use in bioinformatics and in clinical omic studies, but also from RNA-sequencing

experiments.

2. Outlier profiling methods for cancer studies

The idea that genes are often deregulated in only a subset of patients, especially in cancer studies, led to the development of an interesting method called *Cancer Outlier Profile Analysis* (**COPA**) (Tomlins et al. 2005, MacDonald and Ghosh 2006). Outlier genes are defined as the ones that show signals very different from the average in a subset of samples in some or in all of the studied classes (**Fig. 1-I-4A**). Indeed, the difference between an outlier and a normal gene is that the outlier has a modified expression only in a minority of the studied samples, indicating a heterogeneous behaviour in such sample subset (Tomlins et al. 2005, MacDonald and Ghosh 2006). This idea comes from a well-known biological event in cancer: genetic translocations lead to up-regulation of oncogenes, which may affect cancer progression or reflect a particular state.

In order to find outlier genes, or outlier features in general, several algorithms have been proposed in the last decade based on different modifications of statistical tests, clustering analysis or resampling techniques applied to either original omic data or multidimensional transformed data (Li et al. 2007, Tibshirani and Hastie 2007, Wu 2007, Baty et al. 2008, Lian 2008, Hiissa et al. 2009, Wang and Rekaya 2010, Mpindi et al. 2011, de Ronde et al. 2013, Yang and Yang 2013, Roden et al. 2014, Noto et al. 2015, Nabavi 2016).

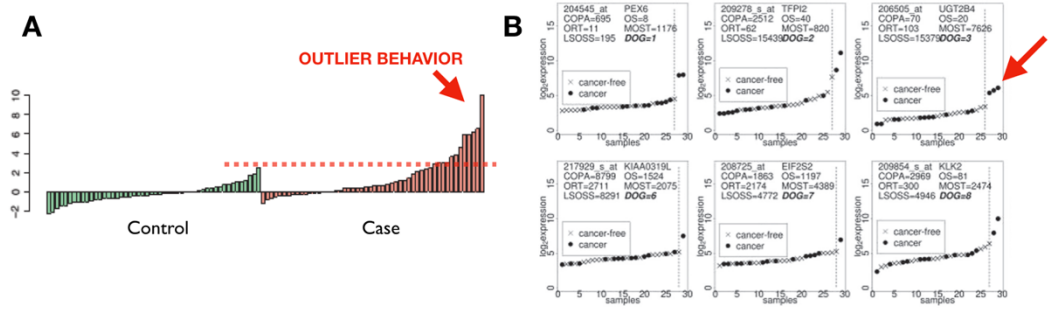


Figure 1-I-4. Modified figures from *COPA* and *DOG* algorithm publications (Tomlins et al. 2005, Yang and Yang 2013). Panel **(A)** describes a hypothetical gene profile from *COPA* point of view, showing a subset of samples with higher expression signal than the highest value of control group. Panel **(B)** shows exactly the same conceptual definition from a *DOG* point of view, providing some examples where *outlier profile* is only supported by one or two samples.

As we mentioned above, the first method approaching this issue was **COPA** (MacDonald and Ghosh 2006). First article describing COPA correctly identified strong outlier profiles of *ERG* and *ETV1* genes, both from *ETS* family transcription factors, in several prostate cancer datasets (Tomlins et al. 2005). This method is intended to discover pairs of genes affected showing mutually exclusive profiles due to genetic translocation event and following an *outlier* profile. Aiming that, COPA is based on classical methods for detecting differences among two group of samples (*t*-tests or Mann-Whitney tests) but including proper modifications: centred and scale each omic profile (typically, rows correspond to genes and columns to samples in our omic data matrix) using median and median average difference (MAD). Thus, given a cut-off for number of *outlier samples*, **COPA** could rank any feature comparing sum of outlier samples for each pair of genes (MacDonald and Ghosh 2006).

2.1 State of art: feature-based methods

After COPA, a wide variety of statistical approaches for ***outlier profile detection*** appeared. Indeed, **OS** method modifies COPA statistic superseding percentile cut-off by the inter-quantile range of the expression data in both groups (Tibshirani and Hastie 2007). Then, **ORT** algorithm modifies OS statistic considering interquartile range only for control or reference samples and, consequently, replacing global median values by median values per category of samples (Wu 2007). Alternatively, in order to consider all possible values for outlier thresholds, the **MOST** method requires case expression data be sorted in descending order. This method carried out trimmed mean for all possible outlier thresholds to consider maximum values as outlier profiles (Lian 2008). Posteriorly, **LSOSS** algorithm postulates that case samples could follow different profiles (activated and inactivated samples) and proposes a model related to fitting least squares of gene expression data. They also include a modified hierarchical clustering method developed to classify the heterogeneous gene activation patterns. Authors of this paper remove median approach for t-statistic distribution of the expression data, and return to mean values for each class (Wang and Rekaya 2010). Additionally, **GTI** method were presented as an alternative to COPA, OS and ORT (similar performance on single studies) for the application on meta-analysis but indicating problems with down-regulated genes (Mpindi et al. 2011). Next, **ZODET** was compared to GTI, which considered to transform each gene profile to z-scores

with respect to control samples (Roden et al. 2014). Following this approach, this method will fail and over-rate those genes whose standard deviation and mean of control samples are low, usually called *flat genes* in the literature.

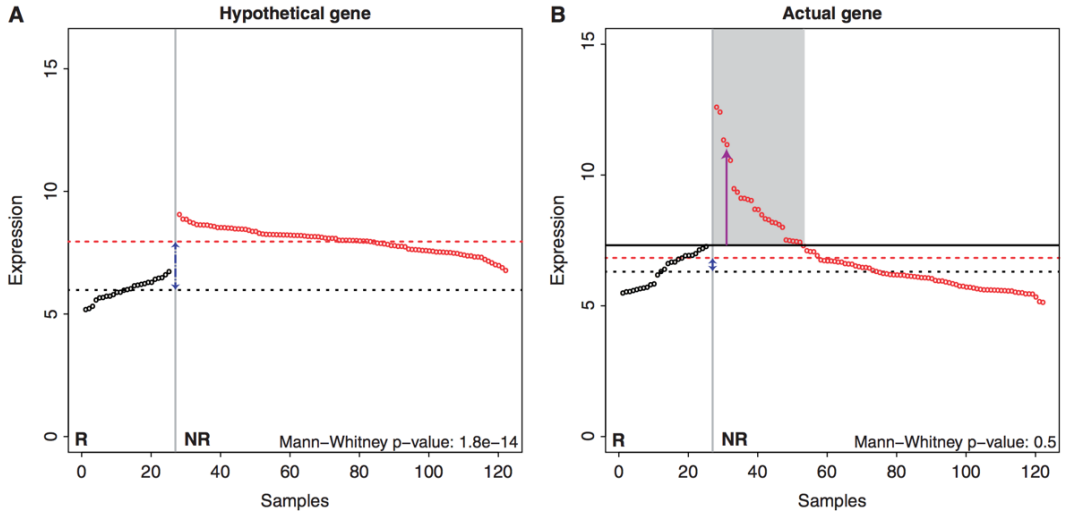
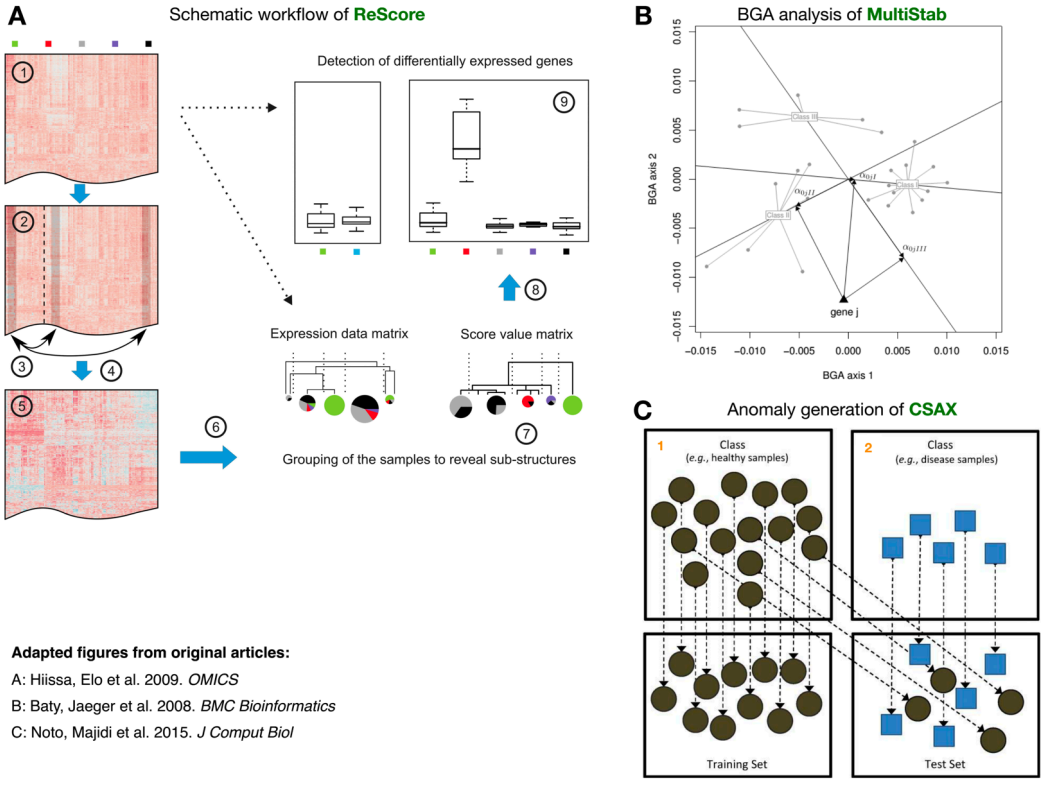


Figure 1-I-5. Original figure from DIDS algorithm publication (de Ronde et al., 2013). Panel **(A)** describes a hypothetical gene profile with DE where any outlier-behaviour is observed within case samples (red dots) against control samples (black dots). Panel **(B)** corresponds to actual gene outlier profile where a subset of case samples is significantly changed respect to control samples value range. The Mann-Whitney p-values are shown at bottom-right corners of each panel. DIDS assumes control as stable profiles with no abrupt dispersion or variability.

Several of these new approaches based on slight modifications of initial proposal done by COPA authors (OS, ORT, MOST, LSOSS) were independently compared by Karrila et al. using the Bhattacharjee dataset (Bhattacharjee et al. 2001) with 139 adenocarcinomas and 17 normal lung samples (Karrila et al. 2011). The comparative was conceived to decide which method could be implemented as a part of a semi-supervised method to discover predictive biomarkers. According to their benchmark, they concluded that MOST was the most suitable algorithm for outlier profile detection due to its stability after resampling and its gene ranking agreement with other methods. Instead, LSOSS outperformed similar to t-test while OS and COPA behaved analogously. However, this study started from an incorrect premise: algorithms for *outlier profile detection* must be stable if a subset of samples was added or removed from the original dataset (a resampling benchmark was established). If this premise was correct, all these methods would give similar scores for genes even if the outlier samples were removed.

Outstanding from all these methods, **DIDS** algorithm (de Ronde et al. 2013) approaches the *outlier profile detection* through a very simple way obtaining notable results in several real and experimental datasets. Given a group-group comparison between control and cases samples (interesting for the study), DIDS considers control samples as stable and reference values from which case values deviations would indicate us if there are samples following an outlier behaviour within case group (**Fig. 1-I-5**). Additionally, the authors proposed a permutation-based *p-value* to determine which particular profiles were not showing an outlier profile by chance (lower number of samples out of control values range would be discarded).



Adapted figures from original articles:

A: Hiissa, Elo et al. 2009. *OMICS*

B: Baty, Jaeger et al. 2008. *BMC Bioinformatics*

C: Noto, Majidi et al. 2015. *J Comput Biol*

Figure 1-I-6. Adapted figures from original papers detailing key points of: **(A)** ReScore, **(B)** MultiStab or **(C)** CSAX methods (structure-based methods).

2.2 State of art: structure-based methods

Apart of these statistical approaches to cancer outlier profile detection, which were focused on the gene profile *per se*, some authors proposed new methods based on data

structure and resampling procedures. In fact, **ReScore** was intended to re-score the original omic data matrix to facilitate the splitting of any omic data in original subclasses analysed (**Fig. 1-I-6A**). Here, initial classes of samples are needed to transform the omic data matrix (Hiissa et al. 2009). Moreover, **MultiStab** was designed for finding out stable biomarkers among categories of samples. Contrarily to the concept proposed by COPA, they were searching stable biomarker for each category of samples. Then, they estimated the stability of all gene coordinates after applying a Between-group Analysis (BGA) (**Fig. 1-I-6B**), which is based on ordination of group of samples similarly to Correspondence Analysis, within a resampling design (Baty et al. 2008). Alternatively, **CSAX** introduced a new way to approach this issue: applying an iterative learning method on a subset of control samples, leaving all case samples and the rest of samples as test set (**Fig. 1-I-6C**). Then, the method would try to identify the control samples within the test set and assess the anomaly (and the precedence) due to the previous mix (Noto et al. 2015). Roughly, it is important to mention that these methods are heuristic.

2.3 Questionable hypothesis behind current methods

Unfortunately, all those methods designed for *outlier profile detection* are missing a crucial reality here: control or reference samples also present intrinsic differences. As well as intra-tumour heterogeneity for case group, each specific patient genetic background involved as control or reference will be partly responsible for variability within control group. For this reason, basing *outlier profile detection* on stable patterns along control samples may lead to higher false negative and positive rates, even more accused if reference group has lower size (as usual). For example, **DOG** method (Yang and Yang 2013) is able to identify as significant outlier profiles those genes where only 1 or 2 samples deviate from control or reference group (**Fig. 1-I-4B**) while **ZODET** is conceptually based on control samples' stability. In addition to the probable patient's variability within the control group and intra-tumour heterogeneity, those omic techniques focused their samples on a cell population would inflates variability due to cell type differences along any tissue.

3. Theoretical framework

As hinted above, the algorithm proposed in this Chapter I (DECO) approaches intrinsic heterogeneity within a group of samples trying to find out which features are directly related to variability. Due to biological characteristics of each individual, any group of samples could be split in different subgroups attending to the similarities among their omic profiles. In fact, this is main concept of personalized medicine: to discover the similarities and differences among patients to properly diagnose and treat them by stratifying in subgroups.

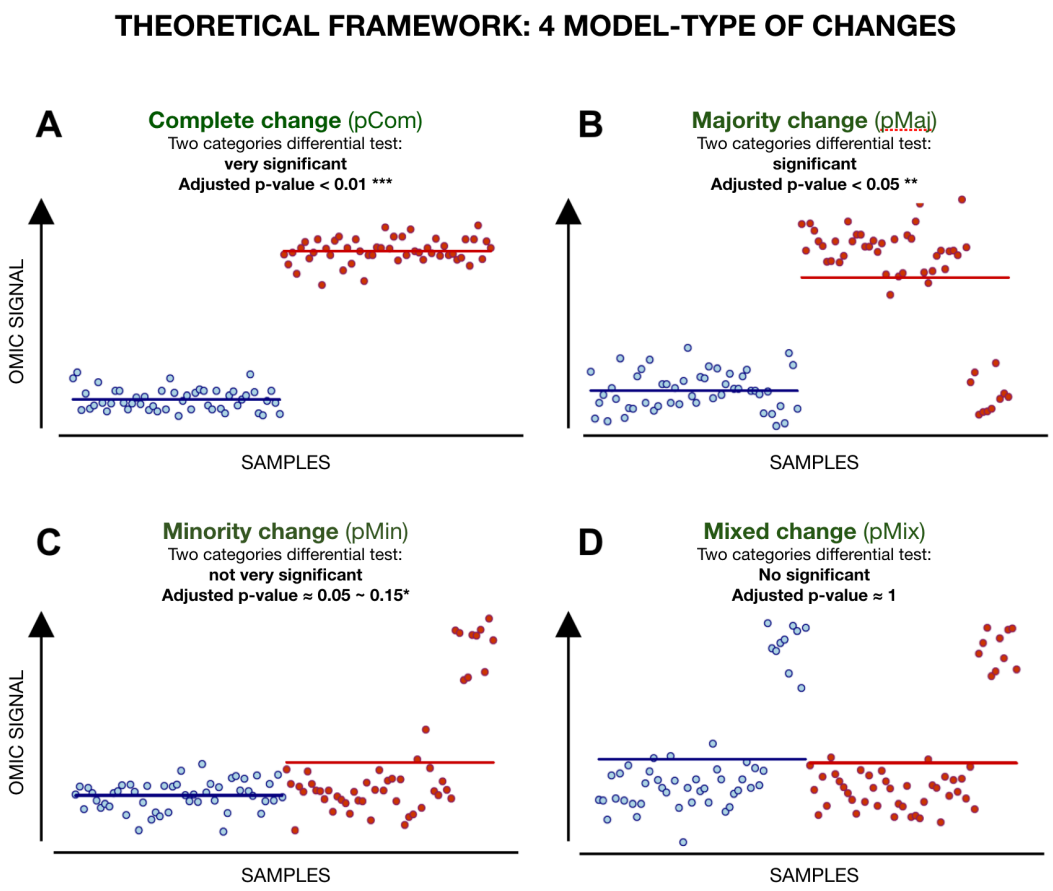


Figure 1-I-7. Theoretical framework presenting four model-types of change expected when comparing two predefined classes: **(A)** complete change (pCom); **(B)** majority change (pMaj); **(C)** minority change (pMin); **(D)** mixed change (pMix). The *adjusted p-values* using FDR correction after *t-test* comparisons are shown in each model-type for a measured variable. The plots represent in blue the signal of such feature for the *control* samples and in red the signal of such feature for the *case* samples. All model-types were extracted from simulated data (10000 features and 80 samples, following a RMA value distribution for *Affymetrix* microarray).

Particularly, the **biomedical investigation is mostly focused in group-group**

comparisons when any omic data analyses were carried out. For simplicity, investigators try to discover which particular transcriptomic, proteomic or genomic landscape are behind a specific cell state, disease or biological background through the comparison against a reference state. However, classical bioinformatic methods of differential analysis (e.g. for transcriptomic data: t-test, Mann-Whitney-Wilcoxon, SAM, LIMMA, edgeR, DESeq, etc) do not deal with heterogeneity as a part of the analysis, failing at some differential scenarios.

Hence, investigators always focus on *adjusted p-value* from multiple testing correction for establishing a cut-off or threshold (usually *adjusted p-value* < 0.05), which may lead to misinterpret results in some cases (**Fig. 1-I-7**). Depending on the sample size of each group compared and the statistical method used, the *adjusted p-value* behind a comparison between two states could range around the typical threshold mentioned above. Thus, any decision made and based on a significance threshold may be inappropriate.

Attending to possible omic scenarios for a single differential profile, we could roughly hypothesize four model-types when a group-group comparison is settled, described in **Figure 1-I-7**. Ideally, we would expect that a subset of features shows a ***complete*** change between two groups (panel A). This scenario is correctly detected by any of current statistical methods and reflected by the *adjusted p-value*. Nevertheless, the *adjusted p-value* could miss relevant information when ***majority*** or ***minority*** change profiles are present in our data (panel B and C). While the *adjusted p-value* of *majority* profiles may be considered as positive biomarker between classes, *minority* profiles should be considered as false positives. Both *adjusted p-values* range depending on group size balance and method used to analyse the data. Finally, ***mixed*** change is correctly defined as true negative by any differential method (panel D). Although this differential change is not directly related to original group of samples, its variability may be associated with another sample information.

In this way, we hypothesize that knowing both significance of differential change and the profile type of any omic feature would greatly increase the interpretability of the results after a common group-group comparison. For this reason, DECO algorithm (detailed along this Chapter I) will be intended to scrutinize and categorize all omic profiles if suitable.

4. Statistical tools: Resampling techniques

In statistics, resampling corresponds to different methods for validating models, estimating statistics from a sample/population or iteratively performing significance tests. These techniques generally provide more accurate results than traditional methods and allow involving a wide variety of modifications, depending on which type of data would be analysed (Hesterberg et al. 2005). As a resampling method, we may highlight bootstrap, jackknife, subsampling, cross-validation and permutation tests. Briefly, the main characteristics are following:

- **Bootstrap** is considered one of the most known resampling methods, consisting of a sampling with replacement (one sample may appear more than one time in a single iteration) to estimate the sampling distribution of any particular statistic.
- **Jackknife** is more often applied for statistical inferences of bias and errors of any statistic. Classically, this resampling scheme is characterized by *n-1* iterations of *n-1* subsets of samples (removing one sample).
- **The permutation strategy** is a resampling method where significance of a test under null hypothesis is assessed through the random rearrangements of sample labels. Monte Carlos testing is one of the most famous permutation tests.
- **Cross-validation** aims to validate a predictive model through the validation of this model on multiple subsets of data (trained with the rest of data).
- **Subsampling** is similar to bootstrap but the resample size is smaller than total and the resampling scheme does not include replacement. Interestingly, it was reported is validity at many common scenarios whereas bootstrap does not.

Different resampling techniques have been broadly adapted in bioinformatics since the expansion of gene expression analysis around 1995, due to the recent advances in cDNA microarray technologies and analyses in early 1990s (Lenoir and Giannella 2006). At that moment, the main purpose was to identify subset of genes which allow the discrimination of different group of patients, mostly focused on unsupervised experiment designs where no initial classes were defined. Later, once the classes were properly characterized, the bioinformatic field turn to the class-predictor analysis, where the investigator wants to define a minimum subset of features for predicting new samples

belonging to any known category (Molinaro et al. 2005).

Particularly, the feature selection procedure included in our method (DECO) would consist of a subsampling procedure, which will explore all the omic data in order to provide a robust output of variables and significant genes.

5. Statistical tools: Non-Symmetrical Correspondence Analysis

Principal Component Analysis (PCA) is one of the most used techniques for the analysis of gene expression data. Since low dimensionality of omic data was broadly proven (Heimberg et al. 2016), PCA provides information on the overall structure of the data, organizing features and samples according to their similarity and variability, but now based on abstract measures after data transformation (Sanguansat 2012). However, the lack of direct interpretability (more accused if the dataset is heterogeneous) and its dependence on effect size and fraction of samples containing biological signal may lead to meaningless results (Lenz et al. 2016).

If the omic data is collected in a contingency table (absolute frequency matrix), classical tools like Correspondence Analysis (CA) would provide a measure and visualization of how strong are the associations between rows and columns. The CA is based on the decomposition of the index ϕ^2 of Pearson, a symmetric measure of association. However, the omic data compiled (i.e. gene expression, protein concentration, methylation level, etc.) is always intended to explain or define a biological state, pathology or particular group of samples. Therefore, the explanatory variables and response variables are predefined from the beginning. To overcome those misinterpretations, the Non-Symmetrical Correspondence Analysis (NSCA) was introduced by D'Ambra and Lauro in 1984 (Diday 1984). NSCA allows to establish ***asymmetric associations*** among features and samples in a common dimensional space transforming a frequency or contingency table N into a matrix of centred column profiles (Eq. 1.1). Given two categorical variables I and J, NSCA aims to assess how explanatory variable J influences on the distribution of response variable I.

$$\prod = (\pi_{ij}) = \left(\frac{p_{ij}}{p_{.j}} - p_{i.}\right) \qquad \text{(Eq. 1.1)}$$

where $p_{ij} = \frac{n_{ij}}{n}$ is the relative frequency for each absolute frequency n_{ij} of N. Then,

the total variability or inertia explained by each component within a NSCA (features and samples) is summarized by the Goodman-Kruskal τ index (Goodman and Kruskal 1959, Sanguansat 2012).

$$In = \tau_{num} = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{\cdot j} \cdot \left(\frac{p_{ij}}{p_{\cdot j}} - p_{i\cdot} \right)^2 \qquad \text{(Eq. 1.2)}$$

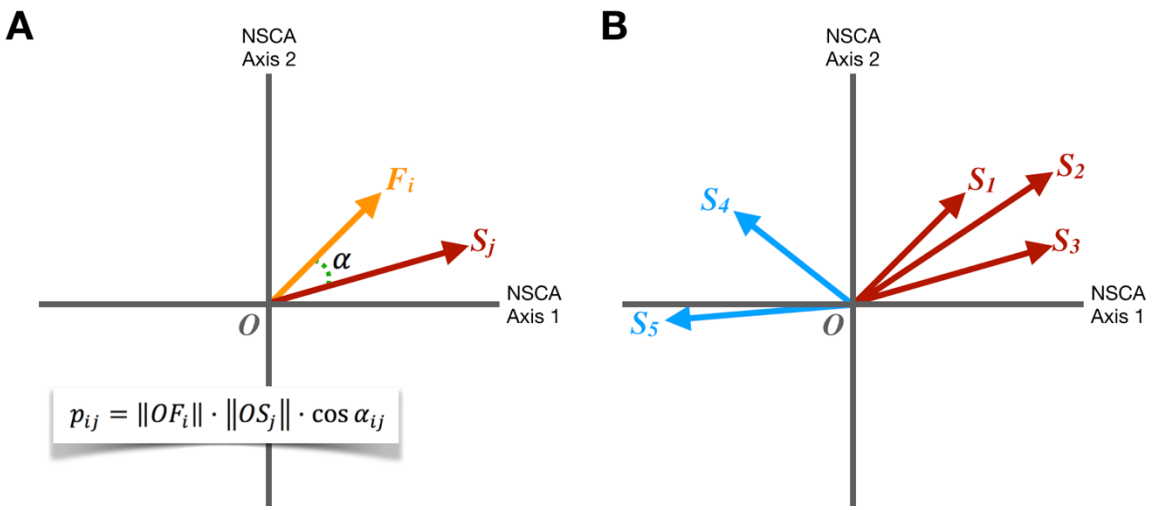


Figure 1-I-8. Inner product calculated by NSCA would orient similar categorical variables in the same direction. **(A)** Inner product calculation between a feature F_i and a sample S_j. Greater absolute inner product would indicate association (positive or negative dependence) while values closer to 0 would signify no association. **(B)** Theoretical example of behaviour for different subgroup of samples.

Eigen-value decomposition of the inertia variance covariance matrix would lead to factorial row and column coordinates definition for all categorical variables I and J. These coordinates would correspond to orthonormal singular vectors from both categorical variables, represented in the same lower (n-1) dimensional space. This space allows to characterize: **(i)** the asymmetric associations among both categorical variables I and J; and **(ii)** predictor or response outliers whose behaviour is significantly different from the rest.

According to D'Ambra and Lauro proposal, to achieve a good quantitative estimation of the predictor-response relationship between two single categorical variables i and j is necessary to calculate the ***inner product*** (**Fig. 1-I-8**). Thus, inner product would reflect not only closeness between i and j but similar direction from origin. In our particular scenario, given a contingency matrix obtained from omic data, a greater inner product

would reflect both analogous direction and vector size for feature i and sample j. Hence, we could infer which feature's profile i (response) is significantly biased by sample j (predictor).

Apart of asymmetric associations, a greater singular vector size after is directly related with τ_{num} contribution (Beh and Lombardo 2014). Any outlier feature or sample behaviour can be inferred depending on this value distribution, also allowing us to discover similar patterns or subgroups within both categorical variables I and J.

It is noteworthy that a deep search in the literature revealed NSCA has not been widely applied to clinical data or omic data analysis: it was used as ordination method for the development of MCIA (Multiple Co-Intertia Analysis) (Meng et al. 2014), or for particular analyses (Ciavolino et al. 2017, Vega-Hernandez et al. 2017). Regardless, the Non-Symmetrical Correspondence Analysis constitutes an excellent tool for analysing dependence relationships among two categorical variables (rows and columns) given a contingency table or absolute frequency matrix relating them. It will provide a meaningful interpretation of exclusive relationships through this predictor-response decomposition, establishing the source of this association and avoiding common symmetrical assumptions where both categorical variables are similarly weighted (Beh and D'Ambra 2010).

MATERIAL AND METHODS

This section presents the experimental and artificial datasets used along of this part of the work. It is important to note that artificial datasets were created for a particular validation while experimental datasets were chosen due to its biological relevance or singular characteristics.

1. Transcriptomic datasets

Due to their wide-knowledge, their easy implementation in bioinformatics pipeline and their accessibility, the analyses, tests, and experiments produced in this Chapter 1 have been applied on transcriptomic datasets from different omic platforms. Both experimental and artificial or simulated datasets will be broadly detailed.

We will compare it against other methods designed to analyse differences and find heterogeneity and outliers in disease sample cohorts: COPA (MacDonald and Ghosh 2006), OS (Tibshirani and Hastie 2007), ORT (Wu 2007), MOST (Lian 2008), LSOSS (Wang and Rekaya 2010) and DIDS (de Ronde et al. 2013) (**Table 1-M-1**) plus the standard t-Test.

Table 1-M-1. Statistical methods focused on finding *outlier gene (feature)* profiles within cancer samples.

Method	Strategy to find outliers	Search for down-regulated outlier features	Weight size of outlier samples	Search for sample subgroups	Reference	Year
COPA	Percentile - MAD	No	No	No	MacDonal et al.	2006
OS	Quantile ordered and cut-off	No	No	No	Tibshirani & Hastie	2007
ORT	Robust t-statistic	**Yes**	No	No	Wu	2007
MOST	Maximum ordered	**Yes**	No	No	Lian	2008
LSOSS	Least sum of ordered subset square t-statistic	**Yes**	No	No	Wang & Rekaya	2010
DIDS	Maximum value from control group	**Yes**	**Yes**	No	de Ronde et al.	2013
DECO	Recursive Differential Analysis and NSCA	**Yes**	**Yes**	**Yes**	*present work*	2018

1.1 Artificial transcriptomic datasets

Several simulated artificial datasets were generated to assess the performance of DECO algorithm and the comparison with other methods. The simulated datasets were designed to have an expression data matrix that included signals for 1100 genes and 40 samples in two classes, 20 in class type 0 (n_1 = 20 controls) and 20 in class type 1 (n_2 = 20 cases). The specific design followed a similar scenario to the LSOSS benchmark (Wang and Rekaya 2010), with differential changes representing two different situations: **(A)** a dataset with a subset of 100 features showing differential expression (DE) for a subset of the case samples (**Fig. 1-R-1A**), reproducing the *minority change* (pMin) profile (**Fig. 1-I-7C**); and **(B)** a dataset with a subset of 100 genes showing differential expression (DE) for a subset of samples of both classes, reproducing the *mixed change* profile (**Fig. 1-I-7D**), so that there is not global DE between classes for these genes (**Fig. 1-R-1B**). Both scenarios describe a balanced experimental design ($n_1 = n_2 = 20$).

According to the experimental designs described above, the simulated expression data include two matrices: **(i)** a matrix with 1000 genes without DE, called matrix F (used for calculation of the False Positive Rates, FPR); and **(ii)** a matrix with 100 genes with DE, called T (used for calculation of the True Positive Rates, TPR). The expression values follow a *Normal distribution* with means $\mu(F)=\mu(T)=0$ and deviations $\partial(F)=\partial(T)=1$, as established in other studies (Wu 2007). Then, we created different *outlier* situations varying number of η outlier samples within T matrix. Thus, given any feature of T, the number of outlier samples showing DE would change among $\eta \in \{1,3,5,7,9\}$ (**Fig. 1-R-1C** and **D**). However, in the **Figures 1-R-1A** and **B**, we only show $\eta=5$ (25% of group samples altered). The differential expression signals were generated by adding a constant value (δ = 2) to the outlier samples but keeping their variability.

Apart from this first simulation using normal distributions, an adapted version of the *simData* function from *optBiomarker* R package (Khondoker et al. 2010) was used. This method allows generating artificial gene expression data with DE signal based on distributions after RMA normalization (Irizarry et al. 2003), and it was applied to test the methods for all 4 model-types described above (**Fig. 1-I-7**). Here, *outlier* sample size was fixed to k=5. The objective of this wide benchmark was determining the ability of each method of detecting and scoring (with each own statistic) these 4 model-types, through the simulation of 8 different feature profiles within a whole dataset composed of 10000

feature's profiles along 40 samples (**Fig. 1-M-1**).

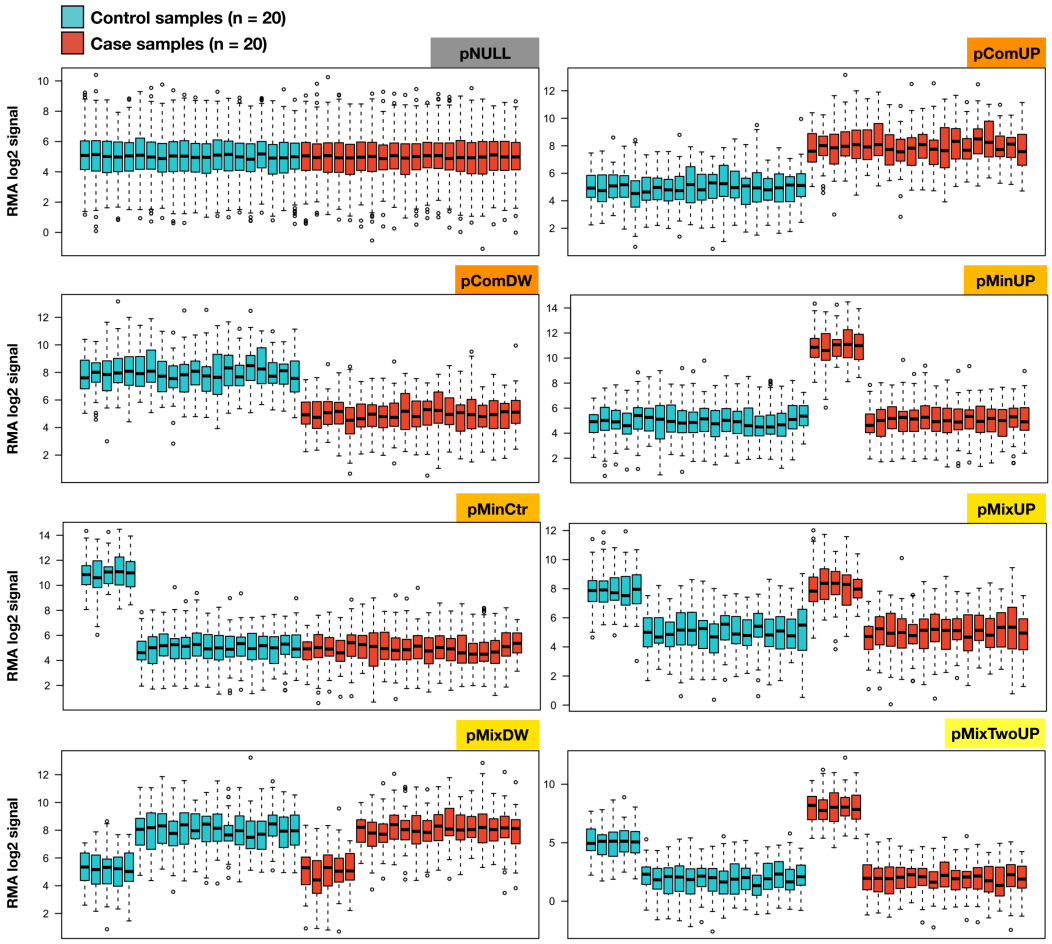


Figure 1-M-1. Artificial ***supervised benchmark*** (40 samples, 10000 features/genes) simulating a RMA log2 gene expression signal, generated with *optBiomarker* R package. **(pComUP, pComDW, pMinUP, pMinCtr, pMixUP, pMixDW and pMixTwo)** Boxplots of 7 positive patterns (4 model-type profiles) generated for both categories (control and case samples) within the whole dataset: pComUP-DW are *complete* changes; pMinUP-Ctr are *minority* changes; pMixUP-DW are *mixed* changes; and pMixTwo is a particular *mixed/minority* change where outlier samples are more changed in case samples. **(pNULL)** Boxplot per sample of rest of dataset, which include features showing no differential expression.

1.2 Experimental transcriptomic datasets

In order to validate and assess DECO's performance and results, five different cancer-related datasets from two different platforms (Affymetrix microarray and Illumina HiSeq) were used (**Table 1-M-2**).

Table 1-M-2. Experimental datasets used to assess current methods to find outliers and DECO.

Disease	Tissue	Platform	Number of samples	Normalization	Subtypes	Experimental design	Control	Case	Reference	Year
Osteorsarcoma (OSC)	Tumor biopsy	Affymetrix Human Gene 1.0	21	RMA log2(signal)	Metastatis, No-metastasis	Supervised (two classes)	No-metastasis	Metastasis	http://bioinfow.dep.usal.es/osteosarcoma	2016
Myelodisplastic Syndrome (MDS-1)	Bone Marrow CD34+ cells	Affymetrix Human Genome U133 Plus 2.0	41	RMA log2(signal)	RAEB1-MDS, RAEB2-MDS	Supervised (two classes)	RAEB1	RAEB2	GSE19429 (GEO database)	2010
Myelodisplastic Syndrome (MDS-2)	Bone Marrow Mononuclear cells (BM-MNCs)	Affymetrix Human Genome U133 Plus 2.0	24	RMA log2(signal)	No leukemia, Low risk MDS	Supervised (two classes)	No leukemia	Low risk MDS	GSE13159 (GEO database, MILE study)	2011
Breast Cancer (BCC-2)	Tumor biopsy	Affymetrix Human Genome U133A	285	RMA log2(signal)	PAM50 (Basal, HER2+, Luminal A and B)	Unsupervised (no classes)	-	-	GSE125055 (GEO database)	2011
Breast Cancer (BCC-2)	Invasive Ductal (ID-BCC) and Invasive Lobular (IL-BCC) carcinomas	Illumina HiSeq	596	log2(RPKM+1)	PAM50 (Basal, HER2+, Luminal A and B); ID, IL-BCC	Unsupervised (no classes)	-	-	Ciriello et al.	2015

The first three datasets correspond to heterogeneous pathologies in which there was no clear gene signature (between subtypes referred in **Table 1-M-2**) described in the literature. In this way, we included: **(i)** an osteosarcoma dataset (**OSC**) including samples from primary tumour biopsies from 21 patients that were treated and followed in the same way, where some of them (n=12) never showed metastasis after treatment but others (n=9) suffered metastasis from the primary tumour; **(ii)** a myelodysplastic syndrome (**MDS-1**) dataset of CD34+ selected cells from bone marrow of 41 patients suffering two closely related MDS subtypes (RAEB1 n=21 and RAEB2 n=20); **(iii)** another myelodysplastic syndrome dataset (**MDS-2**) of mono-nucleated cells from bone marrow (BM-MNCs) of donors that did not have any kind of dysplasia or leukaemia (n=11) and patients with low-risk prognosis MDS (n=13). This subtype of low-risk MDS patients is quite difficult to distinguish in the clinic from individuals with non-malignant anaemias. Interestingly, OSC, MDS-1 and MDS-2 datasets were also considered because classical methods for differential expression between classes did not find any significant result for the subtypes compared (*Subtypes* column, **Table 1-M-2**). Furthermore, we applied DECO on two large transcriptomic datasets obtained from biopsies of breast cancer patients (BCC-1 and BCC-2 in **Table 1-M-2**).

1.3 Data pre-processing

Our method DECO is designed to support any type of omic feature data properly normalized. In this dissertation, we use gene expression data as it is one of the most frequently analysed genome-wide features in published genomic studies. For expression

datasets, the data matrices have to be previously normalized and log2 scaled using any of the well-established methods. RMA normalization method was applied here when data came from *Affymetrix* microarrays platforms (Irizarry et al. 2003).

For RNA-seq data (produced with *Illumina* platforms), TPMs/FPKMs/RPKMs matrices may be the input to DECO after log2 transformation ($log2(TPMs/FPKMs/RPKMs + 1)$); if read counts matrices are available, the *voom* normalization method (Law et al. 2014) or another read counts normalization method should be previously applied.

2. Benchmark of the experimental datasets

In order to avoid well-controlled situations, we used several experimental genome-wide expression datasets from different sources, platforms and cell types (**Table 1-M-2**). In the case of microarray data, the probe-sets were previously mapped to ENSEMBL genes with *GATExplorer* CDFs (Risueno et al. 2010) and normalized with RMA method (Irizarry et al. 2003).

Three methods were implemented to evaluate the results from each algorithm with the experimental datasets. First**, GlobalTest** (Goeman et al. 2004) (implemented in a R package) was used as a test of outcome and response based on the lists of candidate significant gene markers selected by each method. This test delivered the percentage of well-classified samples, the values of GlobalTest statistic and the p-value for the response. Second**,** Principal Component Analysis (**PCA**) (Mardia et al. 1979) also run with the lists of selected gene markers provided by each method and delivered the percentage of samples correctly separated in the principal dimension. Third**,** a **Support Vector Machine** (SVM) method (included in *e1071* R package, (Meyer et al. 2014)) was applied, using a leave one-out sample procedure, to predict the assignment to class of each sample using the lists of gene markers selected by DIDS or by DECO to compare their classification precision.

3. DECO: workflow

DECO provides a comprehensive analysis about heterogeneity present in an omic dataset, describing the dependence between biomarkers and samples. Shortly, this method is composed of three main steps:

1st: Recursive Discriminant Analysis (**RDA**).

2nd: Non-Symmetrical Correspondence Analysis (**NSCA**).

3rd: Integration in a unique and simple heterogeneity ***h*-statistic**.

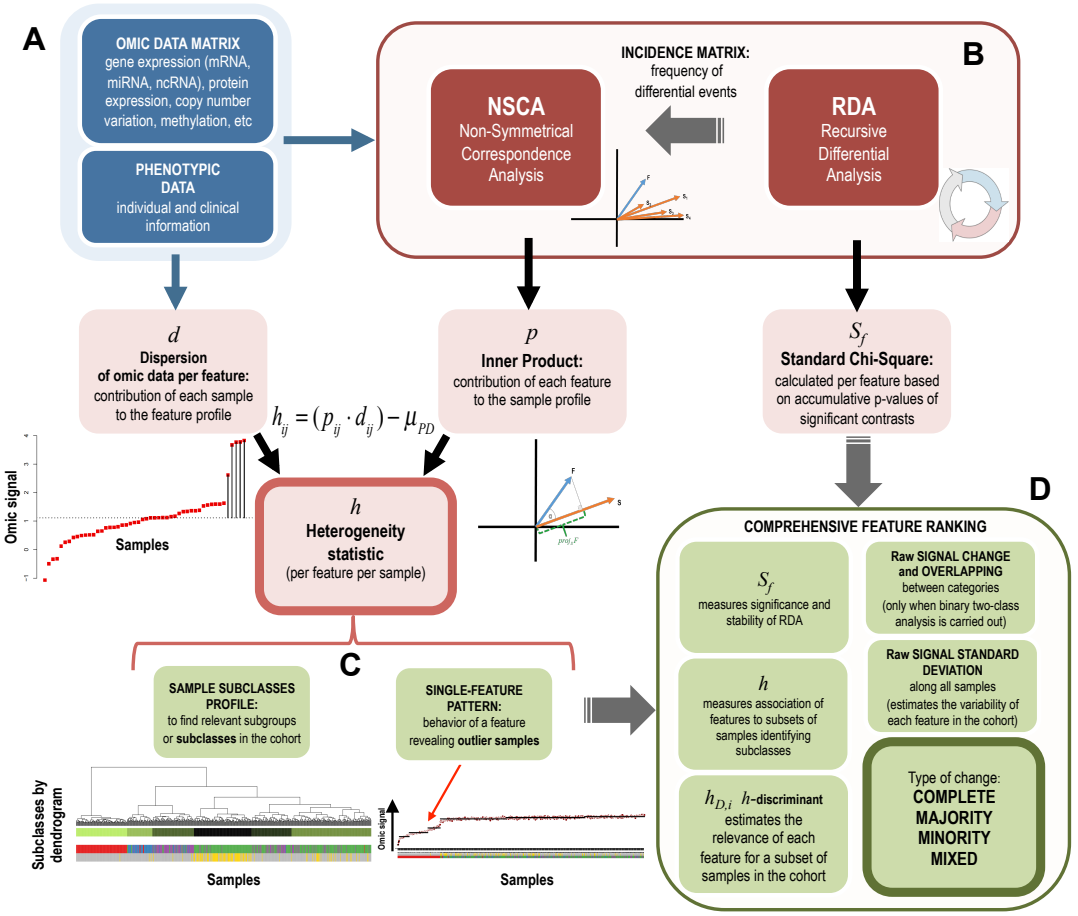


Figure 1-M-2. (A) Description of the dataset to be analysed, where the main parts will be the *data matrix* and the *phenotypic data* that includes all the known characteristics about each sample (e.g. clinical, phenotypical and personal characteristics). **(B)** Two main parts of the algorithm: RDA (Recursive Differential Analysis) and NSCA (Non-Symmetrical Correspondence Analysis). **(C)** The main parameter calculated by DECO, called *heterogeneity statistic* (*h*-statistic), which is determined for each feature in each sample. **(D)** The last step of the algorithm produces a feature ranking based in the values of the *Standard Chi-Square* (S_f) and the *h-statistic discriminant* (h and $h_{D,i}$, measuring the association of features to subsets of samples and estimate the relevance of such association within the cohort to identify possible subclasses).

Accordingly, DECO will recursively explore all the dataset to find out meaningful heterogeneity or variability. Thus, the main results provided by the analysis would be:

(i) Full characterization of biomarkers profiles (**Fig. 1-M-2C**), associated with main classes compared in the study or biomarkers significantly related to other phenotypes. If supervised analysis, they would be categorized in 4 model-types (**Fig. 1-I-7**).

(ii) Feature ranking based on combined parameters (**Fig. 1-M-2D**).

(iii) Subgroups of samples based on h-statistic (**Fig. 1-M-2C**) which show significant variation among individuals (*de novo* subgroups or related to other phenotypical information).

(iv) Possible errors in class label assignment (*mixed* profiles may point to this scenario).

(v) Possible sample and gene outliers.

4. DECO part 1: Recursive Differential Analysis (RDA)

Once we have our omic data properly normalized, DECO will apply the LIMMA algorithm (Smyth 2004) iteratively to perform a Recursive Differential Analysis (RDA), using the empirical Bayes (*eBayes*) method from *limma* R package as basis for differential analysis between samples. This first RDA step allows three types of experiment designs to compare samples, that involve several considerations and which will be referred along of this Chapter I:

(a) binary or supervised comparisons (2 groups) where user wants to compare two categories of samples, control and cases.

(b) multiclass comparisons (> 2 groups) where there are several groups of samples to compare and user wants compare all separately. Here, each group of samples will be iteratively compared against all others.

(c) unsupervised comparisons (no groups) where all samples are compared globally without any predefined category or class.

4.1 Granularity of RDA

RDA step will follow a *balanced resampling without replacement* strategy. Given a omic data matrix and a vector indicating group of samples (or not), an optimum sub-sampling size for each subset of samples must be determined. By default, it has been defined as r **(Eq. 1.3)**, corresponding to the closest integer to the square root of $\min(n_1, n_2, n_3, \dots)$ (for a supervised analysis with classes) or n (for an unsupervised study without classes). This optimum r is the sample size of each subset was determined following the approach of consistency in a variety of resampling situations published by Babu (Babu 1992). Additionally, r could also be defined directly by the user.

$$r = \sqrt{min(n_1, n_2, n_3 \dots)} \qquad \text{(Eq. 1.3)}$$

RDA step is primarily conditioned by the contrast design and r subsampling size. The ability to highlight the different theoretical profiles (**Fig. 1-I-7**) will vary depending on how the analysis was focused on classes or individual samples, what we called the **granularity of RDA**. In this way, it is important to mention that two main objectives could be followed: **(i) a majority** one, which will highlight major, stable but sometimes with lower fold change differences by setting r higher; and **(ii) a minority** one, which will search major and minor differences with higher fold change (**Fig. 1-M-3**).

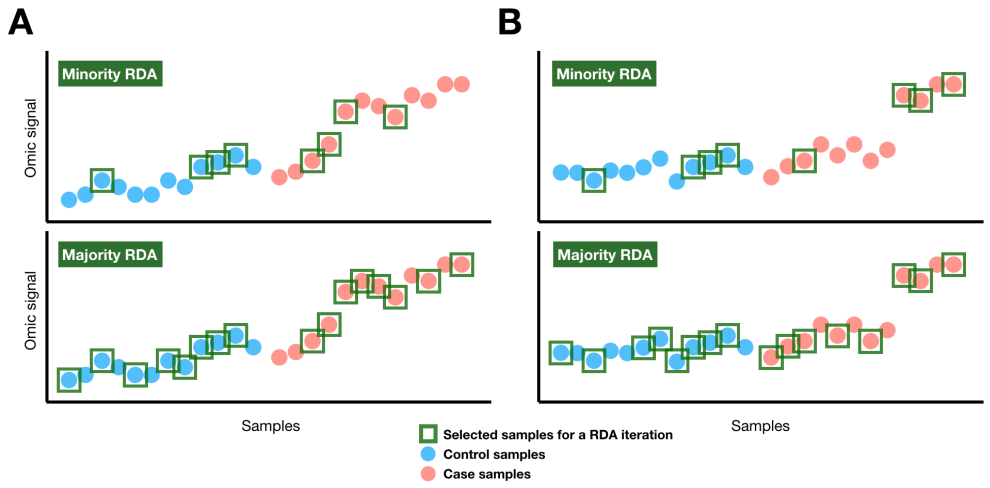


Figure 1-M-3. Theoretical profiles describing one "majority" (A) and one "minority" change (B). The ability of RDA to find out different profiles is directly related to subsampling size (r). **(A)** Majority profile described in Fig. 1-I-7 within RDA step, where green squares would correspond to selected samples for any iteration. Both subsampling sizes (r=4 at upper panel and r=8 at bottom panel) of RDA are more likely to discover this profile. **(B)** Minority profile described in Fig. 1-I-7within RDA step. Here, only lower subsampling size (r=4 at upper panel) would be able to select this profile as positive, while higher subsampling size would miss it due to mean differences between groups of samples.

Once the optimum sub-sampling size has been calculated, the algorithm will randomly generate sample combinations to be tested (by default it will be 10000 random subsets from all the possible sample combinations). Finally, a significant p-value threshold must be defined to consider any differential event as positive (by default to *adjusted p-value* < 0.01) from each output of *LIMMA* (Benjamini and Hochberg 1995).

4.2 Frequency matrix: counting differential events

Once all comparisons have been calculated using LIMMA, DECO will associate feature and samples by counting differential events (DEV): given one comparison between a subset of samples, any feature m showing a significant difference would be associated with these samples. Thus, the algorithm constructs a vector of *p-values* for each significant feature, and a frequency matrix A of features per samples counting the number of times that each feature was significant when a given sample was included in the comparison (**Fig. 1-M-4**). If a binary comparison was previously set up (comparison of two defined classes), up and down changes would be separated per feature within the A frequency matrix, so it would include $2m$ rows of differentially expressed features and n columns of samples. **Equation 1.4A** is used for the *supervised* experimental design, where n_{ij} is the number of repeats that feature i is differentially changed in sample j. In the case of *unsupervised* or *multiclass* design, only the up-changes are counted and a matrix A with m rows is generated **(Eq. 1.4B)**:

$$A = \left(n_{ij}\right)_{2m \cdot n} \qquad \text{(Eq. 1.4A)} \qquad A = \left(n_{ij}\right)_{m \cdot n} \qquad \text{(Eq. 1.4B)}$$

4.3 Summarizing differential events per feature

Next, DECO will calculate a statistical parameter X_f which summarizes the number of positive DEVs, called R, through the *adjusted p-values* of these comparisons following the Fisher's combined probability test (Fisher 1925) **(Eq. 1.5)**.

$$X_f = -2 \cdot \sum_{i=1}^{R} ln(adj.p.value_i) \qquad \text{(Eq. 1.5)}$$

Giving that the parameter X_f follows a Chi-square distribution with $2R$ degrees of freedom, DECO can calculate a new p-value, called here *p-value*(X_f), to identify the

features that have higher significance in the complete differential analysis. This new p-value is also adjusted by the Holm or FDR method (Holm 1979, Benjamini and Hochberg 1995). Finally, standardized values of X_f are calculated in order to compare X_f across different features under the same computational parameters described above **(Eq. 1.6)**.

$$S_f = \frac{X_f - 2R}{\sqrt{4R}} \qquad \text{(Eq. 1.6)}$$

This final statistic S_f provides a robust parameter to rank the differential features according to their significance along the performed RDA.

4.4 Double repeat-threshold

Since RDA would iteratively explore all significant variability inside a omic dataset, there may be a subset of features amounting to a low number of repeats for a large number of samples or vice versa. Thus, a **heuristic and double repeat-threshold** can be applied in order to reduce the number of features susceptible to be the input of next step (NSCA). This threshold has been lately developed and it is composed of two parameters: **(i)** a number of repeats or differential events, called *rep.thr*; and **(ii)** the percentage of samples amounting this number of repeats, called *samp.perc*. Further information about this double **optional filter** and results of its application are provided in Section 9.3 of Results.

5. DECO part 2: Non-Symmetrical Correspondence Analysis (NSCA)

Given a *frequency matrix* A, DECO will apply a Non-Symmetrical Correspondence Analysis (NSCA) (Diday 1984) in order to explore and analyse the patterns and associations between features and samples. If the analysis design is *binary*, including two classes, a separated NSCA for each class will be performed to distinguish the differential events (up or down) that occurred in each class. Instead, a unique analysis will be done for unsupervised and multiclass designs. Further theoretical information about NSCA detailed in Section 5 of Introduction.

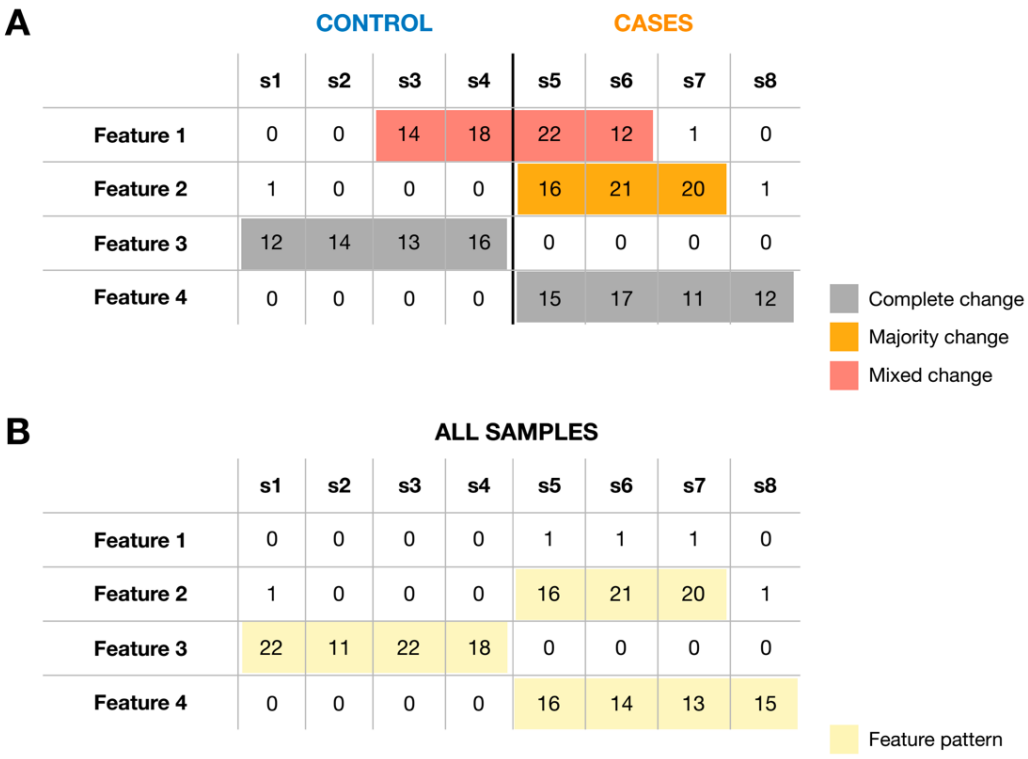


Figure 1-M-4. Theoretical example of how frequency matrix *A* is generated and interpreted. Given a subset of samples for each RDA iteration, DEVs would be annotated to *positive* features. **(A)** Binary or supervised example where different profiles could be found. **(B)** Unsupervised or multiclass examples where only up-regulated differential events are annotated. Feature pattern will generate subgroups of samples after NSCA analysis.

The algorithm calculates the inner product matrix P **(Eq. 1.7)** of the feature and sample vectors in the NSCA space. Column isometric factorization, based on singular value decomposition of matrix A, provides a common $n\text{-}1$ dimensional space to infer individual feature profiles. The inner product between a feature and sample point measures the strength of the asymmetric association (Section 5 of Introduction). Thus, the higher the inner product is, the more dependency of differential feature signal from sample presence could be inferred (**Fig. 1-I-8**). Following Beh and Lombardo notation (Beh and Lombardo 2014) (where *OF* are feature vectors and *OS* sample vectors, **Fig. 1-I-8**) the inner product matrix P can be expressed as:

$$P = \left(p_{ij} = \|OF_i\| \cdot \left\|OS_j\right\| \cdot \cos \alpha_{ij}\right)_{m \cdot n} \qquad \text{(Eq. 1.7)}$$

The inner product matrix P improves the characterization of the feature profiles in complex heterogeneous sample sets in comparison with raw omic data. Additionally, raw

omic data E **(Eq. 1.8)** was used to complement NSCA relational information from P matrix, so a dispersion matrix D including distance from the mean value of each class (if supervised) or of all samples (if unsupervised or multiclass) per feature is calculated **(Eq. 1.9)**:

$$E = \left(e_{ij}\right)_{m \cdot n} \quad \text{(Eq. 1.8)}$$

$$D = \left(d_{ij} = e_{ij} - \bar{e}_{i\cdot}\right)_{m \cdot n} \quad \text{(Eq. 1.9)}$$

All values from both D and P matrices were standardized to have all changes on the same scale. Besides, due to higher values of p were assigned to relevant correspondence among features and samples, P matrix scale was shifted placing the minimum value to 1 to avoid penalization of negative inner product p values.

6. DECO main statistical parameter: *h*-statistic

Next, we combined those new *standardized* scores of D and P, called D^S and P^S, in a heterogeneity statistic or ***h*-statistic** which is intended to reflect both directional influences among features and samples: the relevance of a feature in a sample profile given by $P^S(p_{ij})$ and the relevance of a sample in a feature profile given by $D^S(d_{ij})$.

$$H = \left(h_{ij} = d_{ij}^s \cdot p_{ij}^s - \mu_{D \cdot P}\right)_{m \cdot n} \quad \text{(Eq. 1.10)}$$

In brief, given a feature profile from this H matrix, absolute higher ***h*-statistic** values would correspond to those samples where: **(i)** omic signal is differentially changed (given by D) and **(ii)** the feature has relevance for the classification of those samples given our omic data matrix (given by P).

7. DECO: sample stratification based on *h*-statistic

The global feature profiles derived from the differential events and summarized by the matrix H were used to find subclasses in the sample set. A distance *Pearson correlation* matrix C (Szekely et al. 2007) between sample h vectors is calculated (Eq. 1.11) and hierarchical clustering is applied to group samples into subclasses.

$$C = \left(c_{ij} = 1 - corr(h_i, h_j)\right)_{m \cdot n} \quad \text{(Eq. 1.11)}$$

Later, an iterative function calculating the *Pearson* version of Hubert's gamma coefficient, or γ (*cluster.stats* function from *fpc* R package), per k clusters identifies the highest value for optimal cutting of dendrogram (*cutree* function from *stats* R package). Thus, an optimal number of k clusters-subclasses will be set up **(Eq. 1.12)**:

$$k = \{i : \gamma_i \; is \; maximum\} \quad \text{(Eq. 1.12)}$$

8. DECO: feature profile characterization and ranking

8.1 Overlap statistic

In the case that a *supervised* or *multiclass* design is set up, DECO is able to identify and discriminate which features conform a reliable response for class comparison; and in this way, it is able to segregate specific subtypes of samples within a class or within all the dataset, according to profiles shown in **Figure 1-I-7**. Our method bases this feature classification on how omic signal overlaps between categories. In a formal description, let $e_i = \langle e_{i1}, e_{i2}, \ldots e_{in} \rangle$ (where $(i=1,2,\ldots,m)$) be the raw omic data vector from the matrix *E* per feature. Then, DECO calculates the density function $f(e_i)$ per category to find out how many parts per unit overlap based on approximate integration of the common area under all curves (*sfmisc* R package).

$$O = \int_{-\infty}^{\infty} \left[f\left(e_{i,m_1}\right)\right] * f\left(e_{i,m_2}\right) de \quad \text{(Eq. M-11)}$$

where $*$ denotes complex conjugation. Consequently, features can be assigned to 4 model-types as follows: **(i)** ***complete*** change (pCom), for well-separated feature profiles commonly found by standard methods of differential expression (fixed to $o_f < 0.2$ by default); **(ii)** ***majority*** change (pMaj), for features that show a major change between classes ($0.2 \leq o_f < 0.4$); **(iii)** ***minority*** change (pMin), for features that mark a specific subclass within a category of samples ($0.4 \leq o_f < 0.8$); and **(iv)** ***mixed*** change (pMix) for features whose differential events are not directly related to the compared categories, but

to subclasses within them (0.8 ≤ o_f ≤ 1.0). Example of plot representing this estimation is provided in **Figure 1-R-15**.

8.2 Ranking based on parameters combination

Finally, DECO produces a global ranking of the most relevant features that mark the samples studied based on the average rank obtained from the three main parameters calculated by the algorithm: **(i)** *Standard Chi-square* value S_f, which highlights the most significant changes from RDA; **(ii)** h-statistic range per feature, which indicates how discriminant each feature is, given the subclasses found by NSCA; and **(iii)** both o_f (overlap statistic) and standard deviation of raw omic signal in each differential feature, assessing the variability along samples to allow finding the most stable features that will be consider the best markers for the classes or subclasses found.

RESULTS

Along this Results section, we will highlight: **(i)** the results obtained from the comparison of our algorithm DECO in comparison with several current and classical methods for differential expression or outlier profile detection (using simulated and experimental datasets); **(ii)** the specific application of the algorithm to two large breast cancer datasets; and **(iii)** all issues related to the implementation of the method as an R package.

1. DECO outperforms state of the art methods for finding outliers

Several artificial datasets simulating gene expression data and including different subsets of genes that have differential changes were used (detailed in Section 1 of Materials and Methods) to compare the new method DECO with 7 established methods (described in **Table 1-M-1**). Following this benchmark, we will assess the ability of the RDA first step of DECO to detect *outlier* profiles of each model-type and to score them using *Standard Chi-square*. It is important to remind that *supervised* experimental design is the most common for differential analysis, contrasting any category of interest against a control/reference category of samples. This section will be focused on *supervised* analyses.

All previous methods compared were also applied to these simulated datasets, returning their own scoring statistic for all genes: each method is based on a single statistic of relevance per feature, which is used to rank all genes and prioritize *outlier* profiles. Here, we compared these statistics against our *Standard Chi-square*, which summarizes differential expression events found along the subsampling step or RDA (r = 3, iterations = 10000, *adjusted p-value* = 0.01). **Figure 1-R-1** shows the results of first of these comparisons for two studied types of model-type profiles (according to **Fig. 1-I-7**). Here, the simulated data followed normal distributions, fitting similar distributions as the tested ones by current methods for *outlier* profile detection in their original studies.

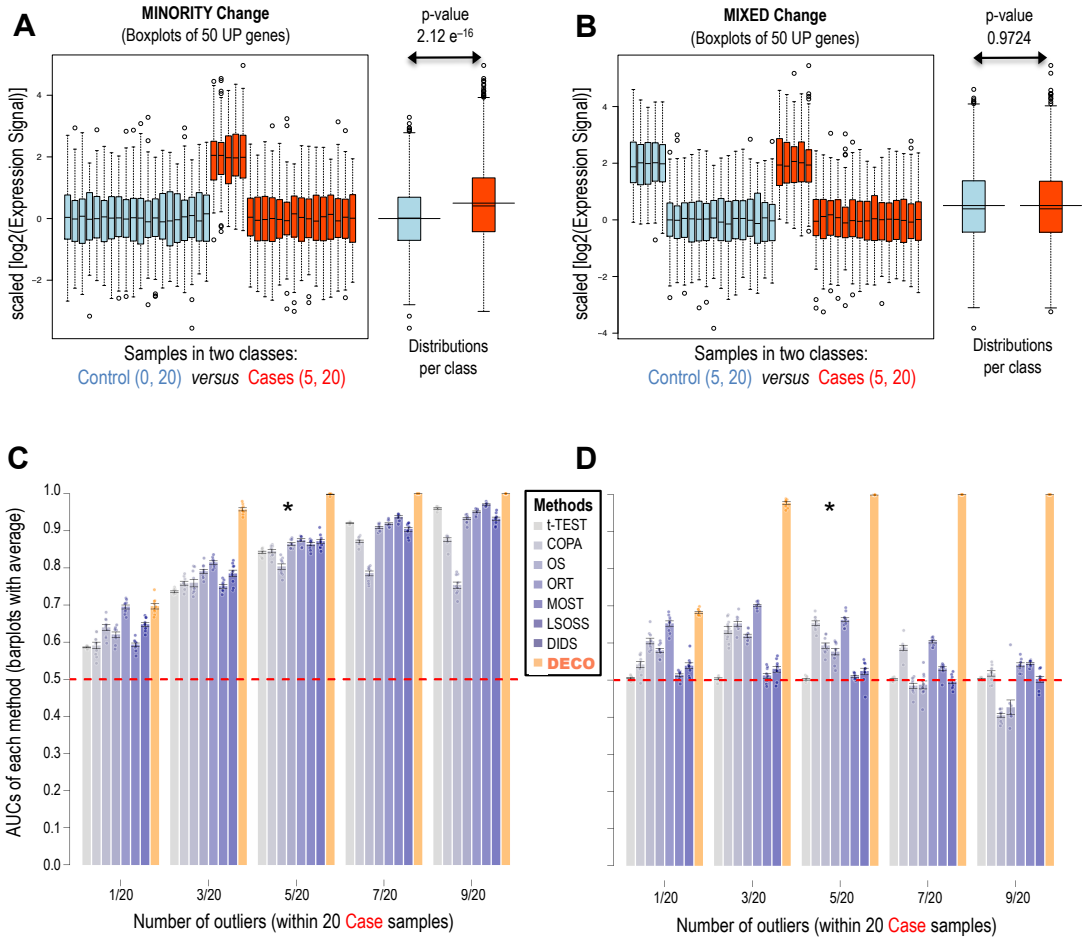


Figure 1-R-1. Comparison of 8 methods (t-Test, COPA, OS, ORT, MOST, LSOSS, DIDS, DECO) used to find significant changes that occur in a minority of samples and in a small proportion of features (≈10%). (**A-B**) Boxplots of artificial expression signal (T matrix) simulating complex DE for a subset of samples (minority profiles and mixed profiles). Both panels also show *p-value* after t-test comparison between group distributions. (**C-D**) Barplot representation including mean and error bars of AUCs after TPR and FPR calculation (10 random simulations). All 8 methods and 5 different number of outlier samples ($\eta \in \{1,3,5,7,9\}$) are represented.

For this purpose, the two most complex model-types detailed in **Figure 1-I-7** were included in this benchmark:

(i) **Minority change**: within case group but not in the control samples (**Fig. 1-R-1A**). We will vary the percentages of samples from case group corresponding to *outlier* profile (from 5%, 1 outlier sample in 20 samples, to 45%) (**Fig. 1-R-1C**)

(ii) **Mixed change**: occurring both within case group and control group (**Fig. 1-R-1B**). It is also explored using different percentages of changed genes (from 5%,

1/20, to 45%, 9/20) (**Fig. 1-R-1D**).

Given an *outlier* sample size (k), the Areas Under the ROC Curves (AUCs) were calculated after 10 random simulations of all the procedures (Section 1 of Material and methods). Mean AUCs and standard error were represented for both comparisons described above: (i) & (ii). The results show that DECO provided better AUCs than other 7 methods along of all scenarios tested (**Fig. 1-R-1**, panels C-D). For the *minority change* case (i), we expected that all methods were increasing their AUCs along higher *outlier* sample sizes (panel C). However, DECO notably outperforms other methods since 3/20 *outlier* sample size was set-up. For the *mixed change* case (ii), we also expected that current and classical methods fail to detect those *outlier* profiles since they are not considering control as a source of heterogeneity (panel D). Indeed, AUCs' values from these methods range around random expectation for TPR and FPR (AUC = 0.5), showing that current methods rely on the expectation that the control samples are stable and they should not suffer anomalous changes. Shortly, the numbers showed that having at least 3 outlier samples (3/20) or more (5,7,9/20) DECO achieves a very good performance giving AUCs > 0.90.

Since the transcriptomic omic data does not follow a normal distribution, we also compared all those methods for *outlier profile* detection using a modified version of the *optBiomarker* R package (Khondoker et al. 2010). This package was developed to simulate RMA distributions for differential expression analyses. Although it was thought to simulate differential expression between two categories (*complete* changes), we accordingly modified this R package to include the 4 model-types detailed above (Chapter 1-I Section 1). Once we modified it, we assembled a wide benchmark including 4 model-types to compare how different methods score and rank each feature. This benchmark included simulated RMA gene expression signal for 10000 features/genes along 40 samples (20 control and 20 cases), where 7 differential scenarios (k = 5) of 50 features/genes were included (detailed in **Fig. 1-M-1**):

(i) **pNull: no differential expression.** It is composed of the rest of the feature/genes (9650g) non-included in any differential pattern.

(ii) **pComUP and pComDW: complete change.** Up and down changes for all samples of each category.

(iii) **pMinUP and pMinCtr: minority change.** Up changes for 5 samples within each

category (pMinUP for case samples and pMinCtr for control samples).

(iv) **pMixUP and pMixDW: mixed change**. Up and down changes for 5 samples within each category.

(v) **pMixTwo: special mixed/minority change.** Up changes for 5 samples within each category, where change is higher for *outlier* samples in case samples. This profile would suppose higher variability of control samples, but still a significant change for *outlier* samples in case samples.

Consequently, we compared the scoring statistic of each method for features belonging to different 8 patterns detailed above and, again, DECO performance was based on *Standard Chi-square* statistic from RDA step. According to original hypothesis of *cancer outlier profile*, we expected that current methods for *outlier profile* detection score *minority* patterns (pMinUP and pMinCtr) higher than other patterns, while t-test and DECO do the same for *complete* changes (pComUP and pComDW). **Figure 1-R-2** represents all gene scores per pattern obtained after applying all these methods. In the figure, each plot corresponds to one method split by the 8 different patterns, where pNULL's segment would correspond to a random subset of 50 gene scores obtained for the whole pNULL profile (9650 features/genes). Interestingly, DECO only disclosed 31 out of 9650g (0.32%) as significant in any iteration of RDA step and, consequently, it was scored by *Standard Chi-square*.

As expected, this benchmark (**Fig. 1-R-2**) was very informative about performance and real assumptions done by each method because it really reflects the particular strengths and weaknesses. Attending to the original *outlier* model-type described by COPA (*minority* change: pMinUP), we can observe a variety of behaviours: while COPA and OS scored pMinUP (up in case samples) as the most significant ones, the other methods (included DECO) prioritized *complete* changes giving them a higher score. We expected this result from DECO and t-test but not from the other methods, specifically designed for *outlier profile* detection. This trend may be confusing for all those who applied these methods on their omic data awaiting maxima scores for *outlier profiles*.

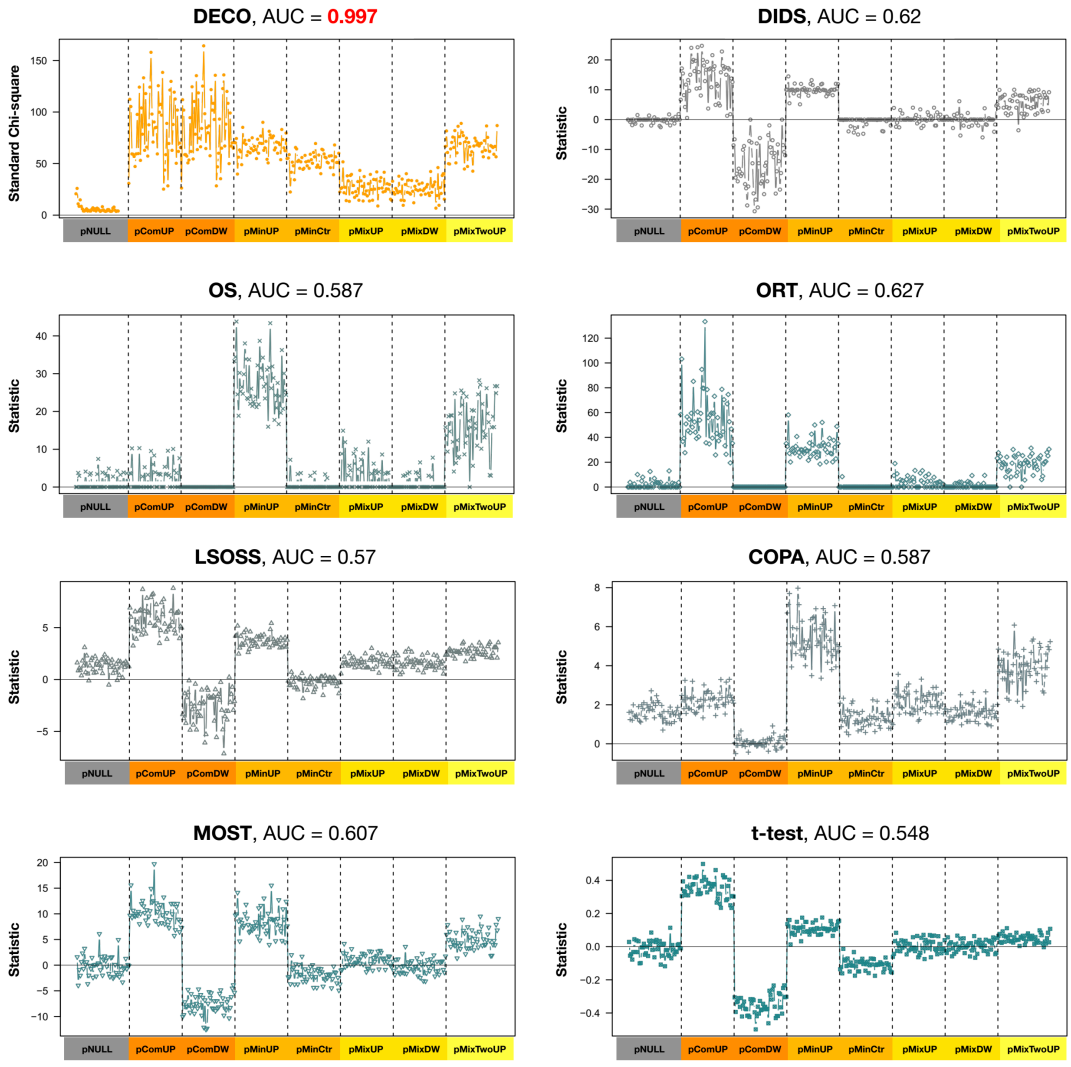


Figure 1-R-2. Plots representing the scoring statistics of each method compared for 8 patterns of simulated benchmark generated with the modified *optBiomarker* R package. Each plot is segmented by patterns (50 gene scores each), where pNULL segment represents 50 random gene scores taken from all 9650 features composing the whole pNULL pattern. AUC calculation (TPR/FPR) is based on scores of differential patterns against whole pNULL (no differential expression).

Additionally, we detailed along the description of these methods (Section 2.3 of Introduction) that they assumed stability for control or reference category, ignoring possible *outlier* samples included within this group. This assumption it is also reflected by pMinCtr scoring, where all methods (except DECO and t-test) were not able to discriminate it from pNull (no differential expression, True Negative features). In this way and trying to provide a meaningful insight of this assumption, we also included the pMixTwo pattern. This pattern

followed a *mixed/minority* profile where there were *outlier* samples in both categories, but case *outlier* samples have even higher expression than control *outlier* samples (**Fig. 1-M-1**). Given this pMixTwo pattern, we aimed to test if a greater variability of control samples is enough to discard these *outlier* profiles as significant. As we can observe in **Figure 1-R-2**, all methods except t-test scored them greater than pNull but lower than pMinUP (original profile mentioned by COPA). Interestingly, our method DECO gave a similar score to pMinUP/pMinCtr/pMinTwo since they are *outlier* minority profiles, solving any effect derived from control or reference samples variability.

Apart of the scoring statistical comparison, it is important to mention that DECO was the only method able to disclose all relevant features (100% of differential genes/features were scored) behind a heterogeneous omic dataset, integrating both classical (SAM, LIMMA, t-test) and *outlier profile* detection approaches. All these significant features have been highly scored (AUC = 0.997) and ranked by *Standard Chi-square* following a logical sequence: *complete*, *minority* and *mixed* profiles. However, if we apply methods for *outlier profile* detection or classical methods, we will only recover a part of the meaningful heterogeneity (**Fig. 1-R-2**), with a lower performance ratio as we showed in the previous benchmark (**Fig. 1-R-1**).

2. Accurate detection of different feature profiles in a large-scale dataset through RDA feature selection and h-statistic.

Once RDA feature selection step was demonstrated to outperform classical and current methods for outlier profile detection, in this Section 2 we will focus on how both RDA feature selection and posterior h-statistic are able to characterize subclasses within an artificial heterogeneous dataset. In any situation, the RDA step was intended to select all those features following one of the 4 model-types described above (**Fig. 1-I-7**) while h-statistic encourages sample stratification using those picked features.

For a correct interpretation and understanding of how sample stratification is improved by DECO, we built another simulated dataset framing a complex situation: 40 samples and 10000 features following a normal distribution, out of which 250 had significant changes following 3 different theoretical model-types (**Fig. 1-R-3**). These differential features, each containing 50 genes, followed 5 distinct profiles:

(i) Two patterns (**p1**, **p2**) of 50 genes following a ***complete change*** in all case samples against the control category. Features in p1 are up-regulated in control samples while features in p2 are down-regulated in controls.

(ii) Two patterns (**p3**, **p4**) of 50 genes changing in a ***minority*** (5/20 samples) of the samples, either in the cases or in the controls. Features in p3 are up-regulated in 5 samples of control category while features in p4 are up-regulated in 5 case samples.

(iii) A profile (**p5**) of 50 genes showing a ***mixed change*** in a 25% of the samples (5/20) in both categories.

The different profiles here presented can be observed in **Figure 1-R-3A**, which shows a heatmap illustrating the whole expression data matrix and the five profiles that affect to only a 2.5% of the 10000 genes. It is important to mention that the expression signals assigned to the genes in the simulated data included a random variability, following the same procedure for previous artificial dataset simulation (Section 1.1 of Material and Methods).

As a reference, we run the well-established and classical SAM and LIMMA methods on the expression data matrix described above, using as threshold a FDR (for SAM) or adjusted p-value (for LIMMA) ≤ 0.05 (Tusher et al. 2001, Smyth 2004). As was hinted along the Introduction of this Chapter 1, these methods were expected to find at least all features describing a ***complete change*** profile between categories, corresponding to 100 genes included in the profiles **p1** and **p2**. The aiming for SAM or LIMMA application was also showing how usual semi-supervised analysis differs from DECO application. A **semi-supervised analysis** would include a classical method for differential expression used to select features and a posterior unsupervised clustering to group samples. Thus, the user expects to group samples as original categories based on this subset of features. Interestingly, SAM was able to find those 100 genes under the significance threshold (result not shown in **Fig. 1-R-3**), but LIMMA found 112 significant differentially expressed genes (**Fig. 1-R-3B**). Such genes corresponded to: 95 genes with *complete change* profile (84.8%), 11 genes with *minority change* profile (9.8%) and 5 genes that do not have differential expression change (4.4% false positives) (**Fig. 1-R-3B**).

On the same data matrix, we applied DECO and the results are shown in 2 steps in

Fig. 1-R-3C and **D**. Firstly, the RDA step selected the most significantly changing features identifying 249 out of the 250 true positives. The heatmap in **Figure 1-R-3C** represents the result of the posterior hierarchical clustering using the raw expression signal of these 249 genes selected by RDA step. As can be seen, these genes are properly arranged in 5 feature profiles (**p1,2,3,4,5**) and the samples are classified in 6 subclasses (**c1,2,3,4,5,6**) according to their corresponding profiles (**Fig. 1-R-3C**). These subclasses within both case and control categories would not have been found if LIMMA subset of features was selected before the hierarchical clustering analysis (**Fig. 1-R-3B**). Therefore, the features selected by *LIMMA* only separate the two main known classes (cases and control) and even one sample of the cases appears in the heatmap misclassified.

Despite the classification of the 249 genes found in 5 profiles and the samples in 6 subclasses obtained using the raw expression signal of the features selected by RDA, the dendrogram displaying the similarity among samples (**Fig. 1-R-3C**) indicated that one of the subclasses (**c6**) did not have a distinct expression profile from the original distribution (**Fig. 1-R-3A**). This subclass shows values that represent small variations from the mean expression signal of the whole dataset. In this way, the samples within this subclass **c6** were poorly defined using the raw expression signal.

By contrast to this representation, **Figure 1-R-3D** shows a heatmap built with the parameter, h-statistic, derived from running the complete algorithm DECO (RDA + NSCA). The h-statistic was intended to improve subclasses separation through the integration of raw omic dispersion and predictor-response information (inner product from NSCA) as detailed above (Section 6 of Material and methods). After this integration, the h-statistic would provide more defined profiles to both selected features and samples. Thus, the **c6** subclass is now well-defined by an increment of **p2** profile and a decrease of **p1** profile.

It is important to note how **c1** and **c4** subclasses had a differential signal that comes from the *complete* changes (**p1** and **p2** respectively) plus another differential signal that comes from the *mixed* changes (**p5**) (**Fig. 1-R-3D**). For those samples, h-statistic values of **p1** and **p2** are lower than for other samples because p5 would exclusively identifies **c1** and **c4** subclasses while **p1** and **p2** are broadly associated with both categories. These kinds of profiles are not achieved using the raw expression signal, and so the results indicate that the h-statistic produced by DECO method is more powerful for sample characterization and stratification.

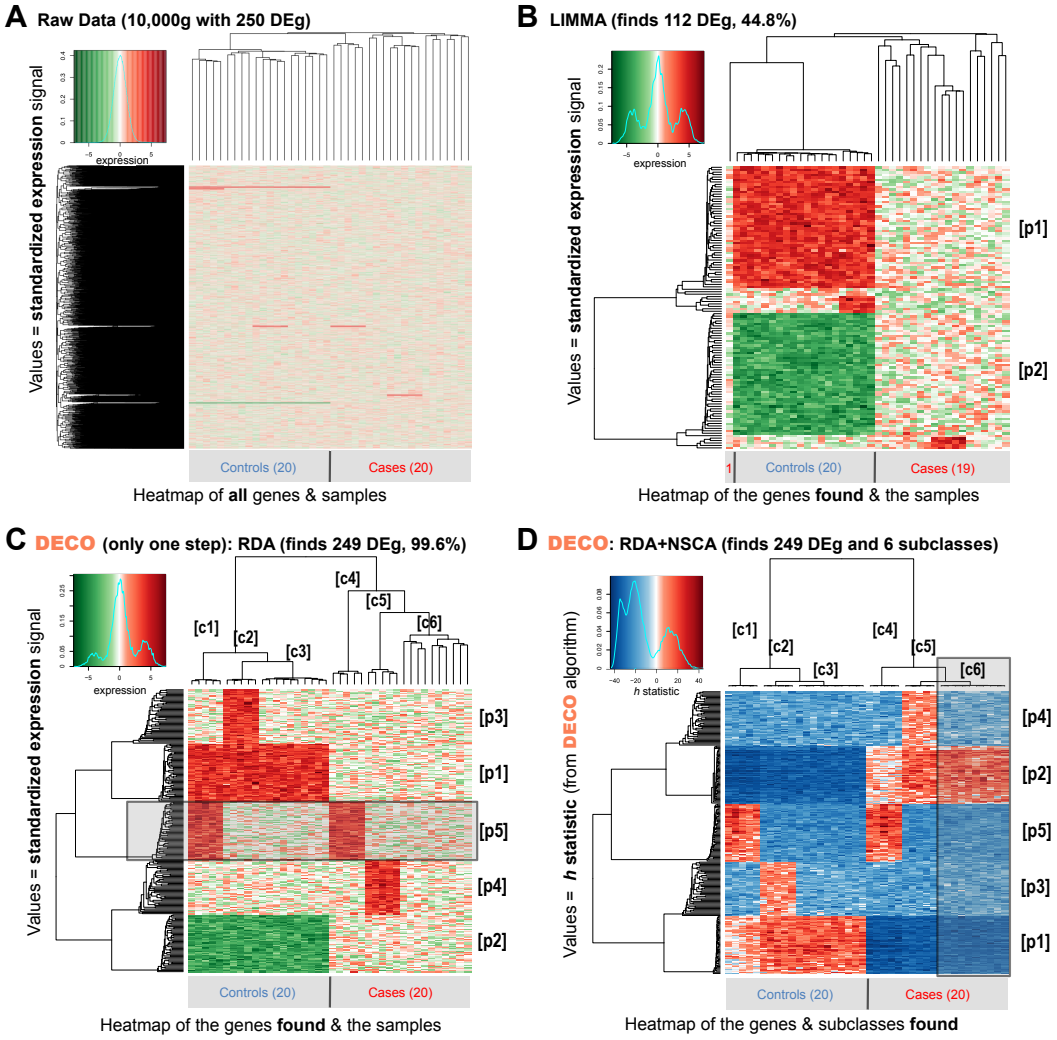


Figure 1-R-3. Analysis of a large simulated dataset that includes 20 cases *versus* 20 controls measuring 10000 genes, including 250 genes following the 3 model-types in different subsets of the samples. **(A)** Heatmap of the full expression matrix. **(B)** Heatmap of the expression data of 112 genes (DEg) found by *LIMMA* method along all samples. **(C)** Heatmap of the expression of 249 DEg found by RDA step. **(D)** Heatmap corresponding to h-*statistic* of all samples and 249 DEg found by DECO method (RDA+NSCA). The whole dataset includes 6 different sample subsets that were found by DECO and characterized according to their gene profiles in 6 "subclasses" **[c1, 2, 3, 4, 5, 6]**. The specific gene "profiles" identified were: **[p1]** profile including 50 genes UP in all controls with respect to the cases; **[p2]** profile including 50 genes DOWN in all controls with respect to the cases; **[p3]** profile including 50 UP only in 5 controls; **[p4]** profile including 50 UP only in 5 cases; **[p5]** profile including 50 UP in both 5 cases and 5 controls (5/20 = 25%).

In conclusion, the results obtained from this benchmark highlights two main advantages of using DECO algorithm: the strength of RDA to select significant variable features (**p1-p5**) and the ability of h-*statistic* to exclusively associate features and samples

(**c6**) within a particular dataset, enhancing the sample stratification. In the next section, further information and implications about how the *h-statistic* segregates subgroup of samples will be detailed.

3. *h*-statistic facilitates patient stratification.

Since the *h-statistic*'s relevance for patient stratification was hinted above, the particular effect on a single feature profile (i.e. any gene expression profile) produced by this integration of both omic dispersion and predictor-response information may be unclear yet. For this reason, we focused on a single and relevant feature profile to visualize the differences of using the raw omic signal (i.e. gene expression signal) or using the new statistical parameter: ***h-statistic***.

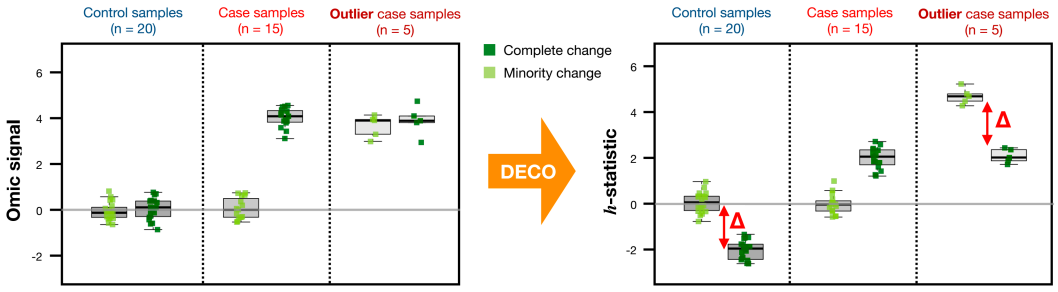


Figure 1-R-4. Boxplot visualization of *h-statistic* effect on two different profiles changing for the same samples (k = 5). The left panel shows artificial expression data (similar to Fig. 1-R-7) for a *minority* profile and for a *complete* profile, while the right panel shows *h-statistic* values for the same profiles. After applying DECO, the *h-statistic* will be increased (absolute values) respect to original omic signal for the *minority* profile on outlier samples, allowing the specific identification of these samples within the dataset, as well as *complete* profile's values for control samples.

It is important to remember that the *h-statistic* was intended to enhance patient classification through the predictor-response information from NSCA. Given a frequency matrix, NSCA is able to establish which features identify a particular subset of samples within the whole dataset. Its particular weight function (**Eq. 1.1**) and posterior singular value decomposition (SVD) provides a view of data structure from the original frequency matrix. In this way, NSCA finds out exclusively dependence structures, avoiding overlapping information if possible. Given two different features (e.g. *complete* and *minority*) accumulating DEVs for a specific subset of samples (**Fig. 1-M-3**), NSCA will prioritize

(higher inner product) the *minority* profile cause of its exclusive (less frequent) association to those samples.

Accordingly, the h-*statistic* will reflect the same behaviour weighting the feature's relevance for each sample but slightly corrected by omic dispersion data (**Eq. 1.10**). Then, if a single feature is showing both high omic dispersion (absolute value) and high relevance (by NSCA) within a sample profile, it will be even higher scored by h-*statistic* (**Fig. 1-R-4**). Consequently, it will encourage those outlier profiles which enhance the characterization of each sample within our dataset. For example, the panel D of **Figure 1-R-4** reveals how h-*statistic* values for **p2** pattern (*complete* change) and c4 subclass of samples are reduced in comparison with **p5** pattern (*minority* change).

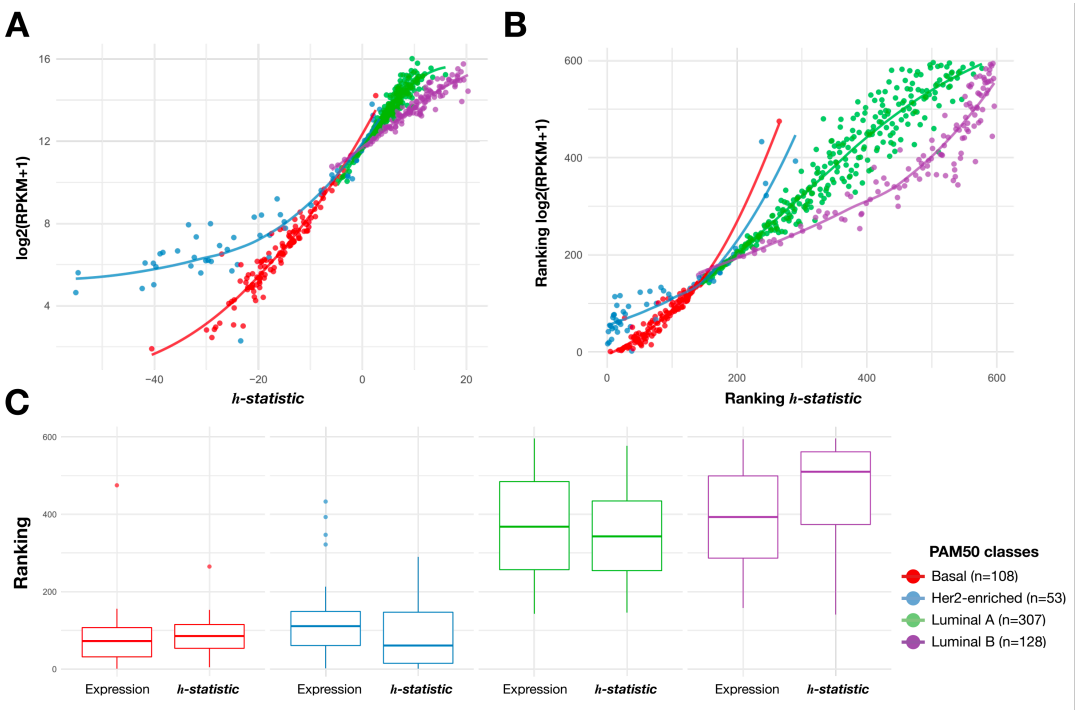


Figure 1-R-5. ESR1 single profile from TCGA Breast Cancer subset (BCC-2) after DECO analysis (r=7; R=1000) through a multiclass design based on PAM50 pseudo-classes (Basal, LuminalA, LuminalB, Her2-enriched). Panel **(A)** represents raw values of both expression data (log2(RPKM+1)) and h-*statistic*, describing different trends by PAM50 pseudo-classes. Panel **(B)** corresponds to the same representation after ranking both expression data and h-*statistic*. Panel **(C)** corresponds to a boxplot of ranked values per PAM50 pseudo-classes. For example, panel A, B and C show how h-*statistic* is able to separate Her2 samples (blue) from Basal samples (red) while raw expression data agglutinates both classes within the same interval, or LumB (purple) from LumA (green).

In order to evaluate how h-*statistic* behaves on an experimental dataset, we simply run DECO (r = 7; iterations = 2000) on a subset of TCGA Breast Cancer dataset (BCC-2

dataset from **Table 1-M-2**) following a *multiclass* experimental design based on PAM50 pseudo-classes: Basal (n = 108), Luminal A (n = 307), Luminal B (n = 128) and Her2 (n = 53). As we mentioned before, if a *multiclass* design is settled, DECO would maintain separated all samples from different subtypes along the RDA subsampling step, providing a clearer interpretation of results per subtype of samples than an *unsupervised* design. In this way, for this BCC-2 dataset, ESR1 was the first ranked gene profile by DECO, highlighting the relevance of the estrogen receptor (ESR1 gene) for the characterization and stratification of breast cancer samples. In **Figure 1-R-5**, we show the difference between the original gene profile based on the raw expression data (log2[RPKM+1]) and the h*-statistic*. Here, we can observe how h*-statistic* segregates the 5 pseudo-classes of breast cancer better than raw data, increasing (Luminal B or Basal) or decreasing (Her2-enriched or Luminal A) the values depending on how relevant is this change for these samples.

In conclusion, the h*-statistic* enhances the stratification of significant subtypes of samples through a numerical transformation to avoid overlap of different subtypes in the same range of values.

4. Identification of markers for disease subtypes in absence of global expression changes: tests on three clinical datasets.

Heretofore, we tested the ability of RDA and posterior h*-statistic* through analysis on simulated omic data. Because these results are not enough to expose how DECO would improve differential analysis of omic data, we also aimed to test and compare our method by means of an extensive comparison on real-experimental transcriptomic datasets. As previously demonstrated (Section 1 of Results), RDA step of DECO algorithm outperforms current and classical methods for differential expression and *outlier profile* detection. In this section, we extended the comparison to microarray transcriptomic datasets from different cancer pathologies. We selected three experimental datasets derived from clinical studies (**Table 1-M-2**), each one composed of two close subtypes of patients.

Interestingly, any of them showed significant differences in their global expression profiles after applying classical methods: SAM or LIMMA (*adjusted p-value* > 0.05) (**Fig. 1-**

R-6A). The three datasets used were: **(i)** an osteosarcoma dataset (**OSC**) including samples from primary tumour biopsies from 21 patients, where some of them (n=12) never showed metastasis after treatment but others (n=9) suffered metastasis from a primary tumour; **(ii)** a myelodysplastic syndrome (**MDS-1**) dataset of CD34+ selected cells from bone marrow of 41 patients suffering two closely related MDS subtypes (RAEB1 n=21 and RAEB2 n=20); **(iii)** another myelodysplastic syndrome dataset (**MDS-2**) of mononucleated cells from bone marrow (BM-MNCs) of donors that did not have any kind of dysplasia or leukemia (n=11) and patients with low-risk prognosis MDS (n=13).

Since the standard methods for differential expression analysis (SAM and LIMMA) did not report any differences, we tried DECO and other methods better suited for discovering subtle differences. Based on previous studies that compared COPA, OS, ORT, MOST and LSOSS methods for cancer outlier discovery (Karrila et al. 2011), we considered MOST as the best of them for different scenarios and used it for our comparison. Additionally, mCOPA (expanded version of COPA also including down-regulated *outlier* profiles) and DIDS were also included in our experimental benchmark because their capability to find outlier genes has been reported (Wang et al. 2012, de Ronde et al. 2013). As described in Materials and Methods (Section 2), two independent tests (GlobalTest and PCA) were set up to assess the relevance of the gene signatures found as significant by each of the compared methods (**Fig. 1-R-6A**). The number of genes found (e.g. OSC dataset: 331 genes found with mCOPA, 1586 with DIDS and 161 with DECO) were always significant ($p\text{-value} \leq 0.05$) according to the respective algorithm. The results obtained for each of the three clinical datasets are presented in **Figure 1-R-6A** as an illustrated table.

As mentioned before, none of the well-established methods were able to find differential genes among the two categories of samples defined in each dataset. Furthermore, the methods that gave differences (mCOPA, MOST, DIDS, DECO) widely differs in the size of gene signatures found. In fact, DIDS always provides by far the largest list. Since it is well-known the difficulty of evaluating several gene signatures if there are no true positives and the size varies among methods, we run the tests using top-100 genes according to *p-value* ranking provided by the only three methods which provide a ranking: MOST, DIDS and DECO.

A **Comparison of 6 methods using 3 experimental clinical datasets to find DE genes between two well-defined classes. Output results evaluated using Globaltest and PCA.**

Osteosarcoma dataset (OSC) Metastasis (n=9) vs Non-Metastasis (n=12)		SAM	LIMMA	mCOPA	MOST	DIDS		DECO	
Signature (genes selected by each method)		**0 g**	**0g**	**331 g**	**top100**	**1586 g**	**top100**	**161 g**	**top100**
GLOBALTEST	p-value	--	--	0.0917	**0.0587**	0.0000194	**0.000152**	0.00436	**0.000617**
	% of correct classification	--	--	80.95%	**76.19%**	90.47%	**90.47%**	80.95%	**80.95%**
	statistic (specificity)	--	--	8.91	**8.32**	15.8	**17.9**	15.4	**17**
PCA	% of variability explained	--	--	69.70%	**52.25%**	42.00%	**47.00%**	72.00%	**70.50%**
	Samples well classified (using 1st component)	--	--	12/21	**16/21**	16/21	**19/21**	18/21	**17/21**
Myelodysplastic Syndrome dataset 1 (MDS-1) MDS-RAEB1 (n=21) vs MDS-RAEB2 (n=20)		**SAM**	**LIMMA**	**mCOPA**	**MOST**	**DIDS**		**DECO**	
Signature (genes selected by each method)		**0 g**	**0g**	**86 g**	**top100**	**1452 g**	**top100**	**441 g**	**top100**
GLOBALTEST	p-value	--	--	0.00555	**0.326**	7.72E-06	**0.0000736**	0.000143	**9.97E-07**
	% of correct classification	--	--	75.60%	**56.09%**	85.36%	**78.07%**	78.07%	**90.24%**
	statistic (specificity)	--	--	8.53	**2.74**	11.5	**15**	15.8	**23.6**
PCA	% of variability explained	--	--	63.00%	**39.42%**	36.50%	**55.50%**	53.00%	**52.00%**
	Samples well classified (using 1st component)	--	--	28/41	**29/41**	33/41	**30/41**	37/41	**37/41**
Myelodysplastic Syndrome dataset 2 (MDS-2) Healthy control (n=11) vs MDS-LowRisk (n=13)		**SAM**	**LIMMA**	**mCOPA**	**MOST**	**DIDS**		**DECO**	
Signature (genes selected by each method)		**0 g**	**0g**	**213 g**	**top100**	**1951 g**	**top100**	**1024 g**	**top100**
GLOBALTEST	p-value	--	--	0.00184	**0.00131**	0.000508	**0.00173**	0.00123	**3.62E-06**
	% of correct classification	--	--	91.66%	**79.16%**	83.33%	**83.33%**	83.33%	**87.50%**
	statistic (specificity)	--	--	10.6	**11.8**	14.5	**17.5**	14.5	**28.4**
PCA	% of variability explained	--	--	44.29%	**55.83%**	47.12%	**64.70%**	58.84%	**66.13%**
	Samples well classified (using 1st component)	--	--	15/24	**18/24**	18/24	**16/24**	17/24	**21/24**

B **DE genes with RANDOM selection of samples**

Osteosarcoma dataset (OSC) (n=9) vs (n=12)	DIDS		DECO	
RANDOM sampling (i.e. no classes) (100 iterations)	significant g in best iter	iters with at least 1 significant g	significant g in best iter	iters with at least 1 significant g
	4492 g	**100/100**	58 g	**3/100**
Myelodysplastic Syndrome dataset 1 (MDS-1) (n=21) vs (n=20)	DIDS		DECO	
RANDOM sampling (i.e. no classes) (100 iterations)	significant g in best iter	iters with at least 1 significant g	significant g in best iter	iters with at least 1 significant g
	2049 g	**100/100**	117 g	**6/100**
Myelodysplastic Syndrome dataset 2 (MDS-2) (n=11) vs (n=13)	DIDS		DECO	
RANDOM sampling (i.e. no classes) (100 iterations)	significant g in best iter	iters with at least 1 significant g	significant g in best iter	iters with at least 1 significant g
	8708 g	**100/100**	8 g	**4/100**

C **GlobalTest outcome with RANDOM selection of genes**

		GLOBALTEST (p-value) (average of the iters)	GLOBALTEST (statistic) (average of the iters)
Osteosarcoma dataset (OSC) (n=9) vs (n=12)	**All genes** (g = 20172)	0.322	5.390
	RANDOM selection of 100 genes (5000 iterations)	**0.412**	**5.369**
Myelodysplastic Syndrome dataset 1 (MDS-1) (n=21) vs (n=20)	**All genes** (g = 20172)	0.019	4.260
	RANDOM selection of 100 genes (5000 iterations)	**0.107**	**4.234**
Myelodysplastic Syndrome dataset 2 (MDS-2) (n=11) vs (n=13)	**All genes** (g = 38048)	0.270	4.820
	RANDOM selection of 100 genes (5000 iterations)	**0.361**	**4.830**

Figure 1-R-6. Results of the comparison of 6 methods (SAM, LIMMA, mCOPA, MOST, DIDS and DECO) applied to find differential expression signal in 3 distinct experimental datasets derived from cancer clinical studies. Yellow boxes indicate the best results. **(A)** Two statistical tests (GlobalTest and PCA) were run to evaluate the signature found by each method. **(B)** This table shows number of positive iterations and maximum number of differentially expressed genes found after applying DIDS and DECO on random datasets (group of samples were mixed). **(C)** Negative control table showing the results of GlobalTest for each dataset when all the genes of the expression data matrix were selected as input or when 100 randomly selected genes were selected.

In this way, we observed that DECO gave best results for the two datasets of myelodysplasia (MDS-1 and MDS-2, **Fig. 1-R-6A**) and a close result to *DIDS* for the osteosarcoma dataset (OSC). GlobalTest is a response-outcome test which allows determining how a given gene set marks the difference between the two sample categories compared (i.e. the gene set provided by each method is used in GlobalTest as *a priori* input group of tested variables) (Goeman et al. 2004). The results of GlobalTest gave best p-values using the top 100 best genes that DECO selected, being better than MOST in all cases and better than DIDS in the case of MDS-1 and MDS-2.

PCA results also indicate that the gene sets provided by DECO are the ones that better assigned the samples to their expected category in the MDS cases. Only in the case of osteosarcoma DIDS seems to be slightly better. To validate that these results, we repeated the differential expression analyses doing a random selection of samples in the two categories and evaluating how many significant genes were found in 100 iterations by the algorithm DECO or by the other method that sometimes performed also well in previous analyses, DIDS. These random tests allowed finding that *DIDS* gave many falser positives than DECO, because it selects many more significant genes that should not be found in a random model (**Fig. 1-R-6B**). The robustness of the GlobalTest was also validated using a random selection of 100 genes in 5000 iterations and showing that the resulting *p-values* were not significant (**Fig. 1-R-6C**).

Finally, we tested the performance of the methods building sample class predictors with a machine learning approach: a *leave-one-out* Support Vector Machine (SVM), using *e1071* R package. This approach was only proved with the best methods according to previous comparisons (DIDS and DECO). This benchmark complements previous results because it allows to evaluate the stability of each gene signature found and also the suitability for each sample analysed. The procedure evaluates the performance of n classifiers (one per sample of each dataset) to determine its correct category (control or case), *leaving-out* such sample and using the rest (n-1) to build each classifier. Thus, each predictor is built leaving one sample out and using the top-25 genes that are selected by each method previously applied to the rest of the samples. In this way n predictors (with n = number of samples in each dataset) were calculated and the probability of assigning each sample to its correct class was determined.

The results from these analyses are showed as boxplots in **Figure 1-R-7**, where we can observe that DECO gave the highest probability and lowest dispersion of true class assignment for all samples in the three experimental clinical datasets studied: median probability values ≈0.86 for OSC dataset; ≈0.84 for MDS-1 and ≈0.95 for MDS-2. These trials were also done using a *random* selection of features to have a random reference in the comparison of the methods. As expected, the random selection gave an average classification of 50% (probability about 0.5) for the two possible classes.

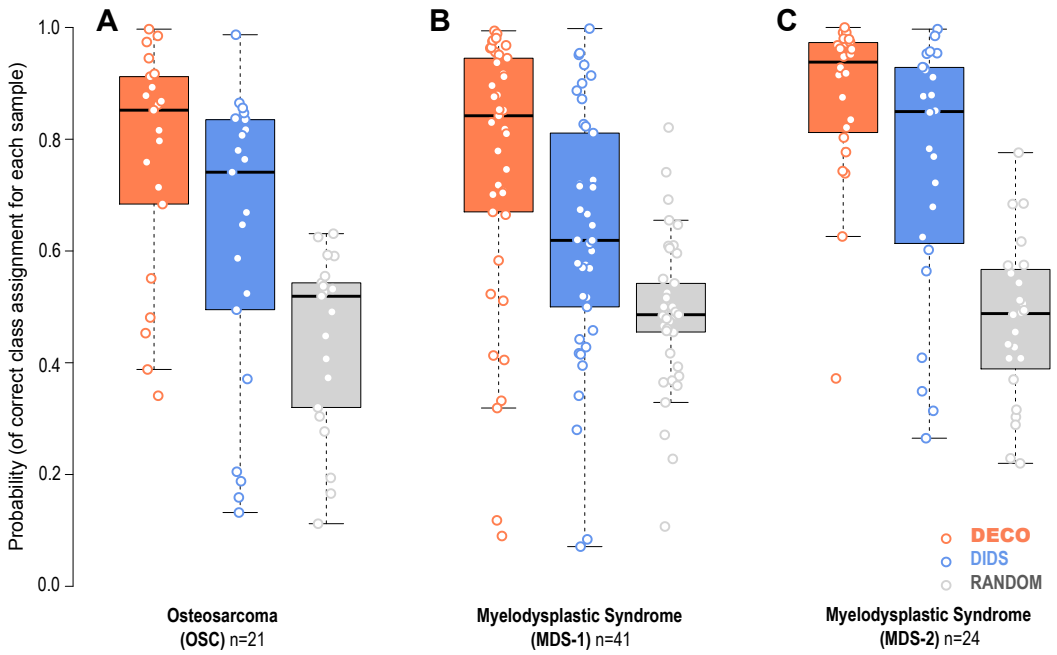


Figure 1-R-7. Support Vector Machine (SVM) predictors built to compare the ability of DIDS and DECO for predicting which class belongs each sample. A *leaving-one* out design was followed: each method provided a gene signature after run on *n-1* samples, then the SVM model was built based on top-25 genes and used to predict the class of the sample outside of design. **(A)** Osteosarcoma dataset, **(B)** myelodysplastic syndrome dataset 1 and **(C)** myelodysplastic syndrome dataset 2 were tested.

The results provided in this section lead us to conclude how robust DECO and its recursive searching of differential features could be on experimental transcriptomic datasets. Interestingly, any classical method (SAM and LIMMA) found differential signal between group of samples due to their lack of attention on outlier profiles. Then, we could see how three independent outcome tests were used to assess gene signatures provided by each method, where DECO outperformed significantly better and meaningful for each sample included in datasets. The recursive analysis included in DECO greatly enhances

the discovery of the 4 model-types hypothesized in this Chapter 1 (**Fig. 1-I-7**), improving the classification of samples even if no global differential signal was found via classical methods.

5. Molecular characterization of hidden factors on a large cancer microarray dataset.

In order to demonstrate that the method not only outperform on simulated data or with small datasets, we tested DECO using two other experimental datasets with more than two hundred samples (**Table 1-M-2**). In this section 5, DECO was applied following an *unsupervised* experimental design which did not presume *a priori* classes or categories.

The first dataset selected was a breast cancer (BCC) collection of 285 samples divided in oestrogen receptor positive or negative (ESR1+/ESR1–) newly diagnosed tumours (**Table 1-M-2**), tested with global gene expression technology (with genome-wide RNA microarrays), taken from the GEO database (ID: GSE25055). This dataset also includes full information about the patients' survival and about their sensitivity to endocrine therapy as well as their sensitivity or resistance to chemotherapy (Hatzis et al. 2011). In this way, the *unsupervised* analysis was carried out using the following as input parameters of DECO: RDA r = 5, combinations = 200000, *adjusted.p.value* < 0.01; NSCA variability explained = 97%, feature threshold = 3 differential in events in at least 5 samples.

After running DECO, 255 genes were selected showing differential expression changes (**Fig. 1-R-9**). The values of all the statistical parameters provided by DECO for these 255 genes are included as **Additional Table 1** (CD of this Thesis), while the complete data matrix corresponding to the h-statistic per gene and sample is provided as **Additional Table 2** (CD of this Thesis). As a whole, the results obtained with this dataset found 6 major subclasses or categories, where primary division of sample dendrogram reveals a deep division between ESR1+ and ESR1- samples. Furthermore, it is important to note that there was a high correspondence between the sample source and a significant subset of genes which marked two subclasses: subclasses 2 and 3 (**Fig. 1-R-9**).

All the clinical samples from primary breast cancer tumours used for this study were obtained by two different groups: collected by the *M. D. Anderson Cancer Center* (MDACC, Houston) or collected by the group called *Investigation of Serial Studies to*

Predict Your Therapeutic Response (I-SPY) (Hatzis et al. 2011). Each of these two groups used a different procedure to isolate the tumour biopsy samples: **(i)** 227 samples were obtained by fine-needle aspiration (MDACC) and **(ii)** 83 samples were obtained by surgical resection of the core biopsy (I-SPY) (Hatzis et al. 2011). We observed in the results provided by the algorithm DECO that a small group of genes marked a clear difference between these two groups of samples isolated in a different way. We tested that this signal was not due to a random selection or to a bad normalization of the data, since more that 95% of the genes did not show any significant difference within these two classes. Therefore, we concluded that those genes were indicating a small change in the expression signal due to some differences in the two isolation protocols used. In fact, according to the h-statistic provided by our method, two of the highest discriminating genes found for these subclasses were haemoglobin β-subunit and δ-subunit (HBB and HBD genes; **Fig. 1-R-8**), which have been recently reported in the literature for suffering a depletion depending on the procedure of biopsy sampling used in patients with breast cancer (Tanamai et al. 2009). All gene expression values in **Figure 1-R-8** have been sorted by h-statistic ranking, showing the power of this new parameter to highlight relevant feature-sample associations.

Together with the signal coming from haemoglobin depletion, the same group of samples showed a strong up-regulation of collagens (COL1A1, COL1A2, COL3A1, COL4A1, COL4A2, COL5A1, COL5A2, COL6A3) which could reveal changes in the extracellular matrix components and could be related to mechanical manipulation of tissue samples, and therefore related to the different isolation procedure (gene patterns 4 and 5, **Fig. 1-R-9**). This effect was not reported in the analysis of the samples published by the original authors (Hatzis et al. 2011), probably because it affects a small number of genes and does not affect to any critical breast cancer associated gene.

As described, we had here a clear discovery of a small gene signature associated with a specific sample subtype (subclasses 2 and 3, **Fig. 1-R-9**) that shows the value of using our new algorithm. It is also important to indicate that doing a semi-supervised analysis or unsupervised analysis with classical methods based on the gene expression signal would not find this signature.

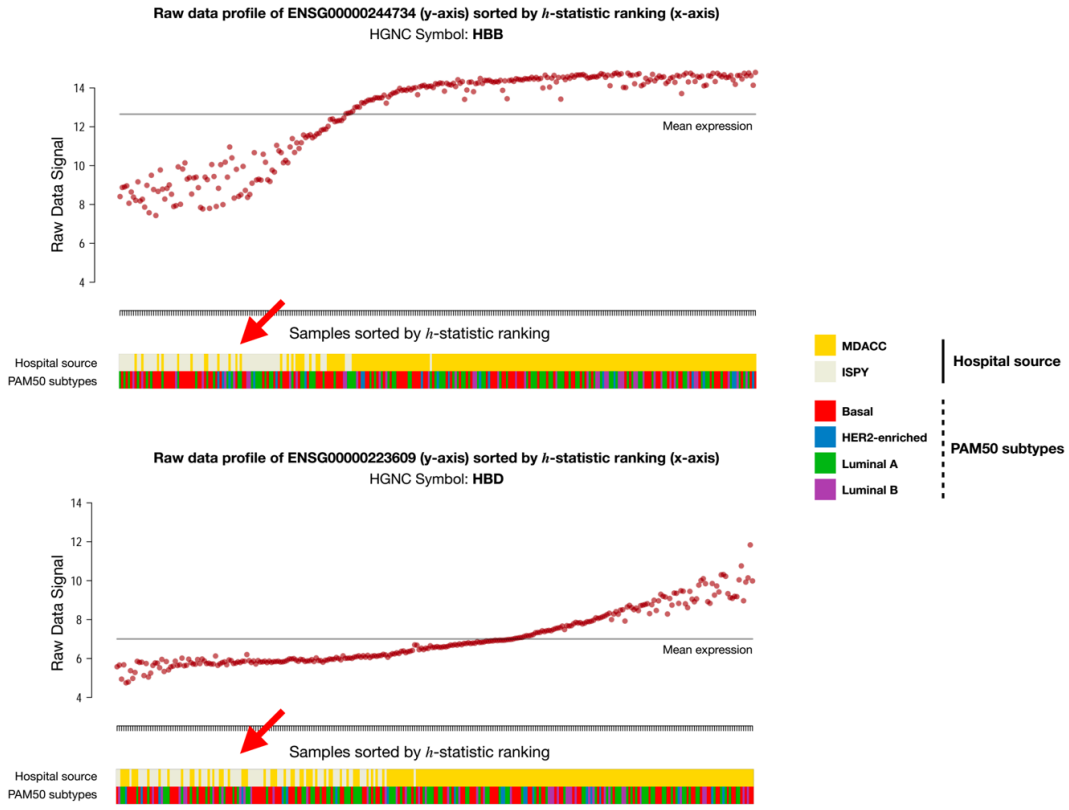


Figure 1-R-8. Gene expression profiles of HBB and HBD genes (haemoglobin β-subunit and δ-subunit) from GSE25055 breast cancer dataset. Both plots have been modified from original representations returned by DECO within its PDF report file (Section 8.3 of Introduction). Red arrows point to samples from ISPY hospital (grey), where the protocol for breast biopsy is different, showing a depletion of the expression of HBB and HBD genes.

According to the standard and well-known subtypes of breast cancer, the results of our analysis showed how the h-statistic provided by DECO found the expected division of samples that follow the PAM50 subclasses (Parker et al. 2009) (**Fig. 1-R-9**). In this way, the method was able to find not only the large differences that marked the separation between basal and luminal-like BCC subtypes, but it also found gene subsets directly related to the other subtypes of BCC that usually are more difficult to separate, like: luminal (A and B) and HER2-enriched (**Fig. 1-R-9**). The method also found specific genes associated with basal or luminal PAM50 subtypes (like: GATA3, TBC1D9, EN1, CA12, NAT1, PROM1 and AGR2) that have been previously linked to the ESR1 status in breast cancer (Parker et al. 2009). In fact, a functional enrichment analysis, using DAVID web tool (Huang da et al. 2009), of the group of genes found by DECO marking the basal BCC subtype showed a high enrichment within specific gene sets that corresponded to basal up-

regulated or down-regulated genes against luminal breast cancer samples (as defined in the Molecular Signatures Database at the Broad Institute, MSigDB, http://software.broadinstitute.org/gsea/msigdb/).

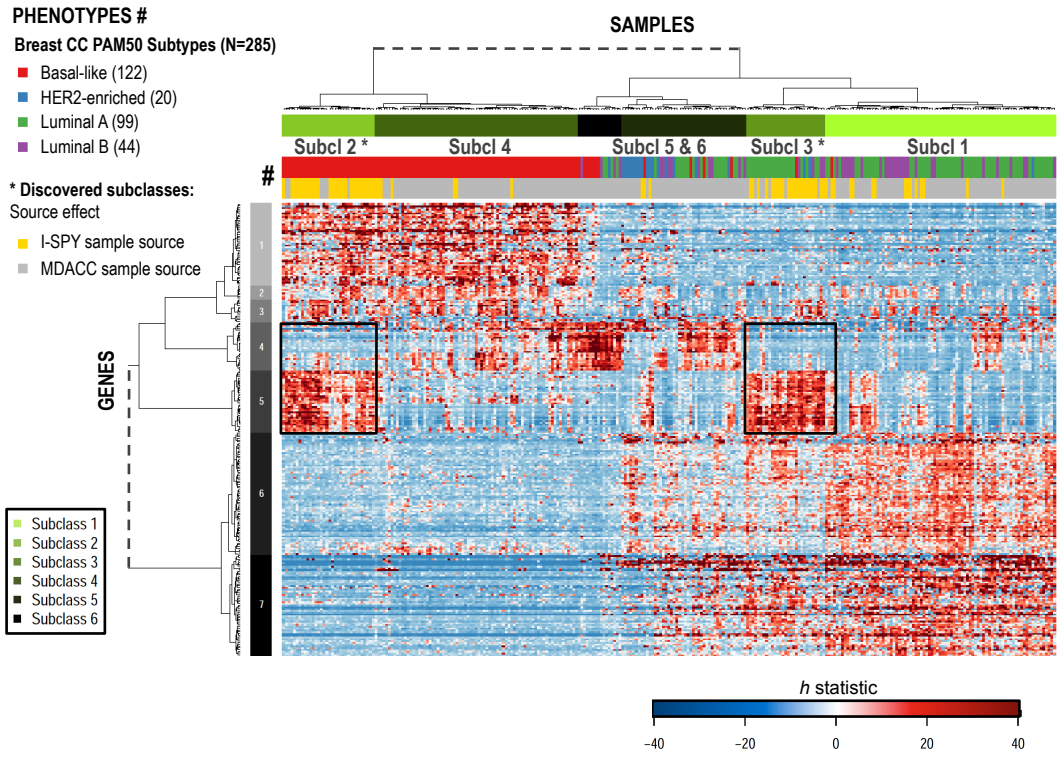


Figure 1-R-9. Heatmap representation of the *h*-statistic of 285 patients and 255 DE genes, calculating the distance matrices using 1-*Pearson* correlation of the *h*-statistic values and posterior Ward's criterion for hierarchical clustering of both features and samples. DECO (unsupervised experimental design, no group of samples) was previously applied on the microarray expression data (obtained from Hatzis et *al.*).

6. DECO matches disease subtypes using an unsupervised design on RNA-sequencing data.

Our algorithm DECO was also applied to another large breast cancer dataset taken from the TCGA database through *TCGA2STAT* R package (Wan et al. 2016), which includes genome-wide expression profiling using high-throughput RNA sequencing (Section 1 of Material and methods). In a recent study Ciriello and collaborators (Ciriello et al. 2015) analysed this dataset characterizing a distinct disease inside the breast cancer tumours corresponding to *invasive lobular* (IL-BCC) subtype. It is clinically and molecularly different to the more common and frequent *invasive ductal* (ID-BCC) subtype. This tumour stratification was not previously investigated because the normal molecular portraits of

human breast tumours, even for the datasets of TCGA (Cancer Genome Atlas Research et al. 2013), followed the most standard classification of breast cancer in 4 subtypes: luminal A, luminal B, HER2-enriched and basal-like (defined by the PAM50 signature excluding normal subclass) (Parker et al. 2009).

Under this scenario, we took 596 breast cancer samples studied by Ciriello et al. (Ciriello et al. 2015) having samples corresponding to each one the 4 main BCC subtypes (307 luminal A, 128 luminal B, 53 HER2-enriched and 108 basal-like), but also including the new tumour subtype classification: IL-BCC and ID-BCC. We analysed this dataset with the algorithm DECO following an *unsupervised* experiment design to test if our method was able to find genes as features that distinguished and separated all the different disease subtypes. For this analysis, we used the original expression RPKM data matrix provided by TCGA, checking a correct normalization and filtering-out 902 genes due to their low expression in all samples (expression signal RPKM < 2). Then, all omic RPKM matrix was properly converted to log2 scale ($log2(RPKM + 1)$). Consequently, DECO was applied without any predefined category of samples, where initial parameters were previously set up to: RDA r = 5; combinations = 1000000; *adjusted.p.value* < 0.01; NSCA variability explained = 80%, feature threshold = 3 differential events in at least 30 samples.

Heatmap representation in **Figure 1-R-10** shows the binary clustering of samples and genes obtained using the h-statistic provided by DECO for each sample and each gene. The method selected 3228 genes that had differential expression changes (according to the threshold indicated above). The values of all the statistical parameters provided by DECO for these 3228 genes are included in **Additional Table 3** (CD of this Thesis), while the complete data matrix corresponding to the h-statistic per gene and sample is provided as **Additional Table 4** (CD of this Thesis). These results showed that the method found 4 subclasses directly related to the 4 BCC PAM50 subtypes: subclass 1 corresponding to basal-like subtype in red (**Fig. 1-R-10**); subclass 2 corresponding mainly to HER2-enriched subtype in blue; subclass 3 to Luminal B subtype in purple; and subclasses 4 and 5 corresponding mainly to Luminal A subtype (marked in green).

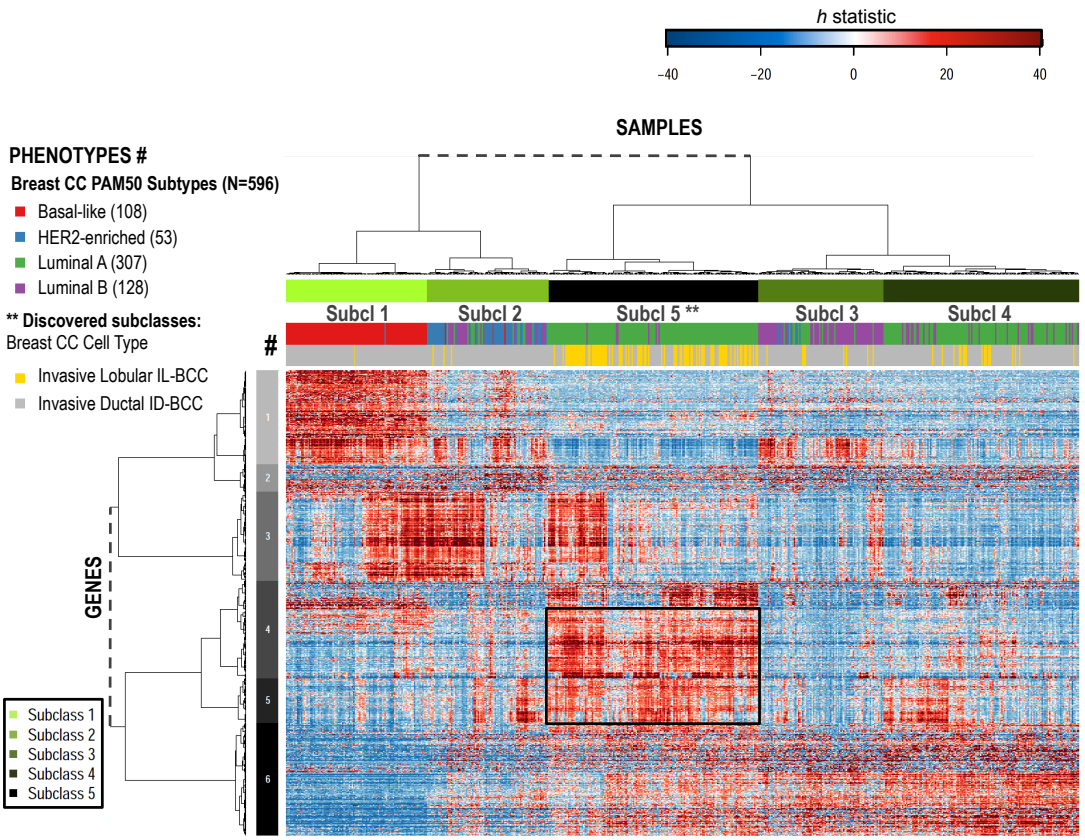


Figure 1-R-10. Heatmap of the *h*-statistic of 596 patients and 3228 DEgenes (the expression data for this cohort was obtained with RNA-seq, from Ciriello et al.). Top dendrogram (corresponding to samples) identifies 6 and 5 main subclasses (marked in A and B respectively). The 4 standard well-known BCC subtypes (usually associated with the PAM50 signature) are labelled with a color panel close to each heatmap, indicating in brackets the number of samples of each subtype.

Interestingly, the method also was able to distinguish inside Luminal A a subtype in that corresponded mainly to subclass 5 and had a distinct gene profile. This subtype corresponded to the recently characterized *invasive lobular* breast cancer (IL-BCC) subtype (marked in yellow in the grey bar in **Figure 1-R-10**). Ciriello and collaborators (Ciriello et al. 2015) published two years ago a comprehensive molecular portrait of the invasive lobular breast cancer (IL-BCC). Some genes found to differentiate *lobular* versus *ductal* breast carcinomas (like thrombospondin 4, THBS4, the thrombospondin receptor, CD36, and multiple cadherins, CDH5, CDH11, CDH17, CDH22, CDH23) (Korkola et al. 2003), were found inside the gene signature that marked the subclass 5 according to DECO. By contrast, some genes that showed significant mutations in IL-BCC, like FOXA1 and TBX3 (Ciriello et al. 2015), are usually up-regulated in all Luminal A samples and so

were not selected as specific markers of the IL-BCC subtype. It is also interesting to remark that previous studies on breast cancer indicated that unsupervised clustering of lobular and ductal breast tumours based on expression profiling failed to distinguish between these two subtypes of carcinomas (Korkola et al. 2003), and this underlines the value of the DECO method to find disease subtypes or classes.

7. DECO multiclass enhances signatures and sample stratification after unsupervised analysis.

It is noteworthy that there is a high and increasing number of publications following a semi-supervised scheme for omic analyses on clinical studies. Under our consideration, a semi-supervised analysis would be composed of a first step to select features associated with the categories of samples of interest (i.e. differential expression) and a posterior analysis to validate this signature through an unsupervised clustering (i.e. hierarchical clustering), which aims to group samples accordingly.

Since the previous section described a particular *unsupervised* scenario of breast cancer where DECO was applied, we actually expected the h-statistic was able to discriminate PAM50 subtypes properly. PAM50 classification was released as a predictive signature of breast cancer prognosis based on the mRNA level of a curated list of 50 genes, dividing the patients into four recognized subtypes: Basal, Her2-enriched, Luminal (A and B) and Normal (Perou et al. 2000). This last subclass was integrated by a particularly heterogeneous group of samples, while the rest of subclasses were well-defined by transcriptomic profiling using microarrays. For this reason, it is reasonable expecting a minimum division among these samples if a transcriptomic omic platform like RNA-sequencing is used.

Interestingly, PAM50 is not the first option for diagnostic of breast cancer patients because the subtypes proposed agree with the absence/presence of three particular biomarkers easily detectable by immunohistochemistry: two hormonal receptors like oestrogen-receptor (ESR1) and progesterone receptor (PGR) and also Her2 gene (ERBB2) (Yip and Rhodes 2014). Indeed, Her2-enriched samples correspond to ESR1-/PGR-/ERBB2+ samples, luminal correspond to ESR1+/PGR+/ERBB2-, while basal samples mostly correspond to the triple-negative case. Notably, it is still an issue that basal-like

samples are not characterized by the presence of any marker, leading to the appearance of multiple analysis referring this question in the last decade (Bianchini et al. 2016, Jiang et al. 2016, Liu et al. 2016, Martínez-Canales et al. 2017).

Since this agreement among PAM50 and immunohistochemistry is not always fulfilled because of intrinsic tumour variability and technical variability (**Table 1-R-1**), a double purpose is followed in this section: explore the transcriptomic signatures of ***pure* breast cancer subtypes** through a combination of PAM50 prognostic value and immunohistochemistry; and characterize in a simple procedure the positive markers for triple-negative (basal-like) samples. Aiming that, we run DECO (*multiclass* design) on a subset of the BCC-2 dataset composed of *pure* samples (**Table 1-R-1**), those meeting the double condition: PAM50 and immunohistochemistry. This subset involved 361 samples divided in three categories (called *Basal_000c*, *Lum_110c* and *Her2_001c*), which have been remained separated in the subsampling procedure or RDA (r = 5; iterations = 10000; adjusted p-value = 0.01). Later, the NSCA step and h-statistic have been calculated after filtering for significant results to finally obtain 693 features/genes and 3 subclasses defined by DECO (*rep.thr* = 10 and *perc.samp* = 5%; threshold explained at Section 9.3 of Results).

Table 1-R-1. *Pure* samples contained in BCC-2 dataset through combination of PAM50 and immunohistochemistry classifications (0-1 code for three biomarkers ESR1/PGR/ERBB2; *x* corresponds to Not assigned).

PAM50 subtypes \ Immunohistochemistry	000c	001c	00xc	010c	011c	01xc	0x0c	100c	101c	10xc	110c	111c	11xc	1x0c	1xxc	x01c	x1xc
Basal	**68**	7	37	4	1	1	2	6	1	1	1	0	2	0	0	0	0
Her2	9	**22**	4	0	1	0	0	3	8	2	0	8	2	0	0	1	1
LumA	2	0	4	0	0	2	0	21	5	9	**201**	35	114	2	1	0	0
LumB	1	0	1	0	0	1	0	13	5	10	**70**	25	40	0	0	0	0
Normal	3	0	4	1	0	0	0	4	0	0	6	2	4	0	0	0	0

Figure 1-R-11 shows three different dendrograms corresponding to different steps of the simple pipeline carried out by DECO. As detailed above, we faced a very clear biological scenario here because of combining PAM50 and immunohistochemistry markers (*pure* subtypes). For this reason, the whole subset of 361 samples from BCC-2 dataset already shows a good stratification of samples based on these *pure* categories (**Fig. 1-R-**

11C). However, we can also observe several misclassified samples (not placed close to the rest of samples from the same category) marked with red asterisks if raw omic data and all features are input to hierarchical clustering. This first result (**Fig. 1-R-11C**) puts forward the idea of combining both predictive tests for enhancing classification of samples and reassures the evidence that each subtype actually comprises a very different biological scenario, as reported in the literature.

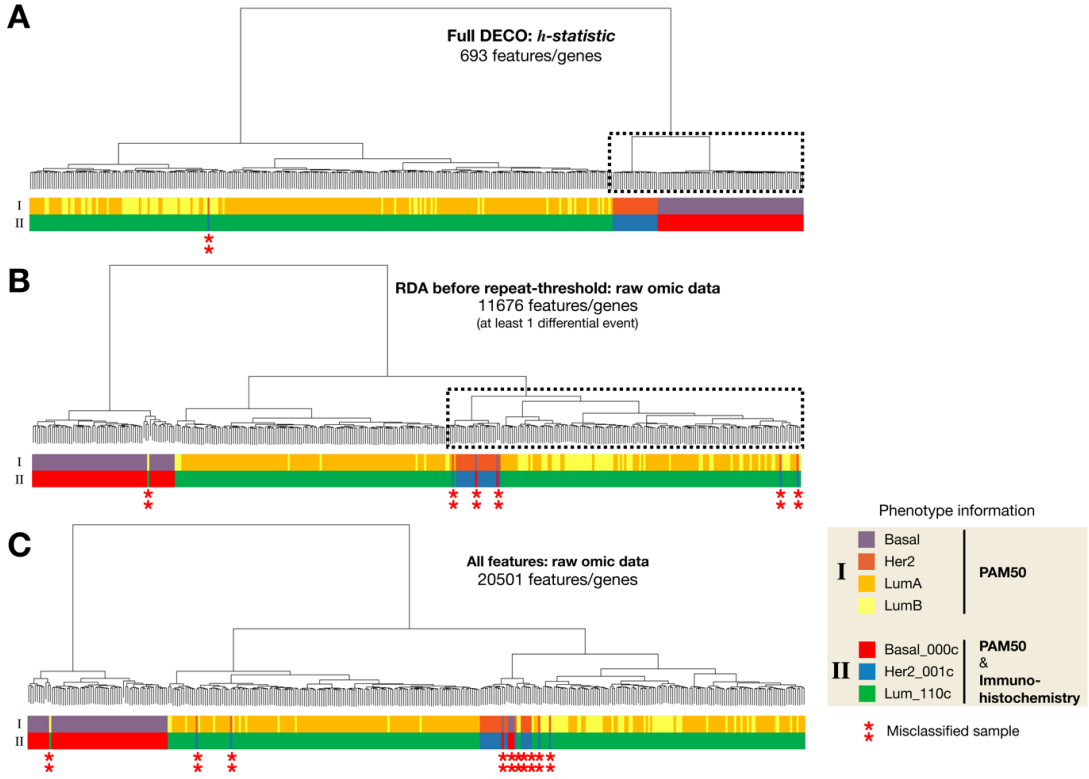


Figure 1-R-11. Dendrograms obtained after hierarchical clustering of: **(A)** *h-statistic* from DECO, **(B)** raw omic data of gene signature obtained via RDA (any feature with at least one differential event) and **(C)** raw omic data of whole BCC-2 dataset. Red asterisks mark misclassified samples and dashed boxes mark relevant parts of the dendrogram. For all panels, distance was based on *Pearson* correlation (distance = 1 - correlation), while hierarchical clustering used *Ward* method.

Consequently, we assumed that the application of DECO would retrieve even a better sample stratification and specific signatures' definition. Thus, in **Figure 1-R-11B** we can observe the corresponding dendrogram to raw omic data of selected features by RDA step (at least 1 differential event per feature). Still, there are several misclassified samples but less than previous (panel C). Interestingly, the hierarchical clustering puts together in the same branch (dashed box, **Fig. 1-R-11B**) samples from *Her2_001c* and a subset of

Lum_110c (both Luminal A and Luminal B). Alternatively, both h-*statistic* used by DECO and double repeat-threshold on features selected by RDA improve the sample stratification, reducing to only 1 misclassified sample and clustering together all samples of *Lum_110c* subtype but separated from *Her2_001c* and *Basal_000c* (**Fig. 1-R-11A**). This is a positive result because two first biomarkers (ESR1 and PGR) correspond to hormonal signalization which it is known to be more relevant, so *Her2_001c* and *Basal_000c* may be more related due to the absence of both biomarkers.

As mentioned before, our main purpose will be the definition of triple negative or basal-like signature. Although it is a very distant subtype from other breast cancer subtypes in terms of gene expression, there is no clear gene signature described in the literature for these samples. Classically, patients including in the triple-negative subtype have poor prognosis and are diagnosed by the absence of the 3 markers mentioned above (ESR1/PRG/ERBB2). No positive biomarker was definitely proposed for the diagnosis of this group but several studies have been published approaching this, as mentioned above.

Since DECO has been proven to improve the original sample stratification, we could review the **up-regulated biomarkers found for *Basal_000c* subclass**. For this purpose, we firstly focused on the average h-statistic value of each feature per subclass provided by DECO. The top-15 up-regulated features or genes assigned to this subclass included: ROPN1, ART3, HORMAD1, GABRP, ZIC1, A2ML1, KRT16, MSLN, PRAME, ROPN1B, FABP7, MIA, EN1 and SOX10 (decreasingly ordered). Several of these genes were reported as breast basal-like carcinomas markers on an independent cohort of patients (Ivanov et al. 2013), specifically EN1 has been proposed as new targetable gene (actinomycin D) in triple-negative breast cancer samples (Kim et al. 2018). Remarkably, two of them are paralogs (ROPN1 and ROPN1B) related to PKA-dependent signalling processes for spermatozoa capacitation, whose high expression have been also associated with melanoma (Uhlen et al. 2017). Moreover, other genes like MIA was broadly associated with different types of cancer (Sasahira et al. 2016).

Interestingly, DECO also corroborated the very clear gene expression **signature for *Her2_001c*** subgroup composed of ERBB2 amplicon at 17q12-q21 DNA region (i.e. ERBB2, STARD3, GRB7, PGAP3 or CDK12) (Kauraniemi and Kallioniemi 2006).

All these results obtained by simple analysis, carried out by DECO on a subset of BCC-2 dataset (**Table 1-M-2**), leads us to conclude that DECO resembles an excellent and

robust bioinformatic tool to analyse and characterize gene signatures of subclasses of samples in a single pipeline (similar to *semi-supervised* design mentioned above), enhancing the sample stratification even if a very clear stratification was present in raw omic data (**Fig. 1-R-11C**).

8. R package: *deco*.

8.1 General environment

DECO was initially programmed in R environment, then a complete R package, called '*deco*', have been produced. Additionally, a detailed R vignette (**Appendix 1**) describing how this R package functions could be consulted, which was also included in the R package.

Table 1-R-2. R package dependencies of *deco* R package.

Package	Description	Version
limma	Linear Models for Microarray Data used for differential expression analysis.	*3.30.13*
snowfall	Library for easy parallel computation in R.	*1.84-6.1*
foreign	Optimized reading and writing files in R.	*0.8-67*
AnnotationDBI	Management of Annotation libraries and objects.	*1.36.2*
Biobase	Management of R objects from Bioconductor.	*2.34.0*
gdata	Advanced management of R matrix and data.frame.	*2.18.0*
lisp	Higher-order programming.	*0.1*
ade4	Tools for multivariate data analysis.	*1.7-6*
locfit	Local regression, likelihood and density estimation.	*1.5-9.1*
sfsmisc	Approximation to numerical integration.	*1.1-0*
gplots	Libraries for representation of results and plot configuration.	*3.0.1*
RColorBrewer		*1.1-2*
scatterplot3d		*0.3-40*
made4		*1.48.0*

Our method was written, developed and compiled within R environment (R-version 3.4.0). The *deco* R package created is compatible with all R-versions since 3.0.1, it is periodically maintained and revised to ensure its correct function. Today, *deco* R package

is available for downloading in Bioinformatics and Functional Genomics group's website (http://bioinfow.dep.usal.es/deco/) and it is also planned to upload it into CRAN R repository (http://cran.r-project.org/).

8.2 R dependencies of *deco* R package

The development of *deco* R package leads to include several R packages for the proper functioning of this method. While *limma* R package is essential for RDA step and differential analysis or *snowfall* for parallel computation, other R packages like *gplots*, *RColorBrewer*, *scatterplot3d* or *made4* were used to improve representation and plot of *deco*'s results. Moreover, other packages were included for management of data (*foreign* or *gdata*), bioinformatic or annotation protocols (*Biobase* or *AnnotationDBI*), advanced statistical analysis (*locfit*, *ade4* or *sfsmic*) or optimized programming (*lisp*) (**Table 1-R-2**). Moreover, the R code for NSCA was adapted from original code produced and published by Beh and Lombardo (Beh and Lombardo 2014).

9. *deco* R package: development and main functions created

Since DECO method is composed of two main steps: RDA and NSCA (Section 2 of Material and methods), the *deco* R package have been written following the same scheme. Two main R functions, called *decoRDA* and *decoNSCA*, would compute both main procedures: the results of the resampling procedure (*decoRDA*) are needed by *decoNSCA* to calculate the Non-Symmetrical Correspondence Analysis and, then, the h-statistic.

9.1 Input data

Given an omic data matrix as input to analyse, the features/genes/proteins must be placed as rows and samples as columns. This omic data matrix is the unique requirement to run *decoRDA* R function. Nevertheless, a named vector indicating which samples are belonging to each group of samples may be also input, if supervised or multiclass experimental design is planned. As hinted above, the omic data matrix must be properly normalized by any suitable method for the omic platform used to generate the data. For example, several normalization methods have been proposed or used for DECO validation, like RMA normalization method for Affymetrix microarray platform, RPKMs for RNA-seq

experiments or CPMs obtained after counts/reads normalization through **voom** normalization method (Law et al. 2014).

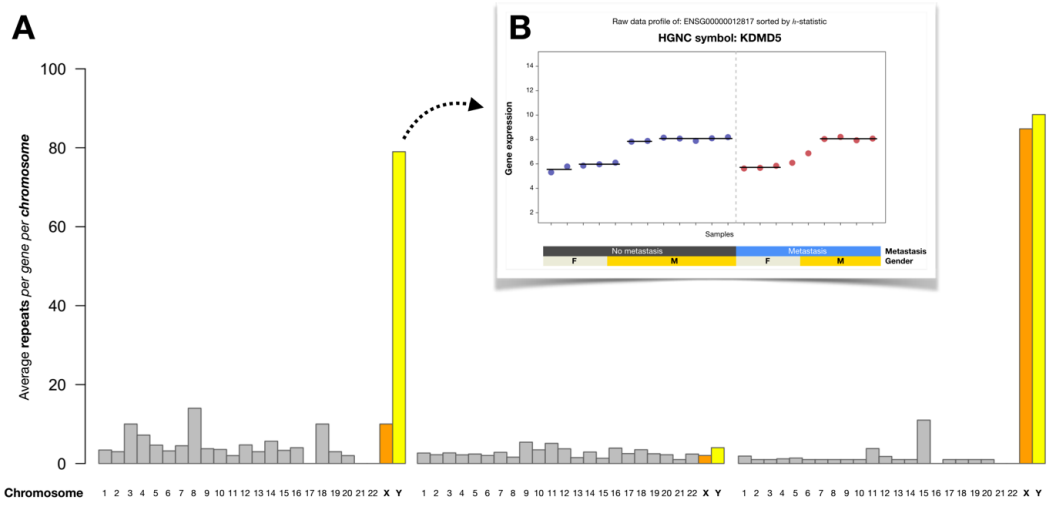
9.2 *decoRDA* R function

The ***decoRDA*** R function would compute the subsampling procedure depending on the experimental design: supervised, multiclass or unsupervised. Anywise, the groups of samples will be compared using LIMMA (Smyth 2004) through pair-wise comparisons for supervised and multiclass experimental design. Given a number of iterations/combinations of samples (all possible combinations or a random subset) and an optimal subsampling size (Section 4 of Material and methods), *decoRDA* will calculate the combinations and compute a LIMMA differential expression analysis iteratively on these subset of samples, saving the statistical output given by LIMMA per iteration and counting in an incidence or frequency matrix which features have been identified as significant for each subset of samples (Section 5 of Material and methods). In case of supervised analysis was chosen, the frequency matrix will be divided by classes to enhance posterior subclass discovery, then both UP and DOWN differential events will be separately counted in the frequency matrix, doubling the number of rows (one UP and one DOWN per feature). Otherwise, differential events will be annotated only in samples where the change was an UP regulation (Section 4 of Material and methods).

These results and the incidence matrix are not directly saved as R objects when the subsampling procedure is executing due to its computational cost and RAM memory consumption. For this reason, the *foreign* R package was implemented to write and read files of a temporary folder, previously created to save these intermediate results. This temporary folder will be immediately removed after *decoRDA* finish.

Interestingly, since the use of omic data is continually growing, a **parallel computation** of *decoRDA* was also implemented to hasten the subsampling procedure. For this purpose, the *snowfall* R package was chosen due to its stability, simplicity and the development of parallel computation alternatives for the gold-standard loop functions in R: *apply*, *sapply*, *lapply*, *mapply* and *tapply*. Moreover, this library and parallel computation were also implemented for the calculation of the **overlap** statistic within the *decoNSCA* R function.

Apart from technical issues derived from resampling procedures in R, others biological considerations were done in order to facilitate the output of *decoRDA* R function. First, we noticed that working with omic data derived from clinical data made the **patient/sample gender** an issue when resampling techniques are proposed (**Fig. 1-R-12**). All genes or features located in sexual chromosomes would be great candidates to be differentially expressed features even when all groups of samples are well-balanced (*mixed* profiles, **Fig. 1-I-7**), leading for the discovery of clear subclasses and confusing results. For this reason, a R function called *AnnotateDECO* has been developed to annotate each feature according to the attributes indicated by user and to the official R/Bioconductor annotation library for the organism of study. This function is called by *decoRDA* if user requires it. By default, a logical variable called *rm.xy* would indicate to *decoRDA* if the annotation information (chromosome of each feature) should be used to remove those features located at X or Y chromosomes.



OSC

	Female	Male
Metastasis	4	8
No metastasis	4	5

MDS-1

Unavailable data

MDS-2

	Female	Male
Healthy	7	4
Low Risk MDS	5	8

Figure 1-R-12. Figure representing differential evidences of sexual genes if they are included in a resampling procedure, like RDA step. Here, DECO was run by default parameters (following similar supervised design than Fig. 1-R-5) on real datasets OSC, MDS-1 and MDS-2 while setting *rm.xy* to FALSE. **(A)** Height of bar corresponds to average repeats (or differential events) amounted by each set of genes found per chromosome. **(B)** A single gene profile of KDMD5 (chromosome Y) obtained after applying DECO is shown. As we can see, it follows a *mixed* profile.

9.3 *decoNSCA* R function

Once *decoRDA* R function has been run, the R object list provided will be the input

for next function ***decoNSCA***. First, this function will remove those noisy features which results from the subsampling subset may be considered irrelevant (Section 4.4 of Material and methods). While the common thresholds are based on a single parameter cut-off which summarizes significance (i.e. *adjusted p-value*), we demonstrated how it may hide relevant results in heterogeneous scenarios (**Figure 1-I-7**). Although *Standard Chi-Square* was proposed for summarizing the number of positive iterations amounted by a single feature, they may be mostly given by a specific group of samples or globally by low number of repeats in a larger number of samples. Thus, this **repeat-threshold** proposed aims to remove:

a) Features amounting a very **low number of differential events on a large number of samples**.

b) Features amounting a **great number of differential events on a very small number of samples**.

Given one differential feature selected by *decoRDA*, this removal step would be based on the number of differential events or repeats amounted by a minimum percentage of samples. By default, those parameters are settled as 3 repeats (*rep.thr* input in R) and 5% of samples (*samp.perc* input in R), then any feature amounting less than 3 differential events for at least 5% of samples will be removed. In this way, it may be considered as a two-dimensional threshold, as it is shown in the **Figure 1-R-12** corresponding to the analysis of BCC-2 dataset (***pure*** classes, *multiclass* analysis from Chapter 1-R Section 7). Attending to relevant features or genes for BCC-2 dataset, we can observe how very significant features for PAM50 or immunohistochemistry classification like ESR1, PGR or ERBB2 are high ranked and would not be removed (**Fig. 1-R-12**). However, this heuristic threshold prioritizes *outlier* genes like CDK18 (for subgroups or small set of samples), while withdraw other less specific like GBJ6, KRT17 or CDH1 (described above: a) situation). In summary, this double threshold would allow us to control which features will enter to the NSCA analysis, avoiding similar situations to those examples given by current methods for *outlier profile* detection, like DOG or ZODET methods (Yang and Yang 2013, Roden et al. 2014), where even a single outlier sample may enable significant results (**Fig. 1-I-4**).

After this removal step, *decoNSCA* will take the frequency matrix (two if supervised analysis) and will search for subclasses of samples calling the internal function *NSCAcluster*, which specifically applies NSCA and later calculates the ***h*-statistic** based on inner product of NSCA and dispersion of raw omic data per feature (Section 5 of Material and Methods). After the h-statistic calculation, another R function called *cophDECO* will define subclasses based on a hierarchical clustering for samples and for features of this statistic. Separately, the **overlap** statistic will be also computed for each feature (if supervised or multiclass experimental design) using raw omic data.

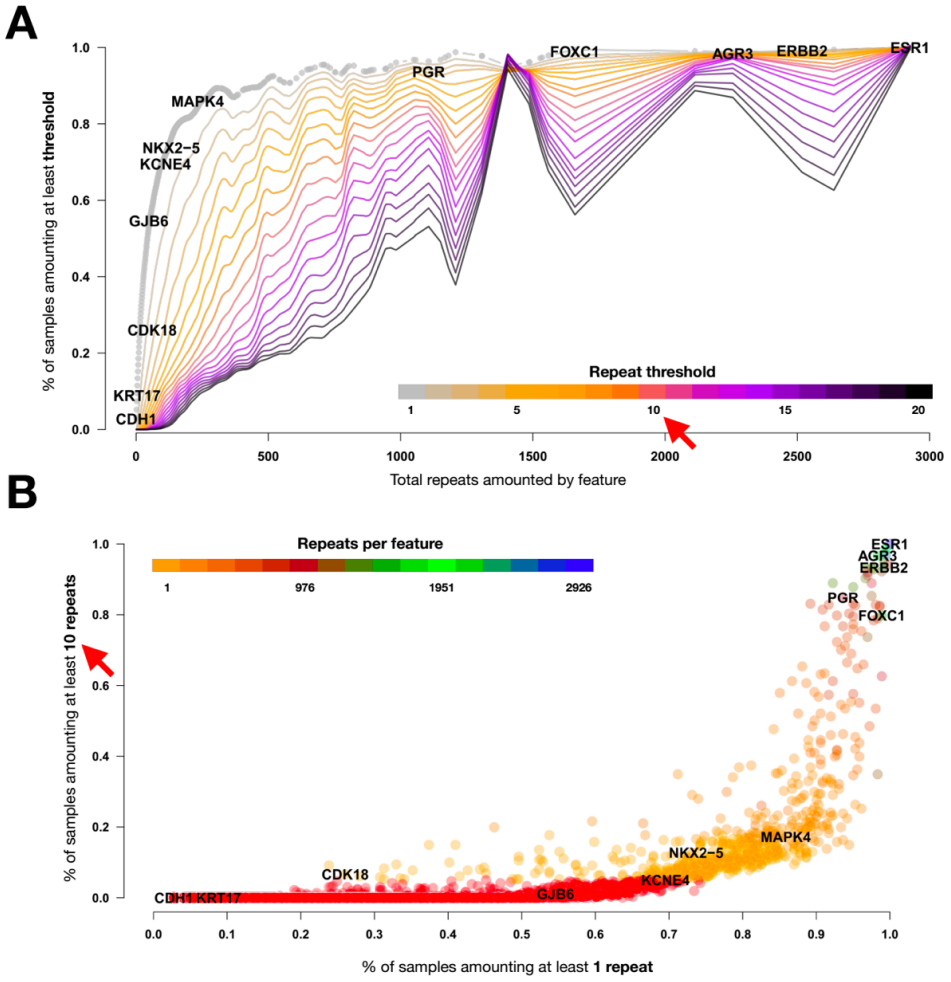


Figure 1-R-13. Example of double repeat-threshold (*rep.thr* and *samp.perc*) proposed for BCC-2 dataset. **(A)** While x-axis corresponds to *repeats* (differential events) amounted per feature remains invariable, y-axis represents percentage of samples amounting at least the repeat-threshold chosen (colour code from 1 to 20 *repeats* here). Trends have been fitted using cubic smooth spline approach (s-par = 0.75). **(B)** Once a repeat-threshold was chosen (10 repeats, red arrows), we represented the *samp.perc* threshold which would filter all these features amounting less than repeat-threshold in a percentage of samples (5% by default). Red dots assigned to filtered features. Both plots included position of relevant genes for BCC-2 dataset.

In order to facilitate management of results provided by *decoNSCA*, an own R object-class was created following the S4 object system and called *deco* class. Notably, this object includes statistical information and a table ranking about each feature, all the information about subclasses found and h-statistic matrix, and also a summary of input parameters given by user.

9.4 *decoReport* and *plotDECOProfile* R functions

Any high-throughput analysis in research should be properly illustrated or represented to facilitate the understanding of other interested researchers. Moreover, we contemplated that some parts of this algorithm may be not easily understood, then it may seem even more complex than really are. Aiming to highlight the most relevant parts, *deco* R package also implements two R functions called *decoReport* and *plotDECOProfile* which will generate a single PDF file including several plots about data and results.

The ***decoReport*** R function produces a PDF file summarizing most of the analysis done by DECO (RDA and NSCA steps). Since a complete vignette is attached as **Appendix 1**, we will just focus on main plots:

(i) First page of the PDF file will briefly **summarize the results** obtained after both RDA and NSCA steps. Number of features selected after RDA, Hubert's gamma coefficient pointing the suitability of subclasses found after hierarchical clustering of h-statistic, top10 ranking of features or subclass membership of each sample are some of this information. This brief summary may be also shown via *summary()* or *print()* R native functions.

(ii) Boxplot of **Goodman and Kruskal's τ** (or inertia) contributions (Section 4 of Introduction) per sample belonging to each subclass found is represented. Here, a similar τ contribution and low dispersion of τ values may be expected for very different samples conforming a subgroup (particular behaviour within the dataset). For example, Basal-like or triple-negative (negative for ESR1, PGR and ERBB2 gene markers) subgroup of samples from BCC-2 dataset integrates a very characteristic biological scenario, aggressive and showing high diversity of outcomes (Jiang et al. 2016, Martínez-Canales et al. 2017), which was reflected via τ contributions.

(iii) A similar plot to **Figure 1-R-13B** remaining how the previously explained **repeat-**

threshold was computed is also provided.

(iv) The **top-50 feature ranking** table is provided in a separate page. If *supervised* experimental design was chosen, then two top-50 tables are provided in order to segregate *complete* and *majority* profiles from *minority* and *mixed* ones. This table is also included in the *deco* R object returned by *decoNSCA* function.

(v) Interestingly, a **top-50 ranking based on *h*-statistic** range is provided (two tables separated per category of samples if *supervised*), revealing how *h*-statistics per feature are pointing to each subclass discovered (**Fig. 1-R-14**). This table is also included in the *deco* R object returned by *decoNSCA* function.

(vi) **Heatmap of *h*-statistic** showing the hierarchical bi-clustering (feature and samples) computed by *decoNSCA* to disclose subgroups of samples and feature's patterns associated with each subgroup. It may include additional information of samples (phenotype) or features if provided as *data.frame* via *info.sample* or *info.feature* inputs to the *decoReport* R function.

Mean *h*-statistic per subclass within ALL samples

Top-50 discriminant features among subclasses found by DECO

		C	B		A				A		D	
ID	SYMBOL	Standard.Chi.Square	Ranking.Scl1.All	h.Scl1.All	Ranking.Scl2.All	h.Scl2.All	Ranking.Scl3.All	h.Scl3.All	h.Range.All	Dendrogram.group.All	Dendrogram.order.All
ERBB2	ERBB2	250.16883	686	−0.30479165	637	−5.213693	1	121.650301691	126.86399	10	690
STARD3	STARD3	254.16898	678	−0.73528688	690	−2.006689	2	100.112357277	102.11905	10	689
ESR1	ESR1	260.10855	3	13.30904594	40	−25.458527	5	−54.404587284	67.71363	8	494
TFF1	TFF1	175.12590	1	15.46013582	1	−47.080384	99	−10.760238985	62.54052	5	326
PGAP3	PGAP3	155.10551	685	−0.31076787	628	−5.359427	4	55.914196666	61.27362	10	691
GRB7	GRB7	147.21103	628	−1.36316059	692	−1.602089	3	57.839528790	59.44162	10	692
AGR3	AGR3	245.27728	2	15.02062449	2	−44.297639	59	−15.458144977	59.31826	9	609
ABCC11	ABCC11	142.91070	83	5.89601692	11	−31.803937	24	25.020932362	56.82487	5	344
ANKRD30A	ANKRD30A	202.13014	4	12.15968652	5	−39.382203	575	−1.066652595	51.54189	9	597
CYP4Z2P	CYP4Z2P	138.63912	54	6.55737078	16	−30.162231	30	21.257240755	51.41947	6	429
TFF3	TFF3	145.42568	9	10.14587851	3	−39.771763	583	1.001777229	49.91764	5	325
C1orf64	C1orf64	202.94726	5	12.09233837	6	−37.210388	395	−2.911101703	49.30273	9	621
AGR2	AGR2	125.08425	11	9.52130702	4	−39.710181	317	3.778780841	49.23149	5	323
TMEM45B	TMEM45B	104.76490	342	3.05612247	118	−19.379598	17	27.670834450	47.05043	5	342
ROPN1	ROPN1	158.80012	59	−6.40872230	13	31.023639	62	−15.160584316	46.18422	4	274
GP2	GP2	138.15179	7	10.91198658	7	−35.018944	101	−10.584329887	45.93093	9	601
MIA	MIA	97.45236	145	−4.77651503	50	24.424357	31	−21.097070941	45.52143	4	270
SCGB2A2	SCGB2A2	45.96344	52	6.58274494	14	−30.728924	70	14.425615008	45.15454	5	363
PNMT	PNMT	44.83636	691	−0.08629879	441	−9.372901	8	35.561705618	44.93461	10	688
PIP	PIP	71.55331	29	7.52730089	10	−33.507816	105	10.290511050	43.79833	6	386
SOX10	SOX10	70.72679	127	−4.96228943	56	24.204366	36	−19.502728782	43.70709	4	269

Figure 1-R-14. Adapted screenshot of *h*-statistic ranking table within the PDF file returned by *decoReport* R function after BCC-2 analysis (*multiclass* analysis). **(A)** Discriminant features will be ranked according to the maximum range of values reached by *h*-statistic per feature. **(B)** *Ranking* columns would rank features according to this relevance or mean *h*-statistic per subclass. **(C)** *Standard.Chi.Square* column just displays this statistic derived from RDA step per feature. **(D)** Dendrogram columns point to which group of feature's dendrogram belongs each feature and which exact position occupies.

(vii) **Heatmap of raw or original omic data** (gene expression, miRNA expression, protein expression, etc) based on the same feature selection done by *decoRDA* and filtered by the repeat-threshold in *decoNSCA*.

(viii) **Single feature profiles** plotted in single pages (if *supervised* or *multiclass* design). These pages include a profile plot showing raw omic data signal but ranked accordingly to h-statistic, which enables the visualization of differences. It is accompanied by information of samples if provided. Moreover, the *overlap* of raw omic signal densities per category of samples compared is also represented (**Fig. 1-R-15**), but not for *unsupervised* analysis. Additionally, a barplot indicating mean and standard error of h-statistics per subgroup of samples found by DECO, in order to quickly visualize which subgroup is supported (marked) by this feature. The profiles to be plotted may be indicated via *id* input, limited to 50 profiles.

(ix) **Biplot** and **3D representation** of NSCA coordinates returned for each sample and subgroup of samples discovered.

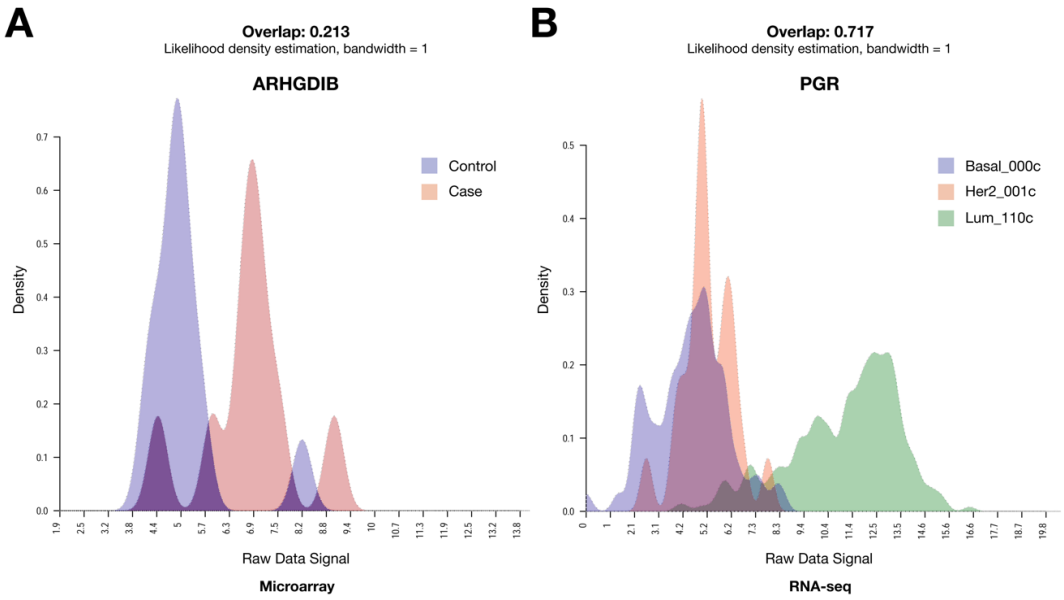


Figure 1-R-15. Adapted screenshot of ***overlap*** statistic plot returned by *decoReport* R function, which will estimate the percentage of overlapping raw omic data signal among classes studied. **(A)** Overlap between non-metastasis (control) and metastasis (case) samples of OSC dataset for ARHGDIB gene. **(B)** Overlap among *pure* classes (Basal_000c, Her2_001c and Lum_110c) of BCC-2 dataset for PGR (progesterone receptor) gene, one of the most relevant in breast cancer.

Additionally, the ***plotDECOProfile*** R function was written for plotting single feature profiles as indicated in (viii) of previous list without restriction of number. However, the *overlap* representation is omitted in this PDF file. Further details about the *deco* R package functions, parameters to control the analysis and R objects returned after each function is provided in the **Appendix 1**, which includes the original vignette specifically written for users of this package.

DISCUSSION

This Discussion presents an overview of the main results presented in this Chapter trying to gain insight into the value of the work done and also reviewing related scientific publications. Our analysis is specifically focused on the strengths and weaknesses of the developed method (DECO) and their main statistics (*Standard Chi-square* and h*-statistic*) for the analysis of complex omic data from heterogeneous sample sets.

1. Recursive subsampling (RDA) provides a robust feature selection both in homogeneous and heterogeneous sample series

Along the Introduction of this Chapter I, we revisited and detailed the main bioinformatic methods and approaches emerged since COPA method proposed the *outlier* profile as a frequent scenario of cancer omic analyses (MacDonald and Ghosh 2006). They proposed the discovery of up-regulation events for a subset of samples when gene expression levels (mRNA) of cancer samples are compared to control samples. In the original publication, they attributed these differential events to genomic translocations of DNA, a very common incident in tumour cells. Particularly, this study was focused on prostate cancer and the fusion of TMPRSS2 and ETS transcription factor genes (Tomlins et al. 2005).

In the singular biological context of cancer, the genomic translocation is one of the many existing sources of biological heterogeneity of tumour cells (Hogenbirk et al. 2016). As mentioned before, individual genotype and phenotypical circumstances, spatial and temporal clonal evolution of tumour cells (even more pivotal if solid tumour) and technical variability (from any high or low-throughput technique) also contribute to a complex scenario where the identification of any relevant source of heterogeneity makes crucial the development of comprehensive approaches (Allott et al. 2016, Rubben and Araujo 2017). However, we mentioned before that most of the current omic analyses focus on *supervised* comparisons (reference samples against case samples) which do not take into

consideration these issues. For this reason, we hypothesized (Introduction, Section 3) a **four model-type scheme** of possible heterogeneous profiles when two categories of samples are compared (**Fig. 1-I-7**), which involves classical or expected complete changes and *outlier* changes between classes.

Given an omic dataset, DECO aims to detect any relevant feature supporting the intrinsic heterogeneity through a subsampling procedure without replacement (RDA). The feature selection process is crucial for any posterior analysis (Singh and Sivabalakrishnan 2015) because it allows us not only select and rank significant features but also place in context which samples are aiding this variability (**Fig. 1-M-4**). Thus, the *Standard Chi-square* was implemented to facilitate the ranking of features found instead a simple counter of differential events (or *Repeats*). It possesses the advantage of being corrected by significance level of each differential event (**Eq. 1.6**): for the same number of differential events per feature, lower p-values (i.e. ≈ 0.001) should be greater considered or scored than greater ones (i.e. ≈ 0.05). Taking into consideration these points, we have demonstrated in Section 1 of Results that RDA step is able to disclose every differential feature and rank them greater than random differences (**Fig. 1-R-2**), following a logical order similar to our 4 model-type (*adjusted p-values* from **Fig. 1-I-7**).

Noteworthy, subsampling and other resampling techniques have been broadly used in many scientific fields for statistic estimation, stability assessment or learning processes. If they are carefully raised (involving previous knowledge, computational cost or suitability of the problem to solve), these techniques will provide very useful and reliable information (Irizarry et al. 2003, Gur-Dedeoglu et al. 2008, Lee et al. 2014). Although *big data* analyses are coming more frequent now, summarizing it into *smart data* remains essential and requires of the development of new exhaustive approaches. Our method DECO, and particularly the RDA step, adds a new scheme analysis on a very acknowledged differential analysis approach like LIMMA and its Bayesian (*eBayes*) method (Smyth 2004), enlarging the suitable profiles from *complete* changes to all our 4 model-types and ranking them accordingly.

An essential point to consider before applying a resampling technique in the differential analysis is the subsampling size: the number of samples per group compared in each iteration. Interestingly, as described in Section 4 of Material and methods, the ability of RDA to detect all 4 model-types is conditioned by this subsampling size, concluding that

a small subsampling size would allow us to disclose all model-types (from *complete* to *mixed*). However, it is important to mention that LIMMA is partially based on t-test statistics, then the sample size of compared samples (subsampling size per iteration) roughly affects to its statistical power (Dobbin and Simon 2005, Stretch et al. 2013). In this way, a greater subsampling size makes RDA more sensible to *complete* changes (often very plane genes: low fold-change among categories), while a lower subsampling size would bring robust *complete* changes and all other model-types (*granularity* of RDA, Section 4 of Material and Methods).

In conclusion, we consider that approaching the feature selection or differential analysis through a subsampling scheme, as provided by RDA, release to gain insight into the significant variability present at any homo- or heterogeneous omic dataset.

2. The predict-response information provided by NSCA in the *h-statistic* notably improves the patient stratification

One of the most common pipelines for differential analysis of omic data described in the literature is composed of **(i)** the first step for select relevant features which discriminate among categories of samples compared and **(ii)** a second step for clustering samples through an unsupervised technique based on previous selected features. Thus, the second step is conditioned by the feature selection process and the raw data input to the unsupervised method. Once we considered that the RDA step provides a precise definition of significant features, we aimed to improve the unsupervised clustering via a proper transformation of raw omic data instead the development of a new clustering technique.

Given a omic dataset, our RDA step will produce a frequency matrix which counts the number of differential events per feature amounted by each subset of samples (Section 4 of Material and Methods). After applying NSCA on this matrix, we would obtain a direct lecture of predictor-response dependence among selected features and samples. Roughly, NSCA would help us to answer this hypothetical question: given a sample(s) present in a feature profile (predictor), would it be responsible for the differential expression of this feature (response). As proposed by Beh and Lombardo (Beh and Lombardo 2014), the **inner product** (p) between features and samples reflects the dependence relationships

behind the inertia analysis (Goodman and Kruskal's τ decomposition) did by NSCA (**Fig. 1-M-6**). Nevertheless, the inner product may excessively weight small *outlier* subgroup of samples (accumulating inertia), which may be inappropriate for a global interpretation of heterogeneity behind an omic dataset. For this reason, we integrated the inner product (p) provided by NSCA and omic dispersion from the mean per feature (d) as described in Section 6 of Material and Methods. The new statistic was called ***heterogeneity statistic*** or ***h-statistic***, which will supersede the raw omic data for a clustering analysis (Section 7 of Material and Methods).

As demonstrated above in Section 2 and 3 of Results, the *h-statistic* enhances the sample stratification, bringing the feature-sample relevance by the inner product slightly corrected by omic dispersion (**Fig. 1-R-3**). Thus, it would reduce the existing overlap of raw omic data depending on the predictor-response information what would improve the output of any clustering technique applied on *h-statistic* (**Fig. 1-R-4**). For example, we could observe how it functions for a single real expression profile like ESR1 gene profile from BCC-2 dataset (**Fig. 1-R-5**). Due to the presence of particularly high expressed genes in Her2-enriched subgroup, like ERBB2, STARD3, PGAP3, GRB7 or CDK12 genes (Section 7 of Results), NSCA would not consider these samples as relevant for differential profile of ESR1. Consequently, *h-statistic* would weight it accordingly, facilitating the subgroup discrimination. In fact, we can observe how different subgroups analysed reduced the value dispersion from raw omic profile to *h-statistic* profile (**Fig. 1-R-5C**).

While there are no similar integrations of raw omic data and other statistics described in the literature, there are a wide range of integrative methods and models for multi-omics approaches (Bersanelli et al. 2016), for example for all these biological processes related to gene expression (mRNA expression, miRNA expression, copy number variation or DNA methylation), several of them aiming to discover hidden biological factors (Ebrahim et al. 2016). Here, the integration proposed is not only very simple and intuitive but also susceptible of being developed in a future towards a multi-omics platform approach, since *h-statistic* was standardized and has no units. In this way, we will briefly discuss the suitability of DECO algorithm for non-transcriptomic platforms in the Section 5 of this Discussion.

3. DECO discloses relevant hidden classes of samples

DECO was successfully applied to a variety of experimental transcriptomic datasets obtained from two different omic platforms: microarrays manufactured by Affymetrix and RNA-sequencing from Illumina (**Table 1-M-2**). Particularly, two out of these five transcriptomic datasets compiled gene expression profiles from breast cancer patients, the most reported cancer in the literature and one of the most frequent cancer in the population. From these two large datasets, we fundamentally found and characterized two **hidden subgroups of samples**: the hospital source (ISPY) for BCC-1 dataset and invasive lobular carcinoma (ILC) for BCC-2 dataset. Initially, we aimed to focus our analysis in assessing the performance of DECO to characterize main subgroups (related to PAM50 classification of samples) if *unsupervised* experimental design was applied.

Surprisingly, DECO revealed ISPY-subclass as a strongly marked subgroup of samples in BCC-1 (**Fig. 1-R-9**), whose technical explication was not reported or introduced by original authors (Hatzis et al. 2011). Additionally, while the characterization of ILC breast cancer subtype is the main purpose of Ciriello et al. paper (Ciriello et al. 2015), they did not deepen in mRNA signature. DECO found a clear signature related to ILC samples in comparison with PAM50 subclasses (**Fig. 1-R-10**) in a very simple pipeline, due to h-*statistic* properties defined above. In fact, all genes found for both hidden subgroup of samples and others related to PAM50 subclasses are coherent and the functional enrichment analyses of different clusters reveals main well-reported signatures (Section 5 and 6 of Results).

Attending to these results and previous work, we conclude that DECO is a very precise and simple bioinformatic tool for disclosing hidden significant subgroup of samples not directly related to original categories, as well as for *outlier profile* detection of single features (**Fig. 1-M-2C/D**).

4. DECO R package is simple and easy to use

To facilitate the use and applicability of DECO, one of the main objectives of this Chapter I was to develop a full and simple R package containing this algorithm. This R package was called *deco*, contains two R vignettes detailing the use of main functions and may be downloaded from our website (http://bioinfow.dep.usal.es/deco/).

Since DECO method is composed of two well-distinguished steps, we proposed to separate R functions in the same way. As detailed above (Section 9 of Results), first R function, called *decoRDA*, was intended to carry out a Recursive Differential Analysis (RDA) based on LIMMA differential analysis technique. Although several different parameters may be input to control the subsampling procedure (detailed within R vignette, **Appendix 1**), this function only requires an omic data matrix (features as rows and samples and columns) to compute this analysis. The experimental design would be defined by the user and enhanced within this function to adapt it for LIMMA design matrixes. In addition, the computational cost and RAM consumption derived from any resampling technique has been significantly reduced through parallelization of calculations (using *snowfall* R package) and saving intermediate results (Section 9.2 of Results). The output will be clear, including the mentioned frequency matrix for NSCA and a feature table summarizing feature statistics among other objects. Thus, any non-expert user can execute it following the R vignette or help pages included in the R package.

Since the second step fundamentally includes a NSCA analysis, the calculation of h-*statistic* and the hierarchical clustering of samples, we integrated these three processes in the second R function, called *decoNSCA*. In this way, the user would also obtain a clear and fast lecture of results produced by DECO in a single R object (of class *deco*) which would include: the table including all feature-statistics, NSCA outputs and hierarchical clustering of both samples and features. The native R functions *print* and *summary* were accordingly modified to provide a summary if a *deco* class object was input. Finally, the basic workflow is completed by the *decoReport* R function which would provide an extended PDF report including relevant plots for the interpretation of results, very easy to execute and where phenotypic information of samples will be input to match significant patterns to original data.

For these reasons, we thought our *deco* R package is functionally coherent, simple and adapted to current computational needs of many bioinformatic users. This R package

was successfully tested on Linux, Windows and MacOSX operative systems and will be also uploaded to CRAN R repository (cran.r-project.org).

5. Suitability of DECO method for non-transcriptomic omic platforms

Since LIMMA is one of the most used techniques for differential expression analysis and has been broadly adapted to R environment, the application on other omic platforms is increasing every year. Indeed, there are many publications in the last decade reporting its use for proteomics differential analysis due to its simplicity, performance and dealing with variability of proteomic data (Margolin et al. 2009, Ting et al. 2009, Pagel et al. 2015, Kuzniar et al. 2017, Basken et al. 2018, Jeannin et al. 2018). In fact, Kammers and colleagues made an extended revision of empirical Bayes method (*eBayes*) of LIMMA and its advantages of use in proteomics and genomics, highlighting its power to detect differentially expressed proteins via inter-experiment variability dealing (Kammers et al. 2015). Although LIMMA is not the first method for differential protein expression analysis, its suitability for proteomic datasets has been broadly demonstrated and, consequently, the DECO applicability on similar datasets to enhance the differential analysis and posterior clustering of samples.

Additionally, other less common omic datasets like miRNA expression or methylation data were also susceptible of being analysed with LIMMA. Since miRNA data may obtained in a similar way that mRNA profiles (microarray platforms like Affymetrix or RNA-sequencing experiments), miRNA analysis has been also related to LIMMA differential analysis in many publications (Thomou et al. 2017, Xue et al. 2017, Mastriani et al. 2018). On the other hand, there are many successful publications where LIMMA have been applied for *supervised* or *multiclass* comparisons of CpG methylation levels in different biological scenarios (Stefan et al. 2014, Wockner et al. 2014, Johnson et al. 2017, Saito et al. 2017, Martorell-Marugan et al. 2018). Interestingly for these two omic platforms, the authors possess the advantage of using the same technique for mRNA levels and miRNA expression or DNA methylation data, which is totally complementary in terms of interpretability of results (gene expression).

In conclusion, since LIMMA has been broadly applied to several different omic platforms (i.e. proteomic, miRNA and DNA methylation data) and it is the functional core of

RDA step, DECO is greatly suitable to be applied on these platforms. DNA methylation and miRNA expression levels are two complementary data to mRNA expression, providing a major insight into the biological scenario behind a particular omic study. As such, we have in mind an advanced development of DECO method for the integration of these particular gene expression related platforms (mRNA with miRNA or DNA methylation) for a same set of samples because the transformation of raw omic data into differential events first and h-*statistic* later is no dependent of any initial unit or statistic.

CHAPTER II

Cohesiveness: a simple and non-parametric statistic for platform-independent feature selection with omic data

BRIEF SUMMARY

Throughout this Chapter II, we propose a novel and simple **non-parametric statistic to measure the proximity of the samples belonging to a category within a quantitative variable**. According to the current results obtained from omic data analysis, our statistic, called ***cohesiveness***, could be used to: (i) select the best features to differentially characterize any category of samples and (ii) determine the most stable feature's patterns corresponding to a category of samples.

INTRODUCTION

Our ability to compile large amount of phenotypic information for clinical or biological samples is being improved every day. Nowadays, we enjoy a great variety of omic technologies to collect high-dimensional data (transcriptomics, proteomics, genomics, metabolomics, etc.) from biological samples of different organisms. As mentioned in Chapter I, the availability of phenotypic data is essential for the proper interpretation of results and posterior classification of individuals in a previous characterized group, especially in the context of *precision medicine* (**Fig. 1-I-1**). However, while we mainly focused on the analysis of heterogeneous (and homo-) omic data in a particular biological context in Chapter I, now, we focus on how dealing with large phenotypical datasets, when high-dimensional omic data are available, to select not only discriminant but also stable patterns within biological features.

1. Feature selection in bioinformatics

Feature selection (FS) has been subjected to numerous studies and development of methods for multiple scientific fields, considered as a crucial step for posterior prediction of particular conditions or group of samples or conditions. Particularly, the increasing trend of high-dimensional data produced by biomedical and biological research has notably contributed to the development of new tools and applications for omics data. For instance, a simple search of "feature selection" in PubMed database (www.ncbi.nlm.nih.gov) returns 4328 studies of which 2488 date from the last five years.

Here, they dealt with "***large p, small n***" problem (where *p* corresponds to independent features and *n* samples), very common in biomedical research when omic data was generated. Especially, it has been broadly examined by machine learning methods as a first step of algorithms to reduce the number of biological features (Wang and Fu 2006) (gene expression, methylation, protein concentration, etc.) measured without altering the intrinsic and relevant structure of omic dataset (Hira and Gillies 2015).

Consequently, the feature selection procedures implemented in any bioinformatics pipeline should be carefully chosen to enhance the posterior interpretability of the results (Saeys et al. 2007). Although there are other problems associated with any classification and feature selection procedure, we would like mentioning how often the big size of omic and biological data render classification algorithms useless due to the computational cost, particularly all those which include combinatorial procedures or randomization of features (Chen et al. 2012, Bolón-Canedo et al. 2015, Kourou et al. 2015).

Interestingly, most of these approaches has been proven on microarray datasets due to its popularity, easy accessibility, and management (Tan et al. 2014, Hira and Gillies 2015). Moreover, application of feature selection procedures is preferred to data or dimensionality reduction techniques, since it is easier to interpret original omic data instead transformed data (Wang et al. 2016). This kind of data is intended to contain ***flat features*** (**Fig. 2-R-1**), where features are assumed to be independent. Classically, feature selection methods for *flat features* have been split into:

- **Filter:** evaluates the association of each feature with a categorical variable through a statistical significance analysis without using any classification algorithm (Liu and Setiono 1996). It could be *univariate* or *multivariate* (when multiple features are evaluated). Fisher's score, ReliefF, mRmR, Wrank or Trank methods are several classical examples of *filter* methods (Bradley and Mangasarian 1998, Nigam et al. 2000, Baldi and Long 2001, Liu et al. 2002, Robnik-Šikonja and Kononenko 2003).
- **Wrapper:** evaluates a set of features considering a predefined classifier to select a subset of relevant and non-redundant features (**Fig. 2-R-2**). However, its computational cost is higher than filter methods (Yu and Liu 2004). Random forest (Breiman 2001) or posterior modifications of random forest, such as like Recursive Feature Elimination (RFE) (Svetnik et al. 2004), are highly-studied and applied wrapper methods for feature selection.
- **Embedded:** evaluates the features at the same time that constructs the predictor classifier, combining the advantages of both filter and wrapper methods. Moreover, there are less computational intensive than wrapper algorithms (Liu and Yu 2005). Lasso or Supper Vector Machines (SVM) are examples of embedded methods used with omics data.

Alternatively, there are feature selection methods for ***structured features***, where features are also independents but there are relationships among them, retrieving a simple or complex scenario. Group Lasso, ANOVA or multinomial regression are several examples of this second type of feature selection, which take into account these relationships. However, we will focus on methods for *flat features* due to its similarity with the method proposed along of this Chapter II.

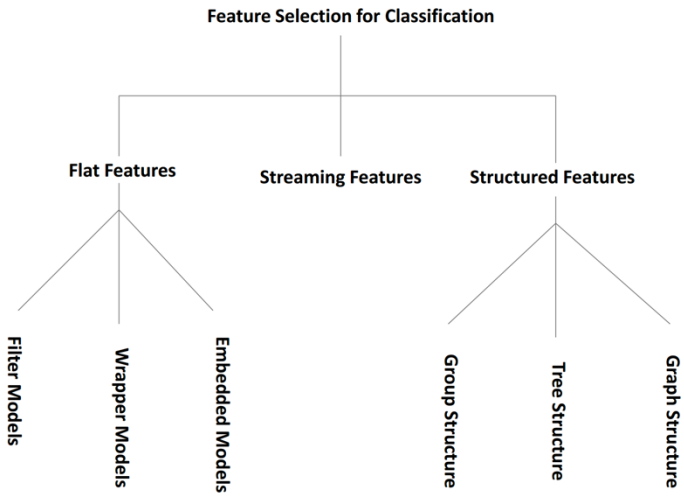


Figure 2-I-1. General classification of feature selection methods. Scheme from (Tang et al. 2014).

Apart from feature selection methods, there are methods focused on feature extraction, as mentioned above. These methods are based on the supposition than given a large matrix (i.e. omic data) following the common "***large p, small n***" scheme, we may reduce the redundant information via transformation of original data into new features, which would comprise relevant information in a reduced set of features. This concept is known as **dimensionality reduction**. Several well-known techniques are Principal Component Analysis (PCA), Multidimensional scaling (MDS), Linear Discriminant Analysis (LDA) or Non-Negative Matrix Factorization (NMF). However, these methods are out of the scope of this Chapter II since they are not related with the simple method proposed.

Interestingly, for feature selection methods, the choice between *supervised* or *unsupervised* design (no classes of samples to compare) within a pipeline establish the main bottleneck to choose a proper technique. Currently, omic studies gather detailed

knowledge on the samples analysed, composed by categorical data, leading to expect changes among different states or classes (*supervised*) to define its biological context (Bermingham et al. 2015). In this way, feature selection procedures within a *supervised* pipeline would try to avoid redundant information while maximizing differences between classes (Koch 2014, Bolón-Canedo et al. 2015, Singh and Sivabalakrishnan 2015).

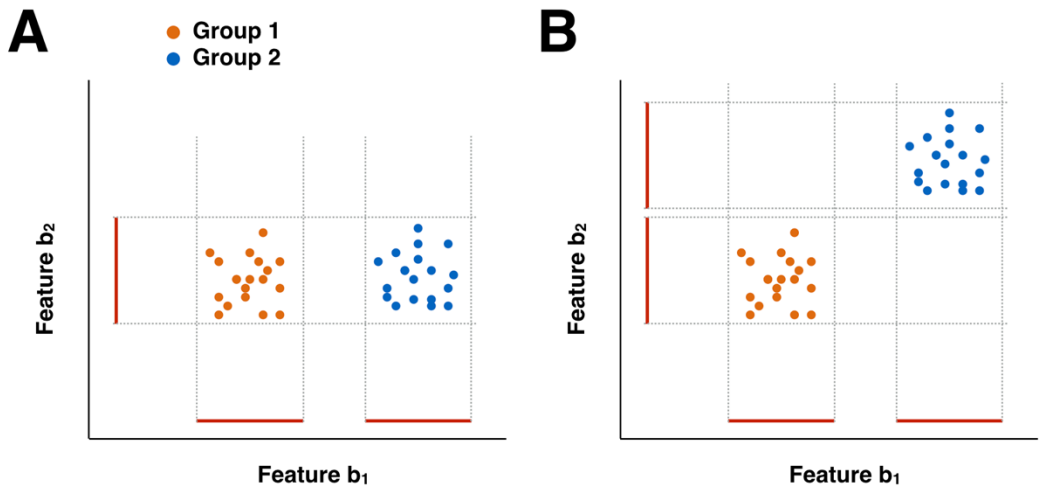


Figure 2-I-2. Adapted figure (Ang et al. 2016). **(A)** Feature b_2 is irrelevant and non-informative for the classification of group 1 and 2 of samples. **(B)** Feature b_1 and Feature b_2 are redundant since they provide the same information for the classification of group 1 and 2.

2. Categorical data analysis in bioinformatics

Categorical variables have two primary types of scales: **(i)** variables having categories without a natural ordering are called nominal; **(ii)** and categorical variables do have ordered categories are called ordinal. In clinical studies, phenotype encompasses a wide range of data: disease risk, ethnicity, age, sex, medical treatments or pathology classification. These categories would allow to the scientist to discriminate among samples in order to characterize each one with a subset of relevant or differential features (Berger et al. 2013). In order to characterize each categorical variable by associating certain continuous variables, multiple statistical approaches have been traditionally applied in bioinformatics, such as discriminant analysis, MANOVA, correlation analysis, or any of differential analyses mentioned in Chapter I. Thus, the feature selection would be directly conditioned by samples' categories and comparison among them, obtaining variable results

depending on the experiment design. However, there is a previous question which anyone carrying on a statistical analysis of omic data should answer: could our categories follow an order? Indeed, the user should choose a proper method, like stage-like (Aibar et al. 2016) or time-course approaches (Nueda et al. 2009), and avoid classical methods to improve the interpretation of the results.

Regarding *unsupervised* methods, clustering is the most widely used technique to group similar samples in biological data, leading to significant discoveries at the beginning of omic analyses (Perou et al. 2000). However, our clustering method choice and the high dimensional nature of biological data compromise the meaningful of the results (Ronan et al. 2016). Any previous normalization or scale transformation of the date could alter the result and, even obtaining a nice clustering result, its biological significance is not always well-determined. In fact, many *unsupervised* feature selection methods are based on data transformation (i.e. combining two or more original biological features in a single one) to enhance the results.

For all those reasons, given categorical and omic data, we will propose a non-parametric, simple and easily-understood statistic -called **cohesiveness**- to assess the ability of any biological feature for classifying samples into phenotypic categories. The **cohesiveness** takes into account both principal concepts of any classifier: contribution (to classification purpose) and interaction (among categories of samples), using two or more categories of samples as basis for its calculations but not taking into consideration the redundancy among features (**filter method for flat features**). The algorithm to calculate the cohesiveness statistic has been written in the R language. The original R script can be found as **Appendix 2**.

MATERIAL AND METHODS

1. Experimental datasets

The gene expression data from the Genotype-Tissue Expression (GTEx) project was used (GTEx v7), which is a RNA-sequencing dataset (Illumina TrueSeq RNA-sequencing) comprising the gene expression signal of 56202 genes (ENSEMBL IDs) and 11688 samples from 54 conditions (51 tissues and 3 derived cell lines). Expression and phenotypical data (www.gtexportal.org/datasets) can be downloaded as TPMs (Section 1.2, Introduction of Chapter I) from GTEx website portal (Consortium 2013). Additionally, a single-cell RNA-sequencing dataset with information of different cell-types of healthy human brain (fetal and adult) was used (Darmanis et al. 2015).

As microarray datasets, an Affymetrix HGU133 Plus2.0 gene expression dataset from patients of diffuse large B-cell lymphoma was used. Additionally, another dataset compiling transcriptomic data from several brain regions was also used (Kang et al. 2011). In both cases, we mapped from Affymetrix probesets into ENSEMBL IDs using BrainArray CDFs (Dai et al. 2005). Then, RMA normalization method was applied (Irizarry et al. 2003).

Table 2-M-1. Transcriptomic datasets used to compare different methods for feature selection in Chapter II.

Disease	Tissue	Platform	Number of samples	Normalization	Subtypes	Experimental design	Reference	Year
Healthy Human Brain (Brain-1)	Pos-mortem sample	Affymetrix Human Exon 1.0	599	RMA log2(signal)	**6 major brain regions - 16 specific brain regions**	**Multiclass**	GSE25219 (GEO database)	2011
Healthy Human Brain (Brain-2)	Adults: epilepsy surgeries Fetal: elective abortions	Illumina MiSeq - Single Cell RNA-seq	466	log2(CPM + 1)	**9 cell types**	**Multiclass**	GSE67835 (GEO database)	2015
Diffuse Large B-Cell Lymphoma (DLBCL)	Bone Marrow Mononuclear cells (BM-MNCs)	Affymetrix Human Genome U133 Plus 2.0	414	RMA log2(signal)	**ABC (n=167), GCB (n=183) and Unclassified (n=64)**	**Multiclass**	GSE10846 (GEO database)	2008
GTEx	Human tissues and cell lines	RNA-seq	11688	log2(TPM + 1)	**51 human tissues and 3 cell lines**	**Multiclass**	https://www.gtexportal.org/home/	2018

2. Methods for feature selection

Along of this Chapter II, several feature selection methods will be used to compare the performance against the proposed *cohesiveness* statistic. For *filter* methods, we selected several different modifications of ReliefF method (Robnik-Šikonja and Kononenko 2003), different metric distances like Euclidean, Hellinger, Angle, DKM (Dietterich, Kearns and Mansour distance) or AUC distances, or probabilistic measurements like Information Gain, Gini's ratio, Gain Ratio or Accuracy. All these feature selections methods are implemented in the *CORElearn* R package. Apart from these feature selection methods, we also evaluated LIMMA (F-test; from *limma* R package) and Recursive Feature Elimination (RFE; *caret* R package) based on Random Forest.

All these methods are referenced in **Table 2-M-2**.

3. Input data

As hinted above, the high-throughput omic data obtained from clinical or biological studies are accompanied by a deep phenotypic characterization of n samples, most of them associated with an interesting factor F. In this way, any factor or categorical variable F could be composed by F categories ($k \geq 2$), which can be characterized through ***cohesiveness*** analysis of biological features. In fact, all samples could belong to one category within a F factor, so we will dispose a omic matrix A associated with n samples (rows) and m omic features (columns) plus one -or more- F factor(s), following the next notation:

$$A = \left[a_{ij}\right]_{n\cdot(m+l)}, 1 \leq i \leq n; 1 \leq j \leq m+l \qquad \text{(Eq. 2.1)}$$

where, given l factors ($l \geq 1$), each cell a_{ij} corresponds to the value of i^{th} sample for the j^{th} feature ($1 \leq j \leq m$) plus the phenotypic information per F factor ($m < j \leq m+l$).

4. Cohesiveness: gap definition and probability function

Let $x = \{1,2 \dots n\}$ be the increasing natural rank of n elements, which could belong or not to a given K category. The subset of elements from x belonging to this K category could be defined as $x^K = \{x_1^K, x_2^K \dots x_r^K\}$, where $r \leq n$. Then, the difference or gap between

two consecutive elements of x^K is calculated as follows:

$$d_i^K = x_{i+1}^K - x_i^K, 1 \leq i \leq r-1 \quad \text{(Eq. 2.2)}$$

Then, the minimum value of the variable $D^K = (d_i^K)_{r-1}$ will be 1 and the maximum value will be $n-r+1$. Given an element of x^K and a value $d_i^K = d$, there are $(n-r-d)$ positions within x in which to place the first element of x^K at a distance d from the second element and the remaining $r-2$ elements could be placed at empty $n-d$ positions (**Fig. 2-M-1B**). Considering that the elements are uniformly distributed and unique in the set x, the probability function of D^K will be:

$$f(D^K) = P(D^K = d) = \frac{\binom{n-d}{r-1}}{\binom{n}{r}} \quad \text{(Eq. 2.3)}$$

for $1 \leq d \leq n-r+1$ and 0 otherwise. The denominator corresponds to the total number of combinations of r elements within a n elements finite vector, while the numerator corresponds to fix one element of x^K and let free the other $r-1$ elements within a $n-d$ finite vector (**Fig. 2-M-1B**). Using Wolfram Mathematica software, we checked that $f(D^K)$ is a probability function because the sum of all probabilities is 1. Additionally, we calculated the mean and the variance of this distribution:

$$E(D^K) = \frac{n+1}{r+1} \quad \text{(Eq. 2.4)}$$

$$var(D^K) = \frac{(n+1)\cdot(n-r)\cdot r}{(r+1)^2\cdot(r+2)} \quad \text{(Eq. 2.5)}$$

Extensively, given a feature with x elements and K categories, we calculated a Z-score Z^K for each K category based on observed mean $\overline{D^K} = \frac{1}{r-1}\cdot\sum_{i=1}^{r-1} d_i^K$, theoretical mean $E(D^K)$ and theoretical standard deviation $\sigma(D^K)$ to estimate the probability of obtaining a specific distribution of gaps d:

$$Z^K = \frac{\overline{D^K} - E(D^K)}{\sigma(D^K)/\sqrt{r}} \quad \text{(Eq. 2.6)}$$

Due to the classical Central Limit Theorem (CLT) (Billingsley 1995), this Z-score Z^K follows a normal distribution $N(0,1)$. Thus, we could use this normal distribution to assign a left-tailed p-value for each Z^K, in order to compare the consistency of all K cohesiveness

given feature x (**Fig. 2-M-1A**). Then, lower p-values indicate that this feature groups closer the elements of a category ($\overline{D^K}$ would be significantly lower than $E(D^K)$).

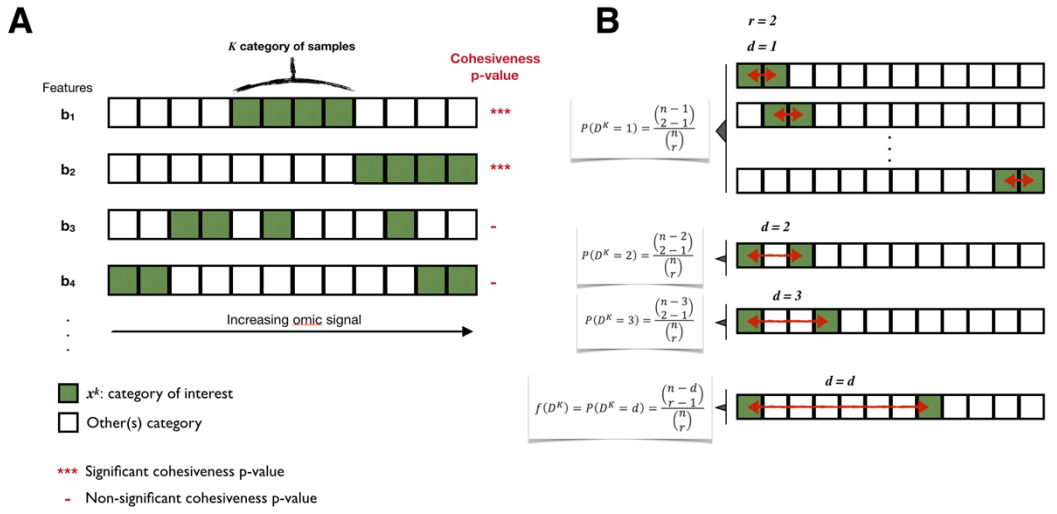


Figure 2-M-1. Theoretical representation of cohesiveness statistic and gap probability. **(A)** Cohesiveness statistic is intended to score higher those tight categories for rank positions, independently of other categories of samples. **(B)** Graphical representation of how the probability of a specific gap between two elements of x^K is calculated.

5. Cohesiveness: optimal significance threshold for multiple categories

After calculating Z^K and its corresponding p-value for each K category of a given feature, we propose a summarization of these p-values to accordingly rank all biological features analyzed. Then, p-values could be summarized using Fisher's combined probability test (Fisher 1925) or Stouffer's Z-score method (Stouffer and Hovland 1949) providing a multiple summarized cohesiveness score C_b per biological feature b and, posteriorly, calculating a p-value for these summarization methods.

Adjustment of p-value for multiple comparisons was proposed using False Discovery Rate (FDR) (Benjamini and Hochberg 1995) on resulting Fisher's or Stouffer's p-value.

6. Cohesiveness: reducing the redundancy of selected biological features

Cohesiveness is intended as a *filter* feature selection method which does not deal with the redundancy of selected features because it does not estimate the relationship or similarity among the biological features analyzed. As a simple estimation of redundancy for classification purposes, we propose to integrate both *cohesiveness* measurement and Spearman's Correlation Coefficient (SCC) among biological features for reducing the redundancy of highly scored features by *cohesiveness*. The correlation matrix of SCCs, called S^b, among all biological features m analyzed would be:

$$S^B = \left(s_{ij} = corr(b_i, b_j)\right)_{m \cdot m} \quad \text{(Eq. 2.7)}$$

Given the square matrix S^b, we could integrate each biological feature vector (row or column) of S^b and the vector of summarized cohesiveness C_b, by multiplying both values, obtaining a corrected cohesiveness. Then, given a biological feature b, we could search the complementary biological feature C_i^* among j elements $(j \neq i, j \leq m)$, which would be characterized by minimum value of:

$$C_b^* = min(C_j \cdot S_{bj}^B) \quad \text{(Eq. 2.8)}$$

Thus, C_b^* resembles the most relevant complementary feature for biological feature b, which integrates both negative Spearman's Correlation Coefficient (SCC) and a high summarized cohesiveness score C_b. Later, we counted the absolute frequency A_b for any biological feature b of appearing as the complementary biological feature of another:

$$A_b = \sum_{b=1}^{m} [C_b^* = b] \quad \text{(Eq. 2.9)}$$

Finally, we added C_b and A_b to calculate a final cohesiveness score C_b^F per biological feature b intended to be used for classification purposes, which allow ranking of all biological features accordingly:

$$C_b^F = C_b + A_b \quad \text{(Eq. 2.10)}$$

7. R script

All calculations required for *cohesiveness* analysis of a given omic matrix and categories of samples are provided as a single and simple R script in **Appendix 2**. The script includes one functions to calculate final *cohesiveness* statistic (trimmed or complete) and p-value per category (Z^K) of samples per feature. Additionally, it also includes an option to compute the summarized *cohesiveness* score C_b according to Fisher's combined probability test or Stouffer's Z-score method.

Table 2-M-2. Methods for feature selection on omic data used in Chapter II.

Method	Concept	Type	R package
Recursive Feature Elimination (RFE)	Random Forest	**Wrapper method for flat features**	**caret**
F-test (LIMMA)	F-divergence based on linear model	**Filter method for flat features (*univariate*)**	**LIMMA**
DistAUC	Different metric used as distance for similarity		**CORElearn**
DistEuclidean			
DistHellinger			
DistAngle			
DKM			
InfGain	Probabilistic models to compare categories of samples		
GainRatio			
Gini			
Accuracy			
Relief	Relief-related methods Different metric used as distance		
ReliefFequalK			
ReliefFexpRank			
ReliefFbestK			
ReliefFmerit			
ReliefFdistance			
ReliefFsqrDistance			
ReliefFexpC			
ReliefFavgC			
ReliefFpe			
Cohesiveness	Original statistic		***Present work*** *(Appendix 2)*
Cohesiveness trimmed	Trimmed version (1% by default)		
Cohesiveness + Spearman correlation	Original statistic and Spearman correlation	**Filter-like method for flat features (*multivariate*)**	
Cohesiveness trimmed + Spearman correlation	Trimmed version and Spearman correlation		

RESULTS

1. Cohesiveness detects stable patterns within variable data

The ***cohesiveness*** statistic is intended to disclose not only differential features among multiple phenotypical categories but also stable feature patterns within homo- or heterogeneous data. In last ten years, there have been several publications related to differential stability or differential stable expression (Hawrylycz et al. 2015), also referred as reproducible gene expression patterns (Huang et al. 2016), relating them to housekeeping functions or preserved and important functions within a specific biological context (Shaw et al. 2011), such as tissues or cell types. These approaches were intended to determine which genes exhibit a reproducible and stable behaviour among the analysed categories.

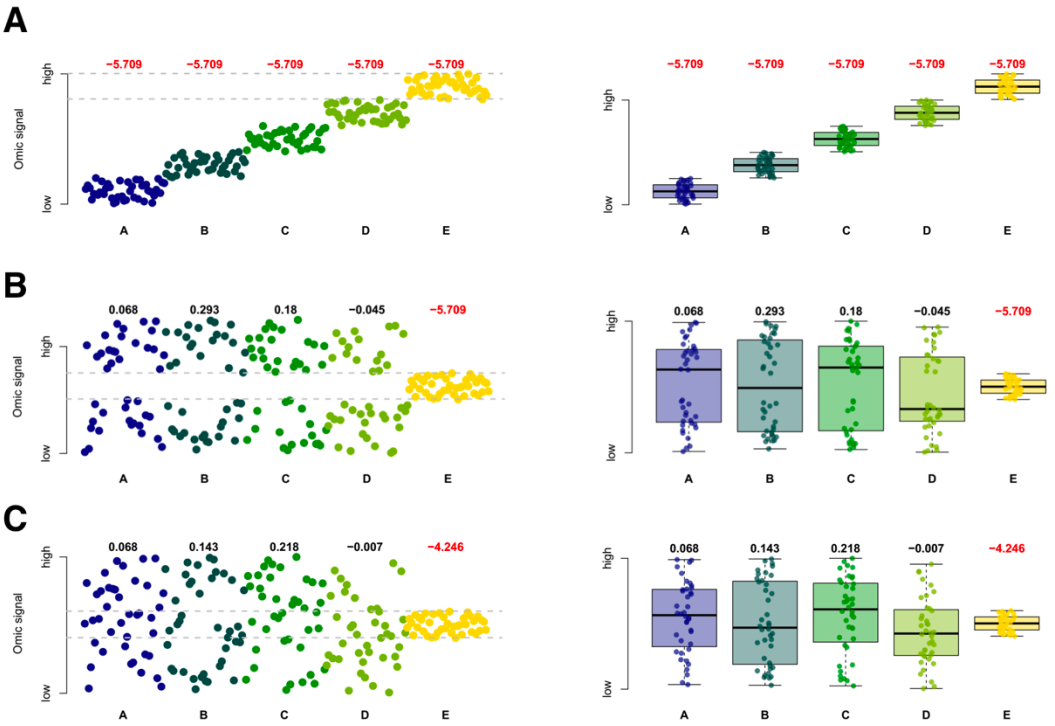


Figure 2-R-1. Example of discovery of stable patterns within very variable signal using *cohesiveness statistic* (Z^K, showed at top of each category and plot. Red values refer to significant ones). **(A)** Non-overlap signal among categories, providing a perfect differential scenario (TP). **(B)** Overlapping (100% of size of D; n = 40) among A-D categories, while E remains non-overlapping. **(C)** Cohesiveness statistics remain high even for E category if the full overlapping signal (100% of size of D; n = 40) is present.

Interestingly, our statistic (Z^K, eq. 2.6) could be used in the same way, identifying stable patterns across categories within any feature. **Figure 2-R-1** shows several examples of stable pattern detection conducted using *cohesiveness* statistic on three artificial feature profiles. We demonstrated how *cohesiveness* is able to find both differential (**Fig. 2-R-1A**) and stable feature patterns within variable or random data (**Fig. 2-R-1B and C**). These stable feature patterns may be no interesting for classification purposes based on mean, median or population's statistics, but they may point to non-deregulated signal in particular conditions or biological scenarios (Prieto et al. 2006).

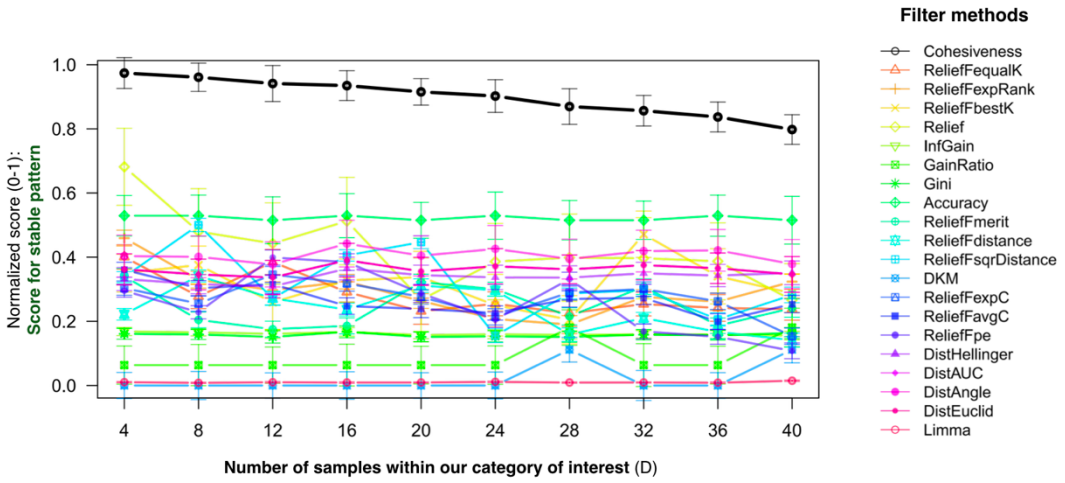


Figure 2-R-2. Stable feature detection by *filter* methods, F-test (LIMMA) and *cohesiveness* statistic (C_b). Each method's differential scores ranged between 0 and 1, corresponding 1 to maximum score (positive feature) and 0 to random scores. Differential Z^K scores from random of this category of interest (category D) were also normalized between 0 and 1. Segments over points represent standard deviation of data.

Figure 2-R-2 displays the performance of all different *filter* methods evaluated (included in *caret* R package) and F-test (LIMMA) against the stable artificial feature (**Fig. 2-R-1A** and **B**). This artificial dataset is based on features shown in **Figure 2-R-1**, containing a *positive* artificial feature with a differential pattern for all four categories of samples (**panel A**), a stable artificial feature for one out of four categories of samples (E positive; A, B, C and D negative; n = 200 samples; **panel B** to **C**), and 100 *random* features for all the categories (n = 200 samples). Additionally, we were iteratively including samples from A-D categories within the range of values of E category (x axis of **Fig. 2-R-2**). Then, we applied all *filter* methods and F-test to these 102 artificial features and calculated all scores for *positive*, *random* and *stable* features. Original scores from *caret* R package

(which includes all these methods) were normalized between 0 and 1 values (1 for *positive* and 0 for *random* features), then we proceeded in the same way with the *cohesiveness* statistic (C_b) for a proper comparison.

As we can observe in **Figure 2-R-2**, there was a great difference among the performance of any of these methods and the *cohesiveness* statistic proposed for disclosing a stable pattern. Here, any value close to 1 for the stable feature would indicate that it has been similarly scored to a *positive* feature (differential feature for all categories), which was used to set up the maximum method's score. Thus, we can observe how *cohesiveness* score for the stable feature and category E reaches values close to 1, indicating that it has been identified as a *positive* feature, while other methods score this stable feature pattern lower than the differential *positive* pattern. All method's scores decrease, as expected, when the number of samples disrupting this stable pattern increases (from 4 samples to 40 samples, similar size to our category of interest E).

2. Cohesiveness as feature selection method for multiple classification of categories

Once we detailed how cohesiveness statistic is able to interpret any single feature, we may define it as a **one VS all** differential method. Because great absolute values of cohesiveness would discover patterns where any or a few samples belonging to other categories are present within, this statistic should be understood as a tool for discovering patterns where a category is significantly different from all others. Thus, we would assume our category is significantly placed alone in a specific range of values (i.e. omic data).

Similarly to Hawrylycz et al. paper (Hawrylycz et al. 2015), we also aimed to determine which genes show a stable expression among 16 brain regions studied by Kang et al. (Kang et al. 2011), as application on a large experimental dataset. These regions were properly analysed by authors, collecting 599 different samples from adult individuals, and grouping these samples into 6 major regions (AMY: amygdala [n=35], CBC: cerebellar cortex [n=34], HYP: hippocampus [n=35], MD: mediodorsal nucleus of thalamus [n=33], NCX: neocortex [n=429], STR: stratum [n=33]). First, we filter 39300 genes into the 10112 most variable genes (IQR filter) and we applied all feature selection methods (**Table 2-M-2**) to select features associated with 6 major regions. *GlobalTest* was used as outcome test to

assess the classification's rate of each signature, as used in Chapter I.

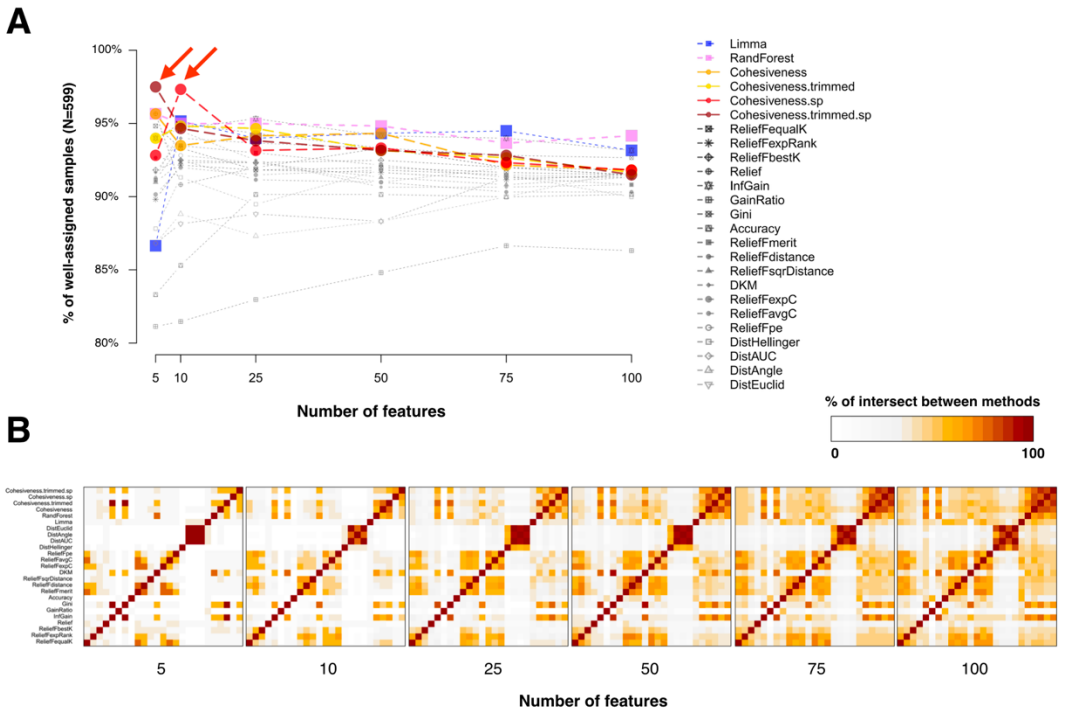


Figure 2-R-3. Classification's rate of all different methods compared, including *cohesiveness*, on Kang's dataset of 6 human brain regions. **(A)** Plot showing the classification's rate (percentage of well-assigned samples) of each method at different number of features, using *GlobalTest* as outcome test to assess the ability of selected features to classify samples. Grey-like colours correspond to methods for *flat* features, blue to F-test from LIMMA, pink to RFE (Random Forest) and orange-red scale to *cohesiveness* alternatives. **(B)** Heatmaps showing the percentage of intersect among gene signatures selected by each method, at different size of signatures.

Figure 2-R-3 displays the performance of different methods for feature selection on 6 major regions of Kang's dataset. Since each method evaluated and ranked all features according to their ability to discriminate among these brain regions, we cut these gene signatures at different sizes to properly compare all them using *GlobalTest* (threshold for well-assigned samples: p-value ≤ 0.001). We expected that Random Forest (RFE) was one of the best methods along different size of signatures, because it takes into account the existing relationships among features (wrapper method) while others do not. RFE outperforms all methods with its top-5 selected features except *cohesiveness* statistic (trimmed = 1%), which classifies 12 more samples than RFE (**Fig. 2-R-3A**). Surprisingly, top-5 from F-test (LIMMA) shows one of the poorest performances. All other methods for *flat* features shows a similar range of classification along different sizes of signatures (90-95% of samples), while *GainRatio* was the worst feature selection method on Kang's

dataset (80-85%).

Once the classification's rates from this large dataset are analysed, we firstly wondered how different are gene signatures among all methods used. **Figure 2-R-3B** tries to compile this information, showing the percentage of pairwise intersect among gene signatures of each method in several heatmaps (one per number of features). Main clusters of intersect correspond to Hellinger, AUC, Euclidean and Angle distances methods, while top-right corner correspond to different *cohesiveness* versions.

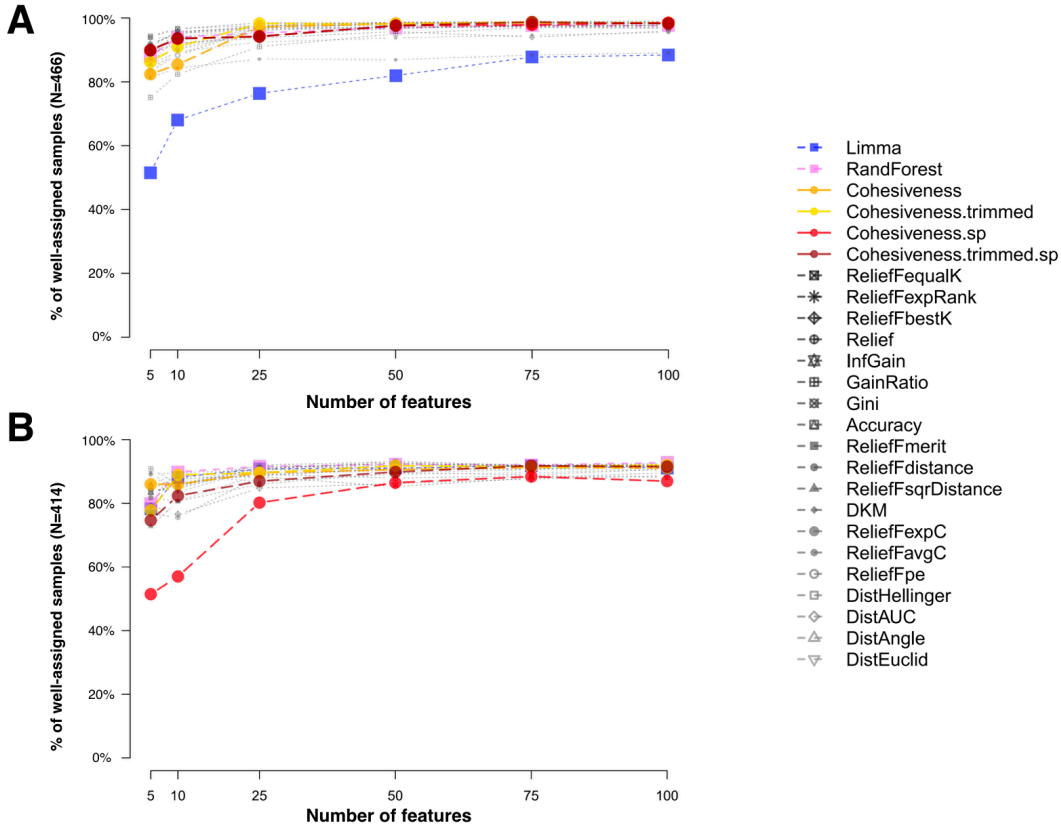


Figure 2-R-4. Classification's rate of different methods for Brain-2 dataset -9 brain cell types- **(A)** and DLBCL -3 different subtypes of pathology- dataset **(B)**.

Additionally, we tested all these methods for feature selection on two more large transcriptomic datasets in order to compare the classification's rates (threshold for well-assigned samples: p-value ≤ 0.001). We used a large dataset from different human brain cell types (astrocytes [n=62], endothelial [n=20], fetal quiescent [n=110], fetal replicating [n=25], hybrid [n=46], microglia [n=16], neurons [n=131], oligodendrocytes [n=38] and OPC

[n=18]), obtained by single-cell RNA sequencing (Darmanis et al. 2015), a RNA-seq variant which allows to sequence all RNA present in a unique cell. Since the single-cell RNA-sequencing technique is very precise, this data is commonly characterized by a notably presence of no reads, making more difficult the statistical approach to differential analyses. **Figure 2-R-4A** shows the classification's rates for this dataset (Brain-2 dataset, **Table 2-M-1**) based on *GlobalTest* results and 9 different categories of samples (brain cell types). LIMMA shows the lowest performance, while trimmed version of cohesiveness statistic and RFE are almost overlapping at all different sizes of signatures tested. In fact, both showed the greatest performances at greater sizes of signatures (75 and 100 features).

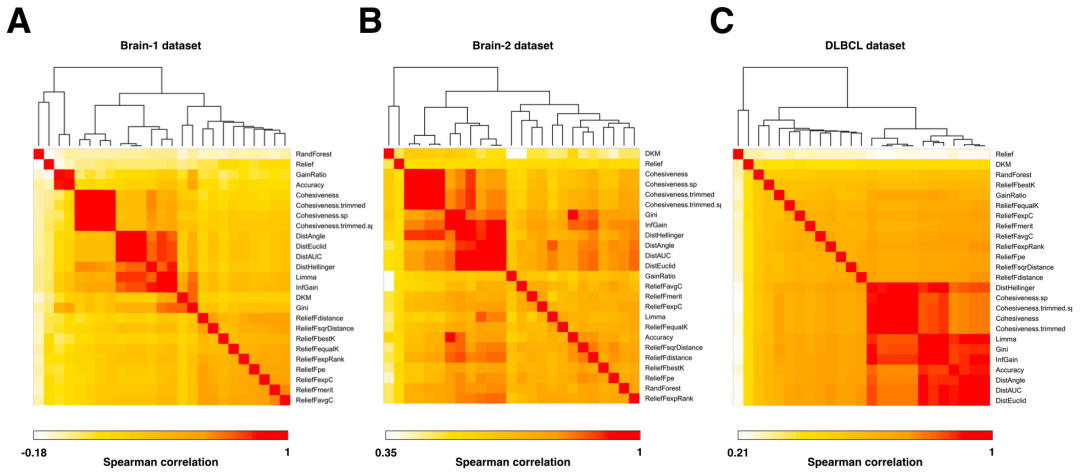


Figure 2-R-5. Heatmaps showing Spearman correlation coefficients (SCC) among different feature rankings obtained by each method (Table 2-M-2) per dataset (Table 2-M-1). **(A)** Brain-1 dataset (Kang et al. 2011). **(B)** Brain-2 dataset (Darmanis et al. 2015). **(C)** DLBCL dataset (Lenz et al. 2008).

Alternatively, we also tested these methods in a large dataset of Diffuse Large B-Cell Lymphoma (DLBCL dataset, **Table 2-M-1**) with only 3 subtypes or categories of samples along 414 different patients (Lenz et al. 2008). As we can observe in **Figure 2-R-4B**, original cohesiveness statistic (without trimmed) performs worse than other methods, what we assumed that may be due to large size of categories (only 3 categories). The probability of finding a lower gap for large categories of samples is greater than for smaller categories, due to the degree of freedom of each case. However, trimmed version of cohesiveness (1% trimmed) outperforms RFE and LIMMA, supporting our previous idea that trimmed version and/or the additional Spearman's correlation attribute A_b could enhance the feature selection at some scenarios, where there are categories with large

number of samples (Brain-1 dataset if 6 major brain regions or DLBCL dataset).

An easy way to compare the similarity among different methods compared (**Table 2-M-2**) could be calculate the pairwise Spearman's Correlation Coefficient of feature's rankings obtained. We have already observed how feature's signatures are related at top levels (**Fig. 2-R-3B**), while we can now compare the global rankings produced by each method in each dataset analysed (**Table 2-M-1**), displayed as heatmap in **Figure 2-R-5**. The four cohesiveness versions are very related each other in all datasets and, posteriorly, closely related to probabilistic and distance methods, like Gini, InfGain or DistHellinger.

3. Cohesiveness finds tissue-specific genes: differential and stable patterns

Genotype-Tissue Expression (GTEx) consortium tries to compile, in a single dataset, the gene expression data from multiple different human tissues using RNA-sequencing technology (Consortium 2013). Nowadays, this huge resource embraces 30 different major tissues and 54 more specific subtypes of human tissues, amounting 11688 different samples. It was intended to help in the identification of tissue-specific gene expression levels and, consequently, their relationship with different biological functions. One of the major issues approached first by authors using GTEx was defining a threshold to identify tissue-specific genes. In this way, Sonawane and colleagues (Sonawane et al. 2017) proposed to calculate a score for tissue-enriched genes (s_b) based on standardized values from median expression of each tissue and all samples:

$$S_{b,tissue} = \frac{med(exprs)_{b,tissue} - med(exprs)_{b,All}}{IQR_{b,All}} \qquad \text{(Eq. 2.11)}$$

Then, they put a threshold of $S_{b,tissue} > 2$ for defining any gene as tissue-enriched for a particular tissue. Sonawane's score assumes that a tissue-enriched gene should show higher expression values than for all other tissues, but it obviates whether the gene is also expressed in other tissues. Instead, other authors calculated the number of tissue-enriched genes based on the number of tissues showing a higher gene expression level than a minimum threshold (i.e. mean(FPKM) > 1) (Uhlen et al. 2015). Thus, they assumed that a hypothetical tissue-enriched gene would not show any expression signal in other tissues. As demonstrated above, *cohesiveness* would highlight both stable and/or

differential features when multiple categories of samples are known. It could be applied as measurement to define tissue-specific or enriched genes, just considering each Z^K calculated per feature.

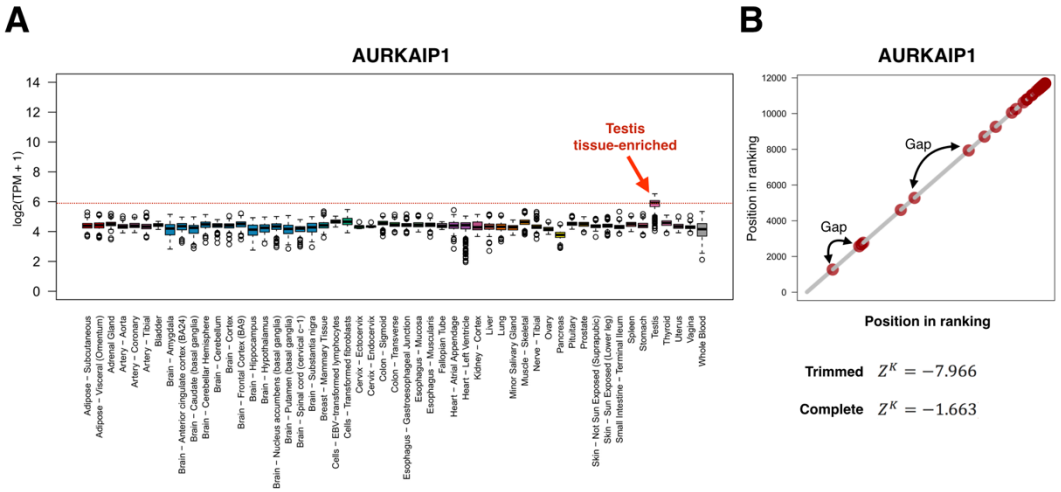


Figure 2-R-6. Expression profile of AURKAIP1 gene per tissue of GTEx dataset. **(A)** Boxplot per tissue using log2(TPM + 1) as unit of gene expression signal. **(B)** Positions in ranking (non-parametric) of testis' samples in red, showing how little deviation in parametric scale (A) may result in large gaps (B).

Since cohesiveness statistic is measuring the mean of all gaps (Eq. 2.4) between elements belonging to the same category, it is sensible to large datasets where n is very high and r is very low. Here, the cohesiveness statistic may be biased by outlier points amounting a great gap from previous sample. For this reason, we also considered trimmed means, replacing $E(D^K)$ by $E_{trim}(D^K)$ to remove those outliers in posterior calculations, including theoretical mean and variance.

Figure 2-R-6A shows an example of tissue-enriched profile (AURKAIP1 gene) for testis affected by a little group of samples, where original cohesiveness statistic assigned as a slightly significant feature with p-value = 0.0438 ($Z^K = -1.663$) while trimmed cohesiveness statistic (trimming 1%, two-tailed) scores it as a very significant tissue-enriched feature with p-value = $8.13 \cdot 10^{-16}$ ($Z^K = -7.966$). In fact, we represented in **Figure 2-R-6B** how samples belonging to testis tissue are distributed by AURKAIP1 gene, showing how a minor portion of gaps have huge values due to the dispersion of several samples and the low dispersion of global expression profile of AURKAIP1.

AURKAIP1 was also correctly identified by Sonawane's score (Sonawane et al. 2017), but there are relevant cases where this score fails because the dispersion of the data and the range of values (close to zero). Interestingly, we demonstrated above how cohesiveness is able to find stable patterns within notably variable omic profiles where other methods fail (**Fig. 2-R-1** and **2**). A tissue-enriched gene may be considered not only as a very expressed gene in a particular tissue respect to others, but also as a constantly expressed gene (low or high) in a tissue while is variable and low expressed. For example, all genes located at sexual chromosomes are candidates to follow this kind of patterns. GTEx includes several male and female tissues, where those genes are intended to be expressed homogeneously but not necessarily at very high expression.

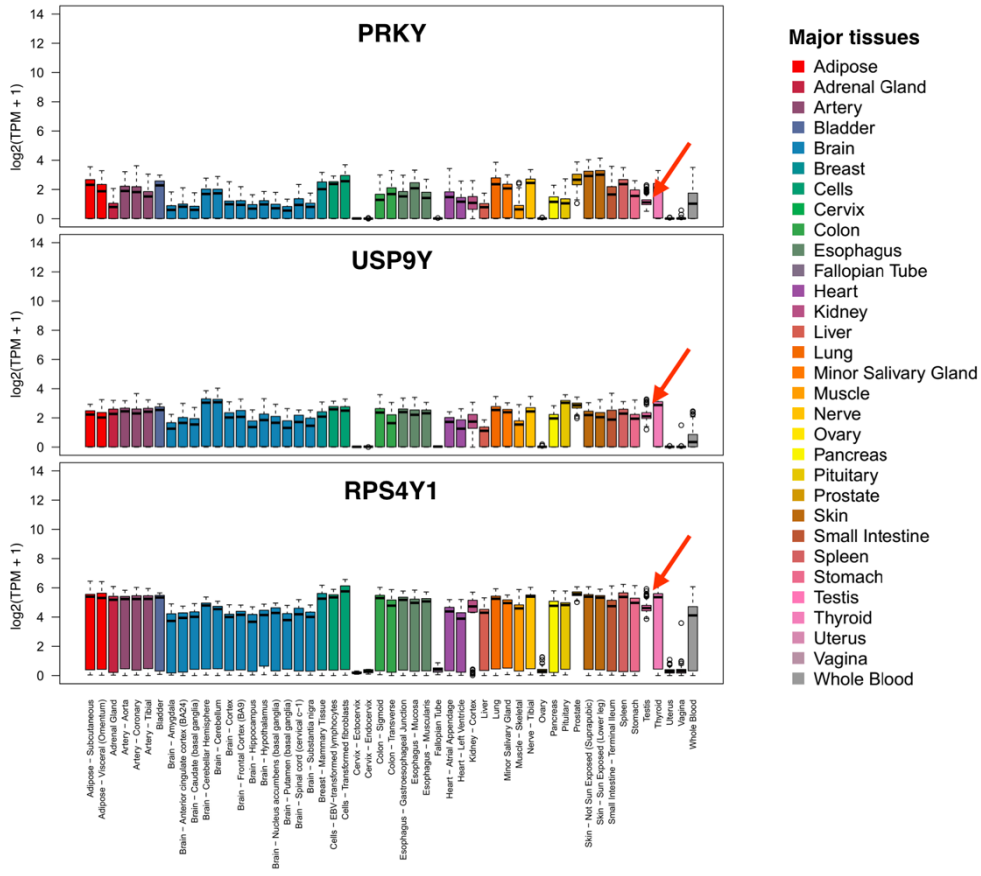


Figure 2-R-7. Tissue-enriched genes within very variable expression profiles. These three examples correspond to genes located at chromosome Y (PRKY, USP9Y, RPS4Y1), enriched in human male tissues like testis and prostate.

Particularly, we looked for stable tissue-enriched genes for male tissues located at chromosome Y or for female tissues located at chromosome X, which range of values are

overlapping other tissues and low, but non-zero. In the literature, several RNA molecules have been reported of causing a phenotypical difference even if only a single copy is detected within the cell (Seiler et al. 2017). Using cohesiveness statistic, we were able to find those patterns, where other methods failed. **Figure 2-R-7** shows three examples of genes located at chromosome Y and showed a very stable expression pattern in testis and prostate tissues, while very variable expression level is present at other human tissues.

DISCUSSION AND FUTURE WORK

Along this Chapter II, we explained and tested a new statistic as a method for feature selection in omic (or not) omic data. This statistic was intended as a simple and non-parametric measurement of the cohesiveness of categorical variables along a quantitative variable, called ***cohesiveness* statistic**. It could be categorized as a *filter* method for flat features (*univariate*, no evaluate a set features), since it does not contemplate the relationship among analysed features to score these features.

Cohesiveness statistic is based on the probability of finding a specific gap or distance among two elements within an ordered quantitative variable. One of the disadvantages of considering rankings instead original values is related to ties. Originally, cohesiveness integrates a *random* method to differentiate these values, but it may entail errors if the percentage of ties is high, because the feature would be considered as a random or negative feature. For that reason, further investigations about how we can approach this issue are required, for example, using a greater trimmed mean, reordering a subset of ties, removing all ties or labelling all those features with a high number of ties as non-discriminant.

However, the **advantages of using rankings or non-parametric measures instead parametric values** are greater than disadvantages:

- Cohesiveness could be applied indistinctly to any omic technology or dataset, it is not a platform-dependent method.
- Cohesiveness is not assuming any value distribution of omic data like classic methods for differential analysis (Section 1.3 of Introduction, Chapter I).
- Cohesiveness is not based on mean or median differences to perform a differential analysis, a minimum gap ($D^K = 1$) directly implies there is no other sample in this range of values.
- Cohesiveness is able to find stable patterns where there is no overlap among categories and sample median or mean may be similar.

As detailed in the Results section, we demonstrated how cohesiveness is able to

perform a differential analysis following a multiclass contrast design (**Fig. 2-R-3** and **4**), classifying the samples in similar success rates to more sophisticated methods, even similar to methods with high computational cost like Random Forest (RFE).

Particularly, we can focus on the results for the classification of samples in order to a better interpretation of how datasets are composed and which methods agree in their selection. The similarity of top and global signatures among all methods, as observed in **Figure 2-R-3B** and **Figure 2-R-5**, remarks a similar agreement with probabilistic and distance-related methods but a high specificity at top signatures, at least for Brain-1 dataset (**Table 2-M-1**). The partial agreement among other methods in terms of chosen signatures may reveal that the 6 major regions of Kang's dataset are mostly classifiable by several redundant genes. This result is supported by the fact that classification's rates slightly varies from lower to higher number of features (**Fig. 2-R-3A**). Alternatively, the classification's rates for Brain-2 and DLBCL improves from lower to higher number of top-features, revealing a less redundant scenario where more features are required to enhance the classification of several samples.

The methods to find tissue-enriched or condition-enriched are very common in the analysis of comprehensive omic datasets, where several tissues or cell types were included. Since the GTEx consortium produced the largest transcriptomic dataset of human tissues, we wanted to test the ability of cohesiveness to deal with a very large omic dataset (56202 features and 11688 samples) for two reasons: computational cost and large gap distributions. Cohesiveness is very fast in comparison with other precise feature selection methods like Random Forest (RFE, aprox. 400x slower) or ReliefF-derived algorithms and it is able to perform a complete analysis of GTEx (54 categories in 56202 features) in approx. 50 minutes. Additionally, the proposed trimmed version of cohesiveness (trim = 1%, by default) allows to improve its performance when the number of total samples is much greater than the number of samples of a category (**Fig. 2-R-6B**).

We can conclude that cohesiveness statistics represents a notably advance respect to current methods in two main issues: **(i)** it is a non-parametric approach, applicable to any omic data set without considering value distribution; **(ii)** and it is able to discover not only differential features associated with each category of samples, but also stable patterns, associated with a specific category of samples, within very variable features.

CHAPTER III

Integration of human protein-protein interaction networks and subcellular localization maps

BRIEF SUMMARY

In this Chapter III, we will develop and explore a new analytical framework for the integration of protein-protein interactomes and subcellular localization data of proteins. Classically, we may think the first requirement for a correct physical interaction should be the co-localization of these proteins in the same space. Nowadays, there is a lack of meeting between human interactome and subcellular localization data, which could be supporting our idea that transient protein-protein interactions and shuttling proteins would be inferred through the integration of both sources of data. Chapter III is based on **HuRI** (*Human Reference Interactome*) and **Cell Atlas** (*Human Protein Atlas* project) resources to gain insights into subcellular co-localization of interacting proteins.

All data, analyses and results provided along this Chapter III were produced in collaboration with Marc Vidal PhD's Laboratory at Center for Cancer Systems Biology (Dana-Farber Cancer Institute, Harvard Medical School. Boston, United States), during a short-term stay of three months in 2017 for collaborating in the last version of **HuRI** (*Human Reference Interactome*) project.

INTRODUCTION

Systems biology summarises our basic intuition about how proteins, genes, enzymes and every molecular compound interact with each other within a cell. Regarding the same genotype of all the cells composing an organism, systems biology tries to understand how different behaviours can be achieved with the same background but different environments. In this way, the evolution of molecular biology along the past century allows us to compile enough information to establish relationships, inferring new biological processes or properties (Vidal 2009). Understanding systems biology as a discipline to produce, integrate and analyse data of possible interactions within a cell, we may use different technologies to obtain data related to genomics, transcriptomics, metabolomics, epigenomics, interactomics, proteomics, etc. Consequently, our ability of analyse this massive or *big data* is conditioned by the proper advances in each technology platform and computational methods (Altaf-Ul-Amin et al. 2014). Developing new and advanced methods for the analysis, interpretation, and integration of several platforms has become one of the most complex challenges, even leading to the **multi-omics era** (Bersanelli et al. 2016, Huang et al. 2017).

As hinted along **Chapter I**, precision or personalized medicine is one of the most promising, controversial and complex challenges in biomedicine. It is based on systems biology in order to diagnosis and prognosticate patients, selecting the most relevant biomarkers to discriminate specific pathological subtypes from each other. Similar to other scientific fields, the progress of personalized medicine and systems biology is closely associated with cancer research because of the magnitude of related publications and research groups involved. This process would generate predictions of biomarkers from modelling, which should be confirmed by experimental evaluation (Werner et al. 2014). Vice versa, advances in systems biology could be directly applied on cancer research or personalized medicine (Sevimoglu and Arga 2014).

Considering that systems biology is dramatically advancing our mechanistic understanding in cancer research, the protein-protein interaction networks allow wider

interpretation of how each protein function depends on other protein activities within the global cell context under a specific functional organization (Barabasi and Oltvai 2004). Protein-protein interaction is defined as a real and physical interaction among proteins though electrostatic forces, which need to be confirmed by experimental techniques. This molecular interaction occurs in a specific biomolecular context of a given cell or organism (De Las Rivas and Fontanillo 2010).

Protein-protein interactions (PPI) have been approaches from very different fields like biochemistry, molecular dynamics or signal transduction. The compendium of whole protein-protein interactions occurring within a cell is usually called *protein interactome*, a very useful resource for signal cascades' research and discovery of therapeutic candidates in the pharmaceutical industry (Sevimoglu and Arga 2014). These networks have been broadly used to describe molecular background of a wide variety of disease and to determine common pathways to similar pathologies (Barabasi et al. 2011, Menche et al. 2015).

1. Technologies to infer protein-protein interactions

Nowadays there are a wide variety of techniques which pursue the identification of protein-protein interactions, several of them resembling slight modifications of an original technique. The chemical and physical characteristics of any of these techniques strongly condition the nature of each protein-protein interaction and, consequently, the confidence behind the experimental validation. While several techniques are based on biophysical methods (electron scattering, biosensor, luminescence, x-ray crystallography, etc.), other methods are based on biochemical properties, imaging technologies, genetic inference or protein complementation, for example (Berggard et al. 2007). Interestingly, the **Ontology Lookup Service** (OLS) from EMBL-EBI provides a useful categorization of these techniques (MI: molecular interactions) compiled in a huge ontology of MI methods (https://www.ebi.ac.uk/ols/ontologies/mi), offering controlled vocabularies to describe different aspects which need to be considered (Jupp et al. 2015).

The main classification of technologies for detection of PPIs is based on experimental characteristics: *in vitro*, *in vivo* or *in silico*. Any of these categories shows a wide variety of approaches within them, providing very distant methods for detecting and,

posteriorly, validating any protein-protein interaction (Rao et al. 2014). Within all them, it is important to highlight mass spectrometry derived methods, modifications of yeast-two-hybrid technique or phage display as widely used high-throughput technologies. Although high false positive rates and noise have been reported from these techniques, they are still considered unbiased techniques which provides a meaningful read of PPIs within a cell. Indeed, several ongoing projects pursue the construction of high-quality human (and other organisms) protein interactomes, using high-throughput techniques, like HuRI, BioPlex, QUBIC or CoFrac (Rolland et al. 2014, Hein et al. 2015, C Wan et al. 2015, Huttlin et al. 2017).

We can also distinguish the experimental methods for detecting protein-protein interactions into two categories: **(i) *binary***, which would be any method able to specifically detect interactions between two proteins only; and **(ii) *co-complex***, which would be any technology with the ability to measure direct or indirect interactions among two or more proteins (complexes) (Yu et al. 2008). Thus, the main criticism on *co-complex* methods resides in their inability to distinguish between physical or indirect interactions between two proteins, leading to the development of computational approaches (spoke or matrix models) to infer real interactions within a complex (Hakes et al. 2007). Among the *binary* methods, the most used technology is Yeast Two Hybrid (Y2H), which is based on the bait-prey (physical interaction) principle to promote the expression of a reporter gene. Mass spectrometry, affinity purification and co-immunoprecipitation are the most relevant *co-complex* methods.

We also may discriminate between **permanent** and **transient protein-protein interactions** within a specific biological context using any of these technologies (La et al. 2013). For example, macromolecular complexes are composed of several proteins which physically interact each other through permanent interactions, while most of cellular processes, such as signalling pathways, phosphorylation, ubiquitination or gene repression involve transient protein-protein interactions. Consequently, since transient interactions are weaker depending on cellular context (spatial and temporal conditions) (**Fig. 3-R-1**), they are more difficult to detect using high-throughput or even more specific low-throughput technologies.

During the last decades, much specific and diverse efforts (low and high-throughput experiments) have been carried out to detect both transient and permanent

PPIs, single or multiple. Thus, several databases, often called *primary* databases, were developed to compile and integrate all this information in unique frameworks to facilitate new analyses and validations. These primary databases, for example HPRD, DIP, BioGrid, BIND, IntAct or MINT (Xenarios et al. 2000, Bader et al. 2001, Zanzoni et al. 2002, Hermjakob et al. 2004, Stark et al. 2006, Keshava Prasad et al. 2009), display and curate all reported PPIs directly from the literature. Unfortunately, it was well-demonstrated that the overlap among them are relative small. For this reason, other recent efforts such as APID (Agile Protein Interaction Data Analyzer), developed in our laboratory, try to integrate all these primary databases, as a second effort to produce full protein-protein interactomes involving every reported PPI (Prieto and De Las Rivas 2006, Alonso-Lopez et al. 2016).

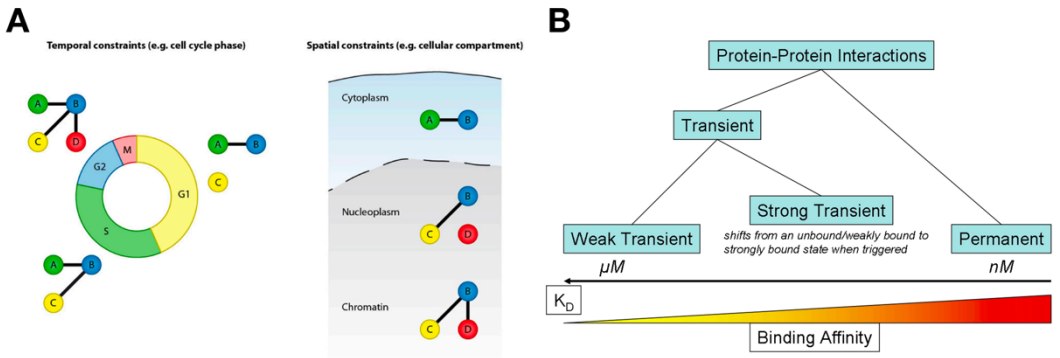


Figure 3-R-1. Transient and permanent protein-protein interactions. Transient interactions are weaker **(B)** and depend on temporal and spatial cellular patterns **(A)**.

2. Protein interactomes

Cells could be represented as complex networks of molecular interactions, providing a meaningful relational space to analyse, validate and infer specific functions (Vidal et al. 2011). Given that the protein is the main functional molecule of any organism, a protein interactome is composed of all reported protein-protein interactions from a specific biological context, cellular type, tissue or organism. Within these networks, nodes are represented by proteins while edges are non-directed because the directionality of these interactions cannot be defined. Importantly, datasets generated through *binary* methods, like Y2H, would contain interactions between two proteins, while *co-complex* methods would bring a mix between direct and indirect associations. Not only these two types of networks differ in terms of global properties, but also may reflect different relationships

among proteins given the same cellular context (Seebacher and Gavin 2011).

Systems biology has been an emerging field for last decades, when graph theory was slowly introduced as theoretical framework to analyse and scale reported molecular interactions behind a particular cellular behaviour (Vidal 2009). Different aspects of topological analyses of large protein interactomes have been very discussed in the literature, where specialists in the field proposed multiple new concepts and ideas. For example, local perturbations within a biological network have been linked to several pathologies (Menche et al. 2015), variants of cancer (Yi et al. 2017), protein targets (Noh et al. 2018) or particular phenotypes (Peng et al. 2018). Local perturbations may be understood as a gene deletion, specific mutation, protein inhibition or any action which produces an alteration in the functionality in one or a group of nodes (i.e. proteins, mRNA, metabolites, etc) within a network (**Fig 3-I-2**). Specially, DNA variants were frequently associated (Sewell and Fuxman Bass 2017).

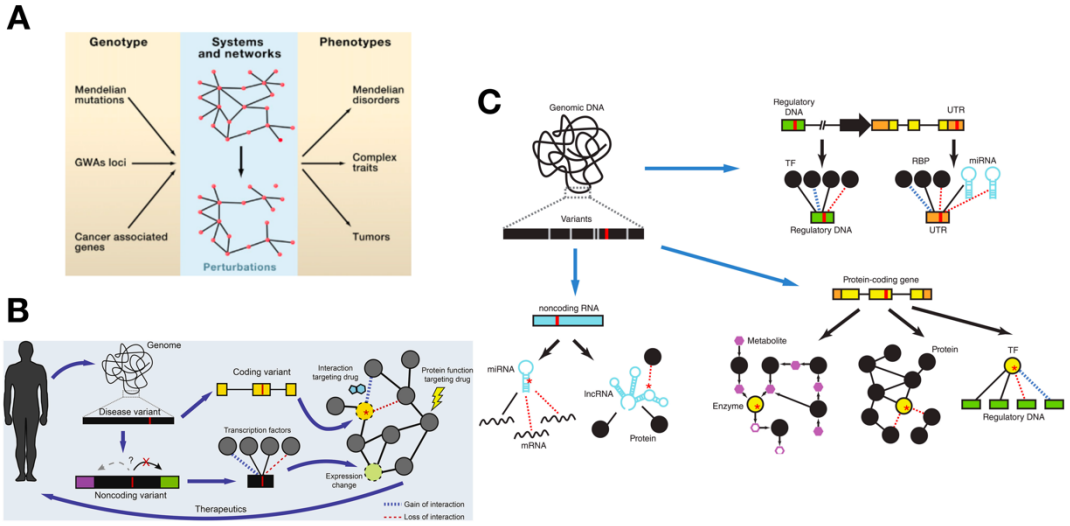


Figure 3-R-2. Network perturbations have been widely associated with different diseases. Figures adapted (Vidal 2009, Sewell and Fuxman Bass 2017).

Alternatively, there is an increasing tendency of integrating networks from different omic sources (i.e. transcriptomics, metabolomics and proteomics), called *multi-omics*, to enhance the ability of systems biology approaches to define particular states underlying any cellular state (List et al. 2016, Dimitrakopoulos et al. 2018). As such, an incomplete biological network would be fulfilled by other omic approaches, if this integration is properly conducted, considering both data properties and prior knowledge.

3. Integrative analyses of subcellular localization and protein-protein interaction datasets

Protein-protein interactomes have been broadly used for predicting subcellular localization (Shin et al. 2009) based on following principle: any physical molecular interaction really requires both molecules to be located at the same cellular space.

Nowadays, subcellular localization information could be classified as experimental (based on antibody imaging) or computational (based on computational prediction from similar located proteins). Although experimental datasets were intended to be systematic and unbiased, a slight nucleus bias was reported for Cell Atlas dataset by authors (Thul et al. 2017). On the other hand, since Y2H methodologies have been largely developed and improved during last decades, protein-protein interactomes derived from this technique may be considered as reliable enough, systematic and unbiased biological networks. Y2H provides a clearer interpretation of the interactome because it is a *binary* method. Thus, combining experimental and systematic approaches from both platforms may be a more unbiased approach for enhancing the prediction of unallocated proteins.

Several approaches related to the proposed integrative framework for human interactomes have been recently published. A simple search in PubMed of "**protein localization prediction**" returns almost 90 different methods proposed since the early 2000's. Several of these proposals are focused on experimental approaches and data (Shen and Burger 2010, Bogachev et al. 2016, Liu and Hu 2016) or annotation databases (Chi and Nam 2012). For example, ComPPI resembles a database of protein-protein interactions compartmentalized in very specific subcellular localizations (about 1600 different compartments) for a variety of living organisms (Veres et al. 2015). Alternatively, a wide variety of computational methods based on network analysis (Mooney et al. 2011, Xu et al. 2013), machine learning (like Supper Vector Machines – SVMs) (Rahman et al. 2016, Almagro Armenteros et al. 2017, Hasan et al. 2017), logistic regression models (S Wan et al. 2015) or entropy measures (Zhao et al. 2015) were also proposed.

In summary, our proposal aims the integration of protein interactomes with subcellular localization data as it would help to predict transient protein-protein interactions (spatial and temporal conditioned), which may remain *undetectable* for antibody detection.

MATERIAL AND METHODS

1. Molecular interactions methods (PSI-MI) and ontologies

Throughout this Chapter III, all PSI-MI terms and controlled vocabulary for molecular interactions methods were based on Ontology Lookup Service (OLS) from EMBL-EBI website (https://www.ebi.ac.uk/ols/ontologies/mi). Due to our experience with literature-based interactomes (Prieto and De Las Rivas 2006, Alonso-Lopez et al. 2016), a manual curation of PSI-MI methods was conducted to generate a robust literature-based human interactome in agreement with Marc Vidal's laboratory at the Center for Cancer Systems Biology (Dana-Farber Cancer Institute, Harvard Medical School. Boston, USA). This new categorisation will be used to produce one of the human interactomes here analysed (LitBM-17) and future analyses for APID interactomes. A table with all PSI-MI methods and categorisation into *binary*, *indirect* or *invalid*, is provided in **Appendix 3**.

2. Human interactome datasets

All human interactome datasets used for the comparison, analysis or integration with Cell Atlas along the Chapter III of this dissertation are described below:

- **HI-III:** human interactome produced as a result of the third mapping of Human Reference Interactome (HuRI) project to provide high-quality maps of protein-protein interactions systematically obtained by Yeast Two-Hybrid experiments (Rual et al. 2005, Rolland et al. 2014), covering around 77% of human genome search space. All proteins have been successfully mapped to ENSEMBL IDs provided by GENCODE database. It is composed of 8189 proteins and 49839 interactions. This last version (HI-III) is currently unpublished, but all dataset is public and available for downloading (http://interactome.baderlab.org/).
- **BioPlex:** human interactome produced using affinity purification-mass spectrometry methodology. This approach is intended as a co-complex approach to find

interactions or associations occurring in macromolecular protein complexes. It covers more than 25% of protein-coding genes from the human genome, resembling a large protein-protein interaction network composed of 10571 proteins and 53074 interactions (Huttlin et al. 2017).

- **QUBIC:** human interactome produced using the quantitative BAC-GFP interactomics technique, called QUBIC (Hubner et al. 2010). This interactome could be also considered as a co-complex approach to detect protein co-complexes. It is composed of 5516 proteins and 29574 interactions (Hein et al. 2015).
- **CoFrac:** human interactome obtained by co-fraction methods, embedded into co-complex approaches to disclose macromolecular complexes. It is composed of 3429 proteins and 16487 interactions (C Wan et al. 2015).
- **LitBM-17:** literature-based interactome created from *mentha* resource, after improving the classification of PSI-MI terms referred in Section 1 of Material and Methods. It was specifically created for the next publication involving HI-III interactome at CCSB Systems biology laboratory of Marc Vidal, PhD. The *mentha* resource involves five different primary databases of protein-protein interactions: MINT, IntAct, DIP, BioGRID and MatrixDB. First, data were filtered to have valid identifiers (ENSEMBL IDs from GENCODE, Pubmed IDs and PSI-MI terms). A single evidence (PPI) consisted of a Pubmed ID and interaction detection method (MI) included in the PSI-MI controlled vocabulary, while duplicated entries from different databases were merged. LitBM-17 only considered PPIs supported by more than two publications or PSI-MI methods classified as *binary* after removing all records included in human experimental interactomes described above. It is composed of 6047 proteins and 13441 interactions. A similar approach to produce a literature-based interactome was done for the previous version of HuRI (HI-II) (Rolland et al. 2014).

3. Subcellular localization data: *Cell Atlas* from *Human Protein Atlas* project

The subcellular localization data used is a part of Human Protein Atlas project (initiated in 2003), which resembles a comprehensive effort for elucidating the whole map

of human proteins along cells, tissues and organs. The researchers used antibody-based imaging for this particular subcellular map of protein localization, called Cell Atlas (Thul et al. 2017), while other omic technologies like mass spectrometry, systems biology or transcriptomics were used for other purposes (Uhlen et al. 2015, Uhlen et al. 2017).

The subcellular protein map compiled by Cell Atlas is composed of 12003 proteins (mapped to ENSEMBL IDs) located at 32 different subcellular localizations, manually grouped by similarity into subcellular meta-compartments (**Fig. 3-R-4B**). Additionally, there are four levels of reliability in Cell Atlas: *Validated*, *Supported*, *Approved* and *Uncertain*. Any protein annotated only as *Uncertain* to one or more subcellular localizations was not considered for the analysis and integration.

The whole dataset can be downloaded from the Human Protein Atlas website as a text file (https://www.proteinatlas.org/about/download).

4. Statistical analyses

The **Fisher's exact test** was used as enrichment tool to determine if significant overlaps are present when 2x2 contingency tables have to be assessed, also represented as Venn's diagrams. Odds ratios are provided in log scale and p-values associated will be corrected using False Discovery Rate (Benjamini and Hochberg 1995), if it is necessary.

Additionally, **Z-score transformation** (Eq. 3.1) after network randomization will be used to determine whether a real or experimental statistic is significantly different to random distribution, using the same data.

$$Z = \frac{x - \mu_{random}}{\sigma_{random}} \qquad \text{(Eq. 3.1)}$$

where x corresponds to experimental/real observation, μ_{random} corresponds to the average of random observations and σ_{random} corresponds to standard deviation of random observations. Empirical p-values from normal distribution will be calculated (right or two-tailed) and False Discovery Rate applied if required (Benjamini and Hochberg 1995).

Regarding **correlation analyses**, the Spearman Correlation Coefficient (SCC) was calculated when non-parametric analysis of relationships was needed, while the Pearson Correlation Coefficient (PCC) was used in parametric relationships. The Weighted Pearson Correlation Coefficient (WPCC) was also calculated for particular scenarios where weights

are relevant to properly understand the results. The significance (p-values) for all these correlation metrics were calculated approaching correlation metrics to Student's t-distribution (degrees of freedom: n – 2), using the native R package called *stats*.

5. Network randomization

The network randomisation of HI-III interactome (undirected network) was conducted using *BiRewire* R package (Iorio et al. 2016). This procedure was iteratively run to generate 1000 random degree-preserved networks, which were used as basis to produce null distributions of different statistics of HI-III interactome.

6. Network analysis, integration and visualization

All statistical analyses, integration and visualisation of human interactomes and Cell Atlas were conducted in R environment. Several R packages like *igraph*, *reshape2* or *ggplot2* were specifically used for the visualisation and management of human interactomes in R. Cytoscape software was also used to represent biological subnetworks (Shannon et al. 2003).

RESULTS AND DISCUSSION

Here we will present the current results and analyses obtained from this ongoing collaboration with Marc's Vidal laboratory from CCSB (Dana-Farber Cancer Institute). For this reason, we unified all results and corresponding discussion of this Cell Atlas and HI-III integration, for a better understanding of currently done work and main gaps to be fulfilled. Moreover, we briefly discuss about the agreement reached for the manual categorization/curation of PSI-MI methods before the creation of a reliable human literature-based interactome (LitBM-17) to compare against HI-III from HuRI (Human Reference Interactome) project.

1. Categorization of PSI-MI terms to produce a reliable literature-based interactome

Although systematic protein-protein interactomes could be produced through several methodologies, we would like to focus on one of the most common and well-understood techniques to infer physical protein-protein interactions: experimental assays like the Yeast Two-Hybrid method (Y2H) (Fields and Song 1989, Bruckner et al. 2009), which is unbiased to the current knowledge and reports physical interaction between two proteins (no macromolecular complexes).

Our lab has been interested, during the last decade, in developing reliable protein-protein interaction interactomes for different organisms based on available literature information. In this way, our last published version of APID (Alonso-Lopez et al. 2016) summarizes and curates information from other databases exclusively related to proven protein-protein physical interaction single experiments, providing a simple and complete database including different confidence intervals based on previous knowledge. Previous studies about APID and literature interactomes led us to collaborate with the ongoing HuRI (Human Reference Interactome) project, in order to improve and curate the classification of PSI-MI terms into *binary*, *indirect* or *invalid* categories.

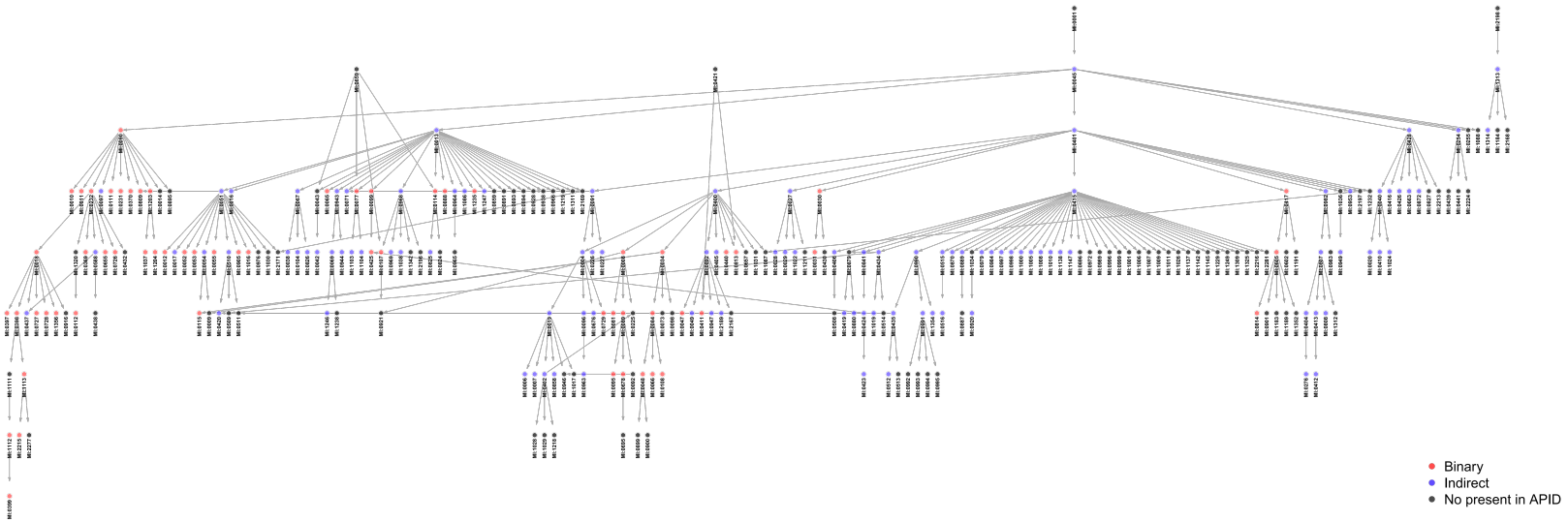


Figure 3-R-1. PSI-MI methods present in APID and also in LitBM-17, specifically created for the comparison with human Y2H interactome: HI-III (HuRI project). Binary and indirect methods are differentiated by colours.

For this reason, we aimed to categorize frequent PSI-MI terms to facilitate their posterior use for producing *binary* and/or *indirect* interactomes, adding posterior thresholds of confidence related to the number of publications or different methods supporting each PPI. Particularly, this new classification was applied for producing the literature-based interactome (LitBM-17) used to compare with the last version of human interactome of HuRI project (HI-III). Since Y2H is considered a *binary* method, we were interested to produce a literature-based interactome only including PPIs supported by *binary* method more than one time (one publication and two methods; or two publications using the same PSI-MI method), following previous scheme of comparisons of last version of HuRI (Rolland et al. 2014). Finally, the new proposed classification of PSI-MI terms is shown in **Appendix 3**.

Interestingly, our agreement about *binary*, *indirect* and *invalid* terms led us to conclude that current ontologies for PSI-MI terms may be confusing and imprecise. **Figure 3-R-1** took ontology of PSI-MI terms from OLS and mapped all terms appearing at least one time in APID, colouring them according to table in **Appendix 3**. We can observe how several father terms are considered *binary* while not all children are or several relationships among very distant areas of this tree/ontology representation of PSI-MI methods. These characteristics of OLS ontology lead to problems when term-convergence was carried out. In our proposal, similar PSI-MI methods were collapsed along the ontology to avoid mismatch information compiled from different sources. Indeed, since literature publications are individually curated by each primary database, we analysed the agreement among these databases to assign PPIs to PSI-MI methods.

Figure 3-R-2 shows two networks displaying the OLS ontology for PSI-MI methods where information about the number of PPIs reported in BioGrid and IntAct were mapped. As we can see, BioGRID is carrying out a more astringent classification of publications reporting PPIs in PSI-MI methods, occupying only a small portion of this OLS ontology. On the other hand, IntAct database assigns experiments to more specific PSI-MI methods and differently converges at some areas of this OLS ontology. This issue is crucial for meta-databases like *mentha* or APID when any threshold or curation is desired because same records or experiments may appear duplicated depending on which databases are mapped, leading to false entries at different confidence levels.

As detailed above in Material and Methods, the literature binary generated for the

comparison against HuRI (called LitBM-17) was produced only considering PPIs supported by more than two publications or PSI-MI methods classified as *binary*. To avoid these issues, all methods from different sources were collapsed following a meta-group scheme (also shown in **Appendix 3**) instead a father-child relationship derived from OLS ontology. Thus, following a meta-group collapse for PSI-MI assignments should deal with these redundancies and disagreements among large primary protein-protein interaction databases as BioGRID, IntAct, HPRD or MINT.

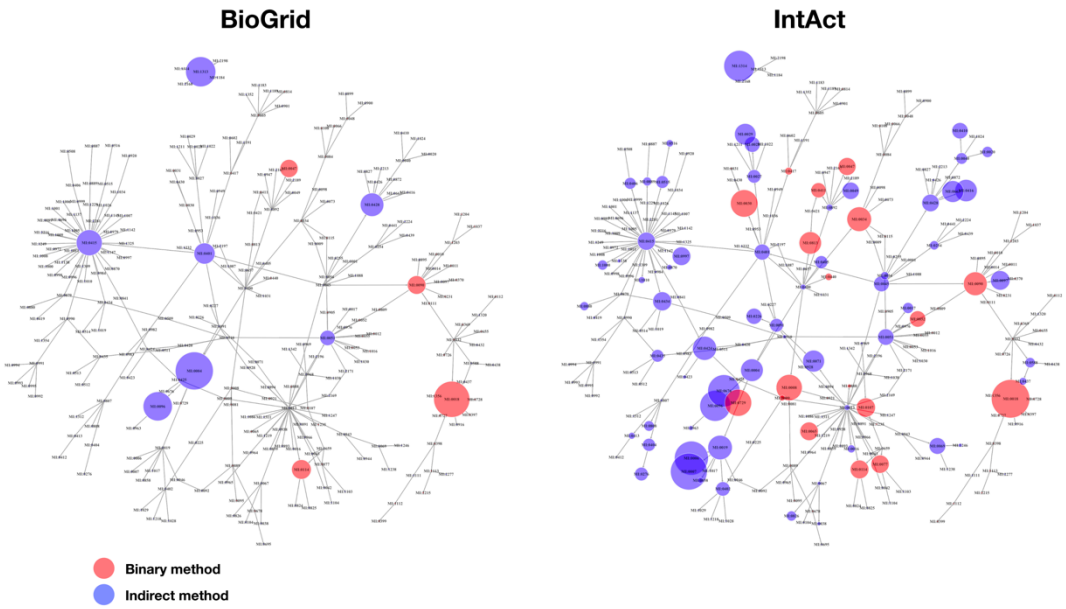


Figure 3-R-2. BioGRID and IntAct number of PPIs mapped on OLS ontology of PSI-MI terms, represented as a network of terms. Size of nodes corresponds to number of PPIs assigned to this PSI-MI method.

2. Comparison of human interactomes

We will analyse and contrast several human interactomes against experimental subcellular localization data of Cell Atlas. Our main objective behind these comparisons is to define an unbiased and high-throughput human interactome for the integration with Cell Atlas information.

First of all, we calculated the overlap between different human interactomes. Although they are not directly comparable because of the technologies used to produce the dataset, **figure 3-R-3** provides us with a brief overview of how distant are these

interactomes in terms of protein-protein interactions and proteins detected.

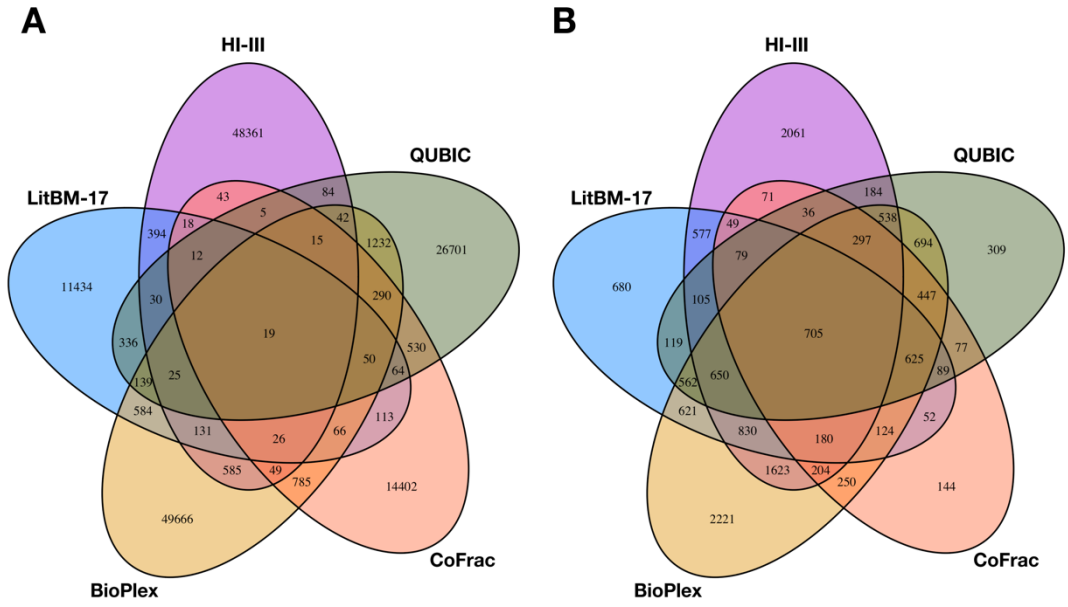


Figure 3-R-3. Overlap of different human interactomes used: **(A)** protein-protein interactions and **(B)** proteins.

As we can observe in **Figure 3-R-3**, there is a very low overlap among different human interactomes for both proteins and protein-protein interactions, as previously reported for old versions of these interactomes (Rolland et al. 2014). Technologies behind these interactomes greatly vary the ability of detecting protein-protein interactions (Y2H, co-complex or literature-based), often making difficult the direct comparison of these networks. However, interactomes systematically generated were reported to be more reliable than literature-derived (especially if low-throughput experiments are more frequent), which involves a strong study bias (Luck et al. 2017).

3. Coverage of the integrative analysis and biases for subcellular compartments

Any technology for detecting molecular interactions is able of showing particular biases depending on their limitations and suitability for the analysis. Since the human interactome of interest (HI-III) was produced through Yeast-Two-Hybrid technology (Y2H), it is important to remind that Y2H is a technique based on bait-prey contact inducing

transcription of a reporter gene. Thus, any revealed protein-protein interaction should be reproducible in the cellular nucleus of yeast, even if these proteins are not usually related or present at this subcellular localization. Indeed, not only the interaction should be reproducible but also bait-prey proteins must be able to enter the nucleus to activate transcription of reporter (Bruckner et al. 2009). For these reasons, some authors reported this specific condition in the literature (von Mering et al. 2002, Bjorklund et al. 2008).

For this reason, we investigated the coverage of these two high-throughput techniques (fluorescent imaging and yeast-two-hybrid) in terms of proteins detected and subcellular localizations assigned to each protein. Since subcellular localizations greatly vary in terms of size and number of functional proteins, we expected a variable number of detected proteins assigned to each subcellular compartment. Additionally, we already mentioned that several authors described a relevant propensity for detecting nucleus-located proteins in yeast-two-hybrid experiments. Thus, similar preference might be expected for HI-III interactome.

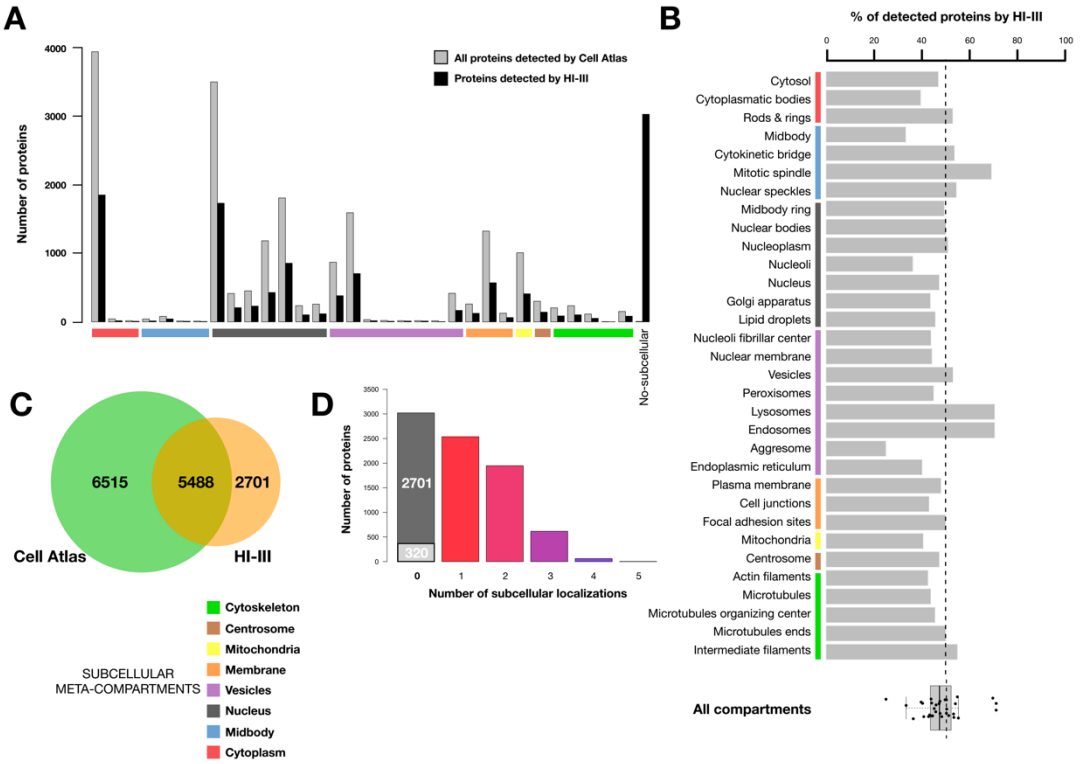


Figure 3-R-4. Coverage analysis of Cell Atlas and HI-III interactome datasets. **(A)** Number of proteins of Cell Atlas and HI-III per subcellular compartment. Labels on *x-axis* are ordered like panel B. **(B)** Percentage of detected proteins of HI-III respect to total number of proteins of Cell Atlas per subcellular localization (average = 47.89%). **(C)** Overlap of detected proteins by both datasets. **(D)** Number of proteins per number of subcellular localizations that may occupy any protein.

Notably, it is important to remember that all protein IDs were properly mapped to ENSEMBL IDs from GENCODE release 27. Furthermore, any subcellular assignments from Cell Atlas were only considered if status or reliability level is *Approved*, *Supported* or *Validated*, but not *Uncertain* (Thul et al. 2017).

3.1 Overlap of Cell Atlas and HI-III human interactome

Figure 3-R-4 tries to summarize main aspects of coverage between Cell Atlas (12003 proteins) and HI-III interactome (8189 proteins). We can observe a notable overlap of number of proteins included in HI-III and with known subcellular localization in Cell Atlas (5488 proteins, 67.01% of HI-III proteins), shown in **Fig. 3-R-4C**. However, 320 out of these 5488 proteins are *Uncertain* assigned, then their subcellular localizations were not taken into consideration (**Fig. 3-R-4D**). Additionally, HI-III's authors reported a high number of interactions depending on keratins and keratin associated proteins (around 100 proteins resembling almost 10000 interactions). The localization of most of these proteins is not known, so we removed them from this integrative analysis to eliminate possible biases. Finally, 5121 out of 5488 proteins entered in posterior analyses, providing a full HI-III interactome comprising 18024 interactions.

Regarding **biases for particular subcellular locations**, we can observe HI-III shows no preferences for any compartment. Panel A of **figure 3-R-4** shows the number of proteins detected per subcellular localization. Noteworthy, several subcellular localizations have a very low number of associated proteins, especially for localizations comprising the whole *Midbody* meta-compartment. Alternatively, the *cytosol* or *nucleoplasm* are the most represented spaces, as expected. However, the percentages of proteins present in HI-III respect to Cell Atlas per subcellular compartment are close to 50% (mean = 47.89%), corresponding more variable percentages to those subcellular localizations with low number of proteins (i.e. *aggresome*, *lysosomes*, *mitotic spindle*, etc). Due to these similar percentages, we can consider that there are no significant preferences for a subcellular compartment in HI-III in terms of detection of proteins, especially relevant for all those nucleus-related compartments which have been previously described as possible favourite subcellular spaces for Y2H experiments.

In conclusion, these results of coverage and biases related to protein's detection

allowed us to continue with all analyses and integration, proving that exists a remarkable agreement between these two resources: Cell Atlas and HI-III interactome.

3.2 Comparison of subcellular biases among human interactomes

Once we have defined the coverage between Cell Atlas and HI-III, we deepened further to estimate if significant biases are present in the human interactome. We demonstrated that percentage of detected proteins by HI-III is very similar along different subcellular compartments (**Fig. 3-R-4B**), but the probability of founding these percentages by chance can be estimated according to the subcellular compartment size and whole interactome size. In this way, we also compared all human interactomes described in Material and methods to disclose biases for all similar human interactomes and compared them.

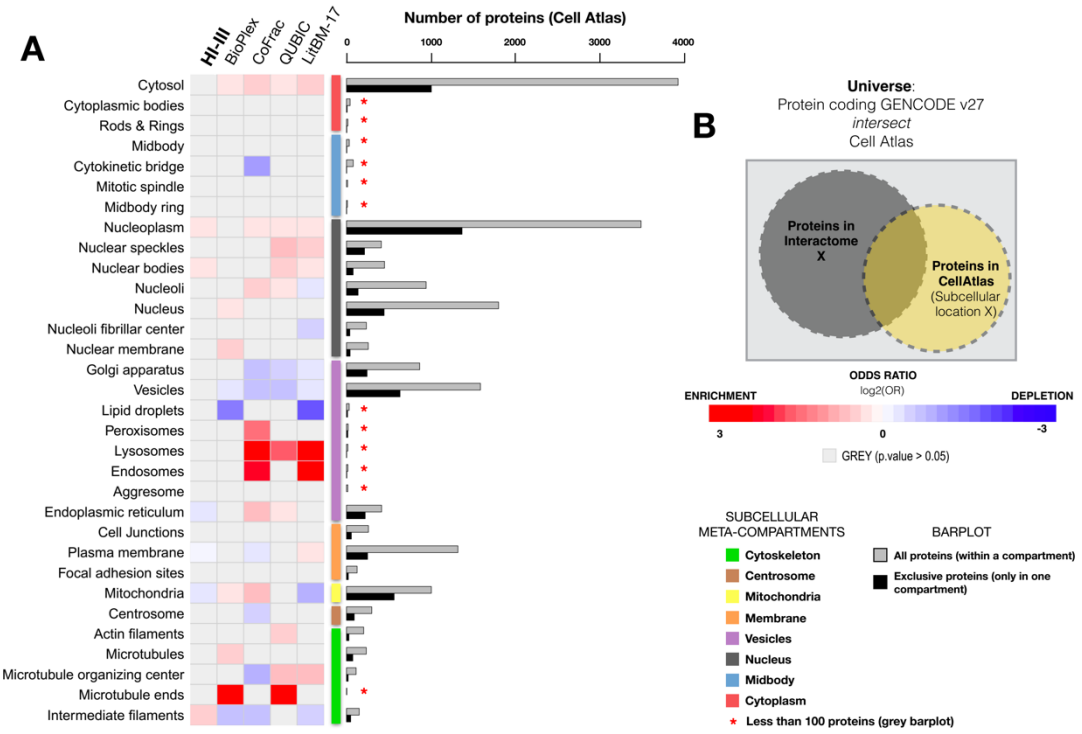


Figure 3-R-5. Enrichment analysis of every human interactome compared (HI-III, BioPlex, CoFrac, QUBIC and LitBM-17). **(A)** Heatmap showing enrichment (odds ratio) for subcellular localizations where red squares correspond to higher odds ratios and blue square to lower odds ratios. Grey squares mean no significant odds ratios. **(B)** Scheme of hypergeometric test calculated, where the intersection of list of protein coding genes from *Genecode v27* and genes detected by *Cell Atlas* were used.

Based on scheme displayed at **Figure 3-R-5B**, we calculated the hypergeometric test for contingency tables (also known as **Fisher's exact test**) on every single human interactome per subcellular localization of Cell Atlas. In this way, the odds ratio provided were transformed into log2 scale. Thus, **figure 3-R-5A** summarizes both enrichment analysis (biases) and number of proteins per subcellular localization per human interactome. We accompanied this heatmap with a barplot displaying number of total proteins and number of exclusive proteins (only in one compartment) per subcellular localization in Cell Atlas. Furthermore, we highlighted in different colours any related subcellular localization according to subcellular meta-compartments.

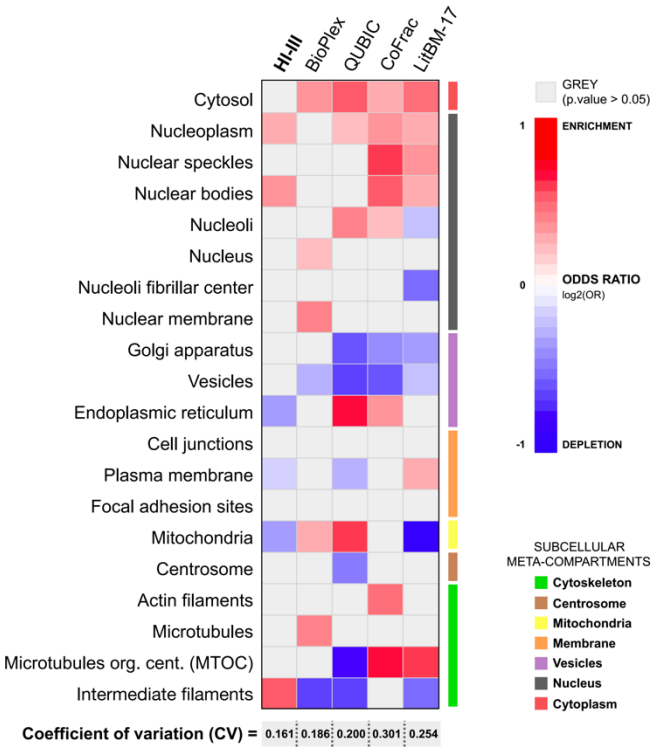


Figure 3-R-6. Enrichment analysis of human interactomes per subcellular localization of Cell Atlas involving more than 100 proteins (20 subcellular localizations). The coefficient of variation of log2(odds ratio) per interactome is shown, highlighting HI-III as the less biased human interactome.

Attending to Fisher's exact test results (**Fig. 3-R-5A**, heatmap), we can observe the different preferences for subcellular compartments of each human interactome (HI-III, BioPlex, CoFrac, QUBIC and LitBM-2017). Similar to previous results, the most relevant enrichment or depletion (odds ratios) correspond to all those subcellular compartments involving less than 100 proteins. For example, CoFrac, QUBIC and LitBM-17 are highly

enriched for *lysosomes* and *endosomes* compartments. Generally, we can also observe how HI-III are less biased than any other human interactome compared, showing a slightly enrichment for *nucleoplasm* and *nuclear bodies* spaces and a slightly depletion for *endoplasmic reticulum* or *mitochondria* spaces. Both results were expected because a lack of membrane-related proteins have been also described for Y2H experiments in the literature, apart from the preference for nucleus-localized proteins.

Regarding the total number of proteins per subcellular compartment (**Fig. 3-R-5A**, barplot), great differences among various spaces led us to remove all those showing less than 100 proteins not exclusively located (grey bar). Any subcellular localization showing a low number of proteins is more sensible to dispersion and suitable of being enriched or depleted. Thus, we recalculated this same enrichment analysis after removing those compartments with less than 100 proteins, showing a similar trend. Here, we also computed coefficient of variation on odds ratios per interactome (**Fig. 3-R-6**).

Interestingly, we can also note how several spaces are more suitable for multi-localized proteins, which were detected at more than one compartment, than others (**Fig. 3-R-5A**, barplot). For example, the *mitochondria* show lower number of total proteins than *plasma membrane* but almost double number of exclusive located proteins, indicating that proteins located in *mitochondria* are more dependent on this space. *Plasma membrane* is intended as one of the most transversal compartments, where occurring an outstanding variety of biological functions carried out by wide variety of proteins (O'Connor et al. 2010).

Subsequently, we can conclude HI-III interactome has notable coverage of subcellular map proposed by Cell Atlas, showing no remarkable biases for subcellular localization even for nucleus-related spaces. Indeed, HI-III is less biased than current human interactomes derived from other high-throughput technologies for defining molecular interactions (BioPlex, CoFrac or QUBIC) and also less than literature-based interactomes (LitBM-17) which comprise both high and low-throughputs techniques.

4. HI-III tends to connect proteins between more related subcellular compartments.

As mentioned above, the classical understanding of physical protein-protein interaction unmistakably implies the co-localization of proteins involved: it is necessary that

both proteins are able to interact in the same biological context. Indeed, several of the current technologies to disclose molecular interactions are just based on co-localization events, like *Fluorescence In Situ Hybridization* (FISH), *Förster Resonance Energy Transfer* (FRET) or *Surface Plasmon Resonance* (SPR) (Dunn et al. 2011).

Particularly for protein-protein interactomes, Thul et al. proposed in the original paper describing Cell Atlas that this resource could be used to improve any existing interactome (Thul et al. 2017), providing an extra level of reliability for those protein-protein interactions occurring in the same subcellular compartment. Otherwise, any protein-protein interaction is less likely to occur. Nevertheless, the quality of co-localization techniques is conditioned by several factors related to biochemical characteristics of proteins or subcellular compartments belonging them. Many transient or very specific molecular interactions may be missed, leading to an incomplete knowledge of the real subcellular map. Consequently, we aimed to test whether **the integration of a complete cell protein-protein interactome and subcellular localization information would empower our ability for discovering transient protein-protein interactions and shuttling proteins**, which are able to multiple interact with multiple proteins in different subcellular spaces.

4.1 Enrichment analysis for shuttling proteins among compartments

Aiming for this purpose, we tested the ability of HI-III to rescue protein-protein interactions within the same compartment or within very related compartments. As a measurement of the relationship between two subcellular compartments, we first calculated the significance for enrichment or depletion on **multi-localized or shared proteins**: proteins demonstrated to be in two or more compartments. Two subcellular compartments would be related if there is a significant enrichment in shared proteins. After mapping subcellular localization to proteins in HI-III (**Fig. 3-R-5**), Fisher's exact test between pairs of compartments were calculated.

Figure 3-R-7 resembles a heatmap which shows the significant relationships (overlap in number of proteins) among all possible pairs, cut-off at two different thresholds after or before FDR correction of p-values from Fisher's exact test. Raw p-values are also shown because several authors reported that multiple p-value adjustment on asymmetrical p-value distributions are limited to two-tailed assumptions (Pounds and Cheng 2006). In

fact, right-tailed test was chosen because we cannot assume that depletion of overlap between subcellular compartments is related to this technique (Y2H) or it is biologically relevant, even if any large bias was found in HI-III (**Fig. 3-R-6**).

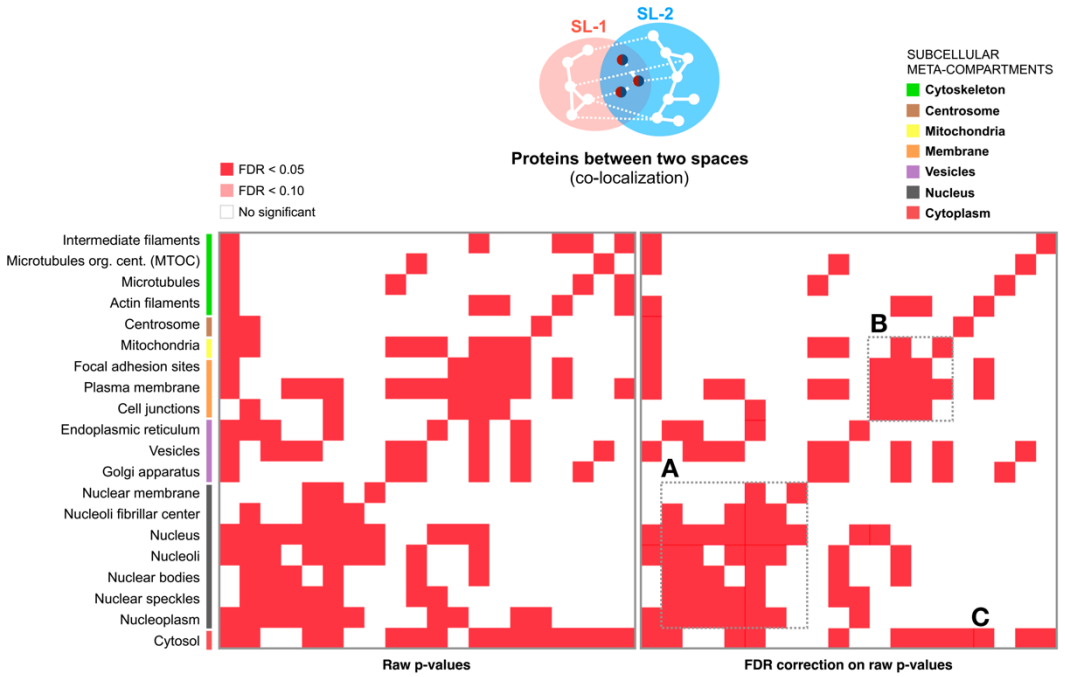


Figure 3-R-7. Enrichment analysis for shared proteins between pairs of subcellular localizations of Cell Atlas, using HI-III as reference. The left panel shows p-values before FDR correction, while the right panel shows after FDR correction. **(A)** Nuclear compartments are very related through shared or multi-localized proteins. **(B)** Membrane-related compartments seem to be very related. **(C)** The *cytosol* is expected to be a common space for protein localization.

Here, we are searching enriched overlaps, then only right-tailed p-values have been considered as significant. As we may expect, there are several compartments which seem to be more related than others. For example, nuclear compartments are very related each other (**Fig. 3-R-7A**), not only due to their biological relevance but also because of the unexpected bias for nuclear subcellular localizations reported by authors in the original Cell Atlas study (Thul et al. 2017). Additionally, membrane-related compartments are also very suitable for sharing proteins due to their physicochemical characteristics and the communication with the extracellular environment (**Fig. 3-R-7B**). Alternatively, *cytosol* is intended to be one of the most occupied cellular spaces, as reflected by HI-III (**Fig. 3-R-7C**).

4.2 Enrichment analysis for protein-protein interactions between compartments

Since **figure 3-R-7** displays the main trafficking of proteins (detected by HI-III) among subcellular compartments, we can infer which cellular spaces tend to communicate each other more often. In this way, we may hypothesise the following: **existing protein-protein interactions between proteins located at one of those commonly related compartments should move to interact from one to another subcellular space**. First, we assessed the agreement between the trend observed for shuttling proteins (**Fig. 3-R-7**) and protein-protein interactions between compartments. Given a pair of subcellular spaces, we considered only those PPIs occurring between proteins annotated to one or both compartments, so interactions within a compartment were removed for this analysis. In this way, only cross-talk among compartment is reflected (**Fig. 3-R-8A**). Z-scores were calculated after randomization of HI-III (1000 random networks) as detailed in Material and Methods, then FDR correction to empirical p-value of z-scores was applied.

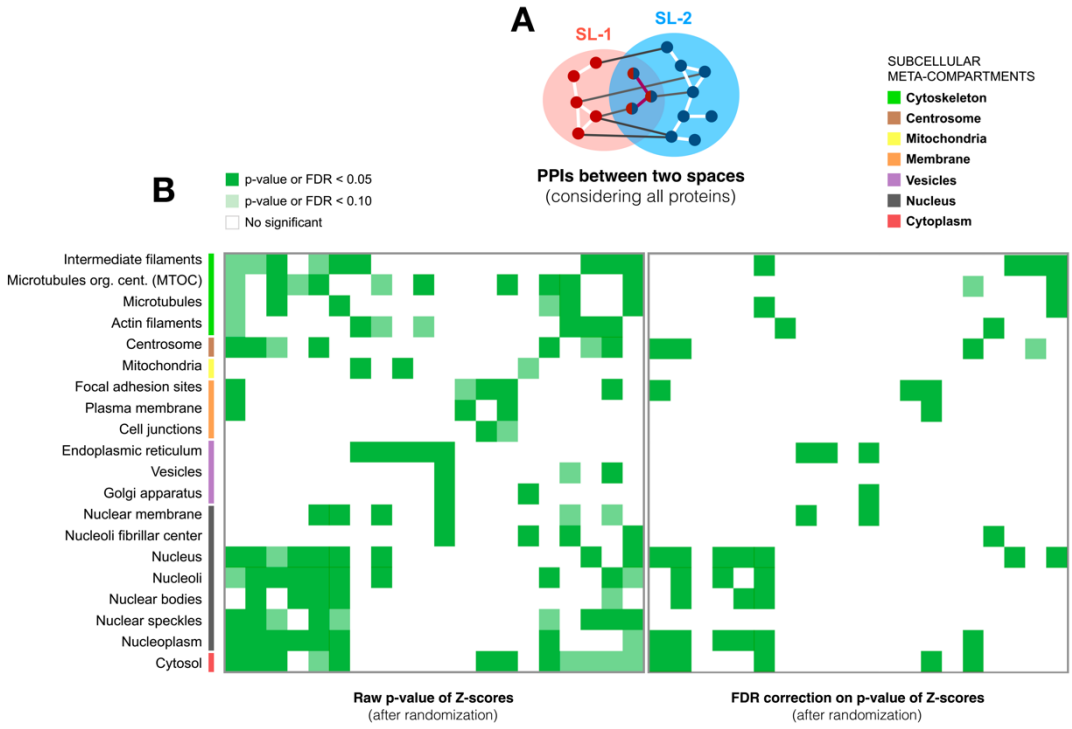


Figure 3-R-8. Enrichment analysis for z-scores obtained after randomization of HI-III, showing the enrichment for PPIs among subcellular compartments of Cell Atlas. **(A)** Scheme of considered PPIs per pair of subcellular spaces, where intra-compartment's PPIs were discarded (white edges) while inter-compartment were considered (dark edges). **(B)** Heatmaps showing significant enrichments for PPIs between two subcellular compartments of Cell Atlas in HI-III interactome.

Figure 3-R-8 shows two similar heatmaps to the previous figure, where significant relationships after randomization of HI-III are plotted. Here, FDR correction and no p-value correction are also shown, accordingly to previous protein enrichment's maps. As we can observe, there is a relevant agreement among protein enrichment (**Fig. 3-R-7**) and PPI enrichment (**Fig. 3-R-8**) according to HI-III interactome. Main relationships like nuclear and membrane-related compartments are also reflected by PPIs of HI-III, which may indicate that proteins annotated only to one compartment are really located at those two compartments.

4.3 Agreement between shuttling proteins and protein-protein interactions between subcellular compartments

Once we have determined the existing enriched relationships among compartments for shuttling proteins (based on Cell Atlas' localizations) and for protein-protein interactions (based on HI-III), we integrated both maps to determine if this interactome is able to report a similar trend. In order to assess if this agreement between previous heatmaps is significant and to highlight the most consistent relationships, we overlapped both maps (cut-off at p-value or FDR < 0.05) and calculated Fisher's exact test (**Figure 3-R-9**).

As we can see, both panels A and B indicate that PPI enrichment follows a similar trend (not expected by chance) with or without FDR correction of empirical p-values from z-scores and odds ratios. In fact, we can observe three main regions: p1 and p3 reflect overlapping regions where protein and PPI enrichment were found, while p2 highlights a PPI's enriched region where *endoplasmic reticulum* seems to be a key compartment for interactions among vesicle-related spaces and nuclear membrane. Furthermore, the odds ratio of the overlap after FDR correction is even higher than previous one, because the most significant PPI's enriched squares (green) keep overlapping protein's enriched squares (red) (**Fig. 3-R-9B**).

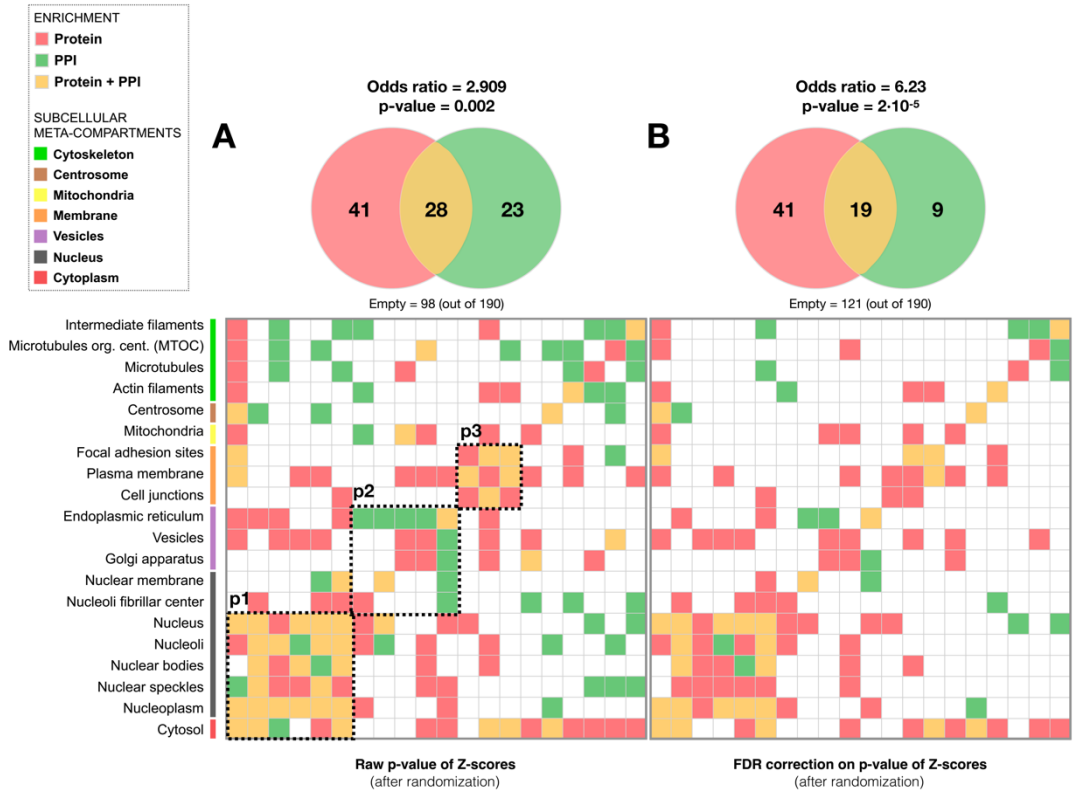


Figure 3-R-9. Combined heatmaps of shuttling protein's enrichment (red squares) and protein-protein interaction enrichments (green squares), before **(A)** and after FDR correction **(B)**. Significance threshold chosen was 0.05 (p-value or FDR < 0.05). Additionally, Fisher's exact test were calculated to determine if there was a significant overlap between both enrichment analysis.

Although these analyses are based on a discrete or binary value (enriched or not based on p-value's cut-off), it may be bias derived from significance thresholds chosen. For this reason, we also evaluated if any parametric or non-parametric relationship is reflected by both enrichments. In this way, we recalculated the z-scores obtained after randomization of HI-III into ratios (ratio = real observed PPIs / average of random PPIs) for making them more comparable to odds ratios. Moreover, both ratios were transformed into logarithmic scale before compare.

Figure 3-R-10 displays a plot facing odds ratios from protein enrichment analysis and ratios from PPI enrichment analysis. We also calculated two statistics for estimating if a positive trend may be inferred: **(i)** linear regression existing (blue line, grey shadow confidence intervals 95%) and **(ii)** Spearman's correlation, which resembles a non-

parametric measurement of relationship. As observed, Spearman's correlation is positive and significant (corr = 0.301, p-value = $3.466 \cdot 10^{-8}$), while linear regression follows an increasing trend (slope = 0.942, intercept = -0.101). Since these two ratios do not follow similar distributions, Spearman's correlation seems to be a more reliable result. According to previous results shown in **figure 3-R-9**, these results also indicate there is a slight trend for HI-III to connect (PPI enrichment) those compartments where exist an enrichment for shuttling proteins (protein enrichment) (**Fig. 3-R-10**). In fact, several relationships among compartments were shown in grey in the figure in the top-right square of the plot, most of them expected due to biological characteristics of each compartment.

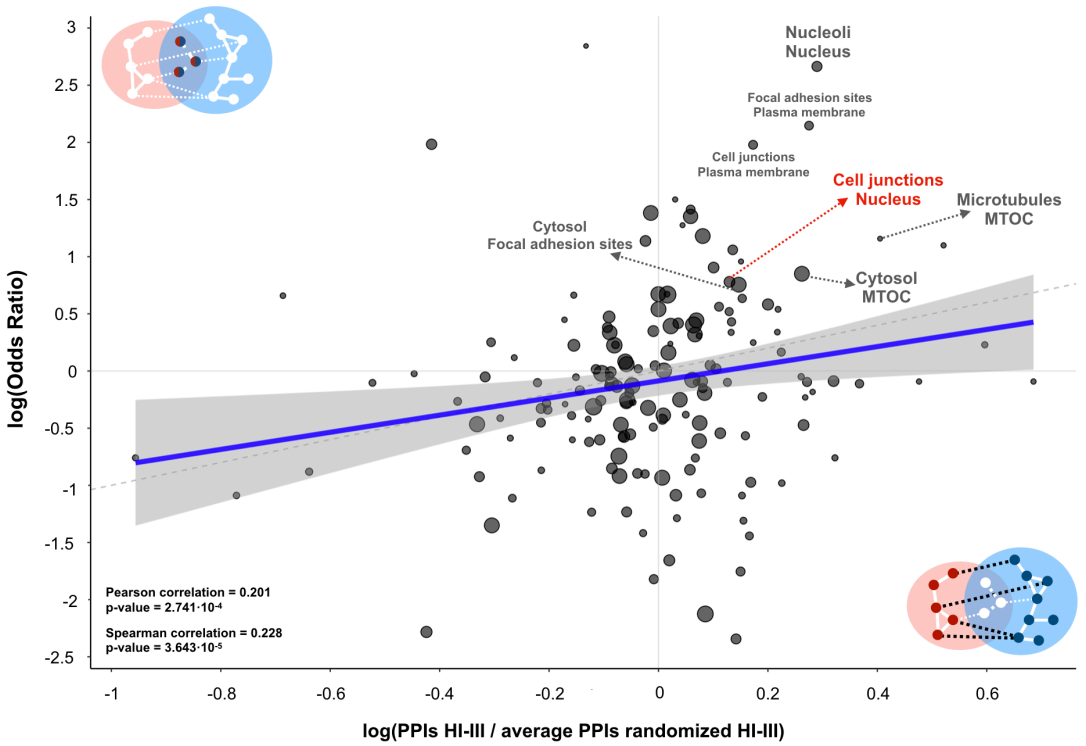


Figure 3-R-10. Scatter plot between odds ratios from protein enrichment analysis (y-axis) and ratios from PPI enrichment analysis (x-axis). Log-transformation was previously applied. Additionally, linear trend (blue line, grey shadow are confidence intervals 95%), Spearman and Pearson correlations between both data are shown (bottom-left corner). Point size corresponds to relative number of proteins included in each pair of subcellular compartments under consideration (graph on bottom-right corner).

The trend shown in **Figure 3-R-10** did not include all those pairs of subcellular compartments where returned odds ratio (protein enrichment, Fisher's exact test) was positive or negative infinitive, indicating a complete or null overlap between spaces, respectively. Indeed, only overlap with itself (diagonal from **fig. 3-R-7**) returned positive infinitive while negative for several pairs. Due the impossibility of considering these values for the trend, we assessed the differences between full or no-overlap of subcellular spaces in terms of PPI's ratios (**Fig. 3-R-11**). As observed, the Wilcoxon-Mann-Whitney test was significant. We proved the number of observed PPIs is significantly smaller if proteins are located in different subcellular compartments that if the proteins share location in two subcellular compartments. In conclusion, this result empowered our previous results of HI-III's ability to disclose interactions among proteins of very related subcellular compartments.

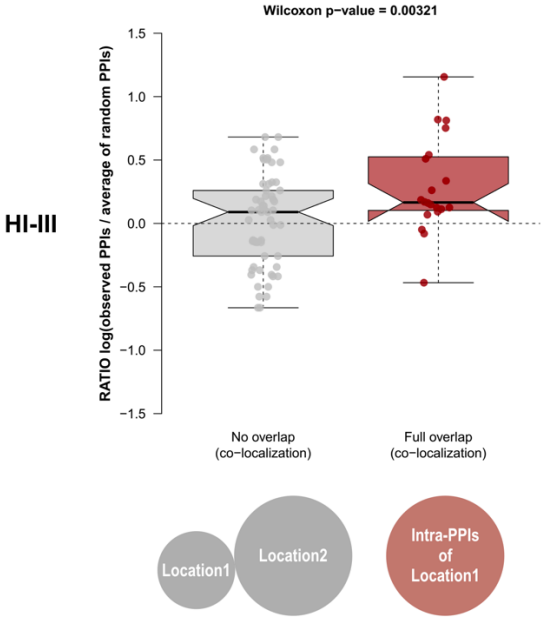


Figure 3-R-11. Boxplots showing the comparison of PPIs between proteins present in two subcellular locations with no-overlap (grey) or with full-overlap (dark red). Wilcoxon-Mann-Whitney test was applied to assess differences between these two distributions.

It is noteworthy the positions taken by greater pairs of subcellular localizations in **Figure 3-R-10** (point sizes are relative to number of proteins), which seem to follow a higher correlation than the rest. In order to test this observation, a Weighted Pearson Correlation Coefficient (WPCC) was calculated using the number of proteins within each pair of subcellular compartments as weights. Additionally, we also put an iterative threshold

to remove smaller pairs of subcellular compartments, then WPCC was calculated by weighting as zero all those pairs showing fewer proteins than threshold. These results are shown in **Figure 3-R-12** (only significant, p-value ≤ 0.05). As observed in this figure, there is an increasing trend of WPCC corresponding to greater pairs of subcellular compartments. Since the *cytosol* and *nucleoplasm* involve the highest number of proteins in HI-III, we can conclude that the trend observed in **Figure 3-R-10** is notably due to these two compartments.

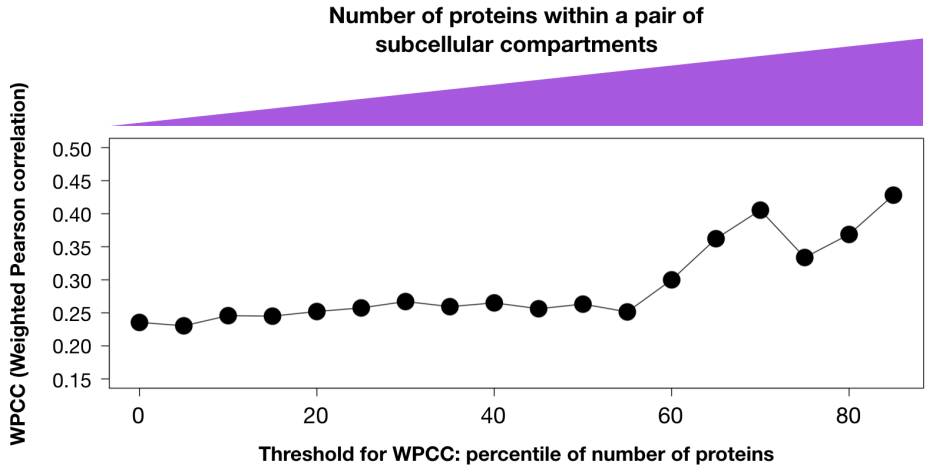


Figure 3-R-12. Weighted Pearson Correlation Coefficient (WPCC) for different thresholds of number of proteins are shown. An increasing trend is observed up to percentile 85 (only significant WPCC are shown).

Interestingly, **figure 3-R-9** could be represented as a network where nodes are represented by these 20 subcellular compartments chosen and edges are significant overlaps derived from shuttling proteins or significant enrichment from PPIs of HI-III (**Fig. 3-R-13**). This cellular network provides a whole perspective of how subcellular meta-compartments are connected if only enrichment of shuttling proteins is considered (panel B) or enrichment of PPIs in HI-III interactome (panel A). In fact, information provided by HI-III seems to be more relevant for grouping subcellular compartments belonging to the same meta-compartment. For example, *vesicles* and *cytoskeleton* meta-compartments are separated only if shuttling proteins are considered, although *plasma membrane*, *cytosol* or *nucleus-nucleoplasm* compartments play a central role within this cellular network in **Fig. 3-R-13B**. In comparison with other subcellular localization databases, this type of cellular

network plot helps to understand how shuttling or shared proteins from different sources are differentially relating subcellular spaces. Binder and colleagues created COMPARTMENTS as a literature-curated database to collect subcellular localization of proteins from different primary resources (Binder et al. 2014). They performed a similar cellular network of shared proteins using their curated database, where *cytosol*, *nucleus* and *cytoskeleton* are very related and separated from vesicle-related subcellular compartments, similarly to our PPI-enriched cellular network (**Fig. 3-R-13A**).

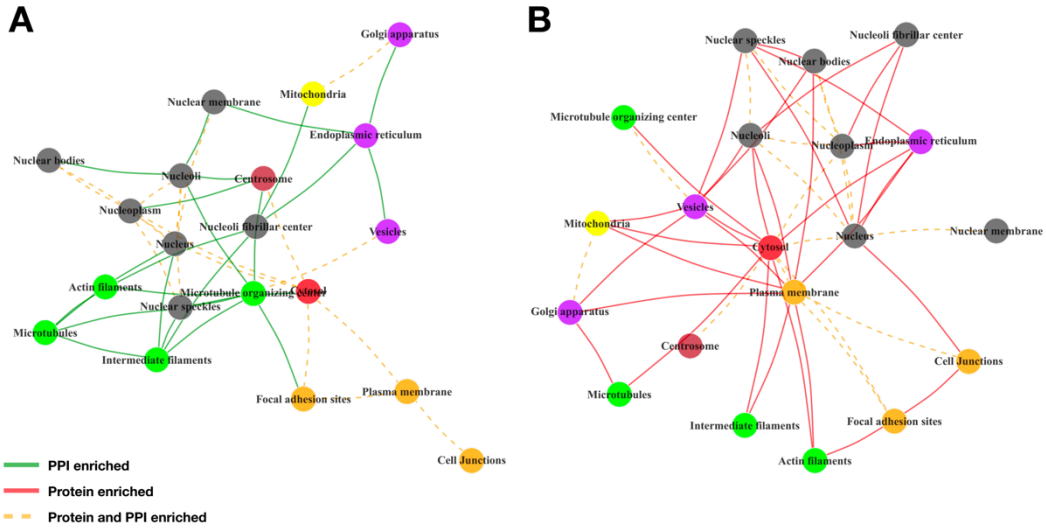


Figure 3-R-13. Network representation of cross-talk between subcellular compartments if only PPI enrichment is considered **(A)** or protein enrichment **(B)** given a human interactome (HI-III). Heatmap from figure 3-R-6A was adapted into network representation.

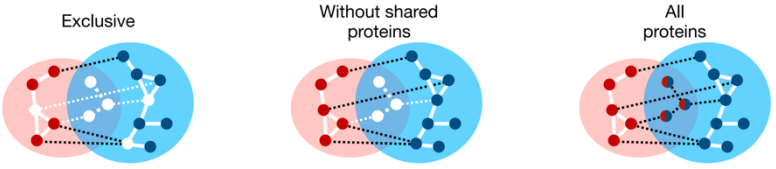
4.4 Three different integrative scenarios for assessing cross-talk

As detailed above, we considered proteins present in both compartments composing each pair of subcellular spaces (**Fig. 3-R-8A**) to produce all posterior analyses of PPI enrichment. However, our results demonstrating a cross-talk tendency of HI-III to connect proteins from very related localizations may be completely conditioned by this assumption. If a great overlap of proteins is found, previous analysis may only reflect a significant number of protein-protein interactions from proteins annotated to both subcellular compartments, bringing any additional information from HI-III to Cell Atlas. Thus, it is also need to determine if there is still a trend when shared proteins are included in the analysis. Otherwise, this complete integration of human interactome and subcellular localization datasets may be irrelevant. For this reason, we reproduced all the previous

analyses shown in Results, following these scenarios:

- **Exclusive proteins:** considering only proteins exclusively assigned to one unique subcellular compartment by Cell Atlas resource (**Fig. 3-R-5A**, black bars in barplot), interacting with other different subcellular space (**Fig. 3-R-14A**). Subnetwork of HI-III only composed of 1516 proteins and 3143 PPIs.
- **Without shared proteins:** considering all proteins located in one out of two spaces composing a pair of subcellular compartments. Thus, shared proteins between two compartments were discarded (**Fig. 3-R-14A**). Subnetwork of HI-III composed of 4125 proteins and 13668 interactions.
- **All proteins:** all proteins included in a pair of subcellular spaces interacting with another subcellular compartment (**Fig. 3-R-14A**). The results of this scenario were shown above. Here, intra-compartmental PPIs among proteins exclusively located at one compartment have been not considered (763 proteins and 904 interactions).

A



B

	Number of proteins	Number of PPIs	Odds ratio between protein and PPI enrichments (discrete relationship)		Spearman's correlation (non-parametric)	Linear regression (parametric)
			FDR corrected	Empirical p-values		
Exclusive	1467	2956	1.81 p-value = 0.519	1.595 p-value = 0.393	-0.08 *p-value = 0.1376*	Slope = 0.08 Intercept = -0.14
Without shared proteins	4125	13668	1.597 p-value = 0.459	1.79 p-value = 0.178	**0.228** *p-value = 3.643·10⁻⁵*	Slope = **0.749** Intercept = -0.008
All proteins	4518	17471	**6.23** **p-value = 2·10⁻⁵**	**3.202** **p-value = 0.00059**	**0.301** *p-value = 3.466·10⁻⁸*	Slope = **0.942** Intercept = -0.101

Figure 3-R-14. Three different scenarios were tested to verify HI-III cross-talk among subcellular compartments. **(A)** Schematic design of three proposed scenarios. **(B)** Table showing results for three scenarios from different analyses described along this Results Section 2.

To summarize all these results and to avoid reproducing same figures using different subsets of data, a table including main results and comparisons is shown in **Figure 3-R-14B**. The second proposed scenario (without shared proteins) still reflects a positive linear trend and significant Spearman's correlation among protein enrichment and

protein-protein interaction enrichment. Thus, we can confirm that HI-III significantly tends to connect proteins annotated to different but related subcellular compartments, even if shuttling or shared proteins are not included in the global analysis.

5. Agreement of given prediction for subcellular localization based on protein-protein interactome (HI-III).

Once we have demonstrated that HI-III is able to disclose protein-protein interactions among related subcellular compartments, we will draw a simple proposal to infer subcellular localization for unannotated proteins included in the analysed human interactome. Based on a protein-protein interaction network, we may suppose that **any given protein should be co-localized in the most frequent subcellular localizations of their first neighbours**. Considering HI-III as reliable human protein-protein interactome, we could look for those proteins whose neighbours are prominently located at a particular subcellular localization. Aiming that, we calculated the enrichment of subcellular localization from partners of each protein reported in HI-III, then Fisher's exact test was applied to define these significant compartments per protein.

First, we evaluated the agreement between Cell Atlas information and HI-III prediction to verify that previous demonstrated trend is also present at protein-level. Thus, we considered only *located* proteins of HI-III on main 20 subcellular localizations mentioned before (**Fig. 3-R-6**), and all those proteins showing a reasonable degree (degree > 3) for this analysis. A total of 2607 out of 8030 proteins remained after these considerations. Moreover, after applying the enrichment test (Fisher's exact test), 1965 out of 2607 proteins analysed showed any enriched subcellular localization within its neighbourhood (at least one subcellular compartment with p-value ≤ 0.05). Finally, we could match both Cell Atlas original subcellular localization and this HI-III prediction based on protein's neighbourhood.

Figure 3-R-15 shows the agreement between the initial Cell Atlas subcellular localization and HI-III prediction (also based on Cell Atlas dataset) per protein. While Cell Atlas assigns 3175 different localizations for these 1965 proteins (grey colour, **Fig. 3-R-15A**), HI-III predicts 2821 subcellular localizations for the same subset of proteins, with an agreement of 600 subcellular localizations for 539 proteins (red colour, **Fig. 3-R-15A**). This

overlap is very significant (OR = 3.556; p-value ≤ $2.2 \cdot 10^{-6}$) and supports our previous results about how using this protein-protein interactome (HI-III) we are notably able to infer the subcellular localization of proteins when enough information is available (degree > 3).

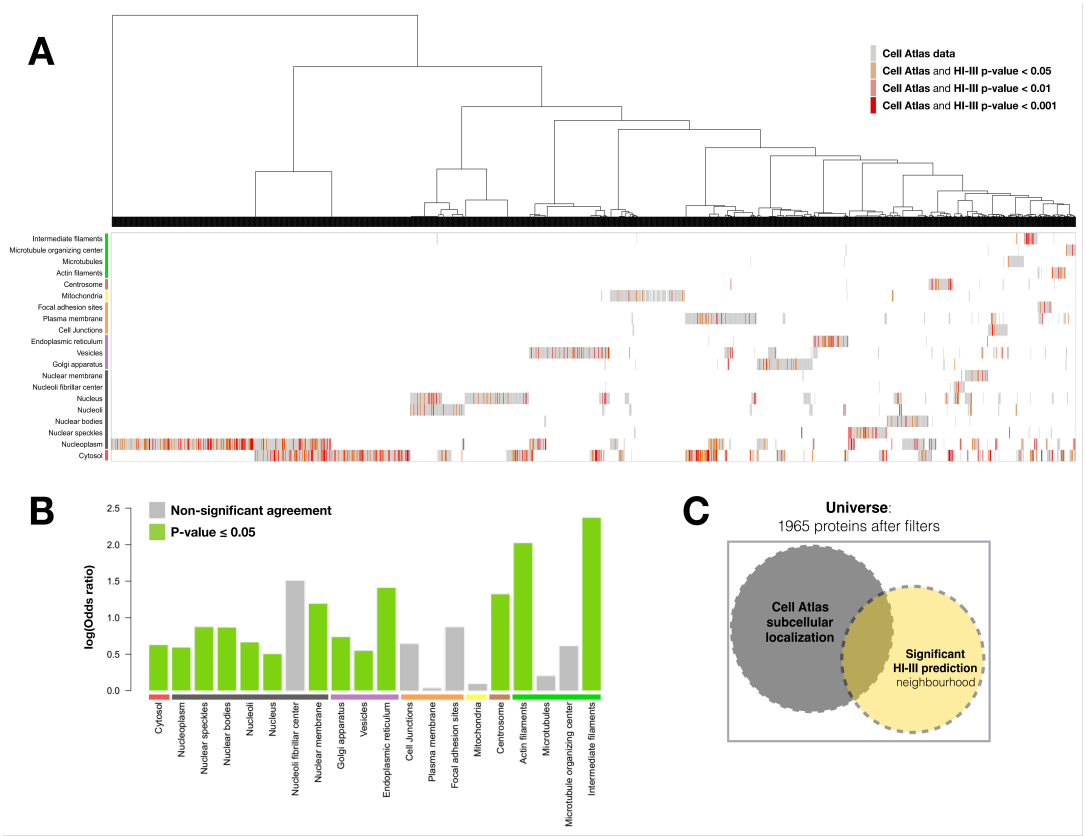


Figure 3-R-15. Agreement between current Cell Atlas subcellular localization and HI-III prediction based on first protein's neighbourhood (1965 proteins). **(A)** Agreement's map ordered by Cell Atlas subcellular localization. **(B)** Significance and enrichment of agreement per subcellular localization. **(C)** Scheme of Fisher's exact test application.

Regarding each subcellular localization analysed, there are several subcellular compartments which this overlap/agreement is no-significant, like all *plasma membrane* or *microtubules* related compartments (**Fig. 3-R-15B**). Interestingly, *intermediate filaments* and *actin filaments* show the most enriched overlaps (**Fig. 3-R-15B**), which may correspond to the great ability of HI-III to detect protein-protein interactions of keratin-related proteins even when they have been removed at first. Several of these proteins are intended to be part of these subcellular localizations. Moreover, about a 60% of these common subcellular localizations are assigned to *cytosol* or *nucleoplasm* (357 out of 600

assignments), which also reinforces previous results highlighting the greatest pairs of subcellular compartments (**Fig. 3-R-11**) as the most representative for the trend showed in **Figure 3-R-10**.

DISCUSSION AND FUTURE WORK

Throughout this Chapter III, we presented the analyses and results derived from an ongoing collaboration with Marc Vidal's laboratory and Human Reference Interactome (HuRI) project. The last version of this interactome (HI-III) has been the subject of analysis and integration with Cell Atlas dataset (Thul et al. 2017), in order to: **(i)** define any possible subcellular localization bias of HI-III and other current human protein-protein interactomes; **(ii)** validate if protein-protein interactions of HI-III are related to significantly connected subcellular compartments by shuttling proteins (multi-localized proteins); **(iii)** assess the ability of HI-III to disclose PPIs between non-co-localized proteins but both localized in significantly connected compartments (cross-talk); and **(iv)** compare Cell Atlas dataset with proposed HI-III prediction of subcellular localization.

As detailed above, we used a high-throughput and experimental resource (Cell Atlas) to define possible subcellular localization bias of HI-III interactome and other current protein-protein interactomes (**Fig. 3-R-4** and **5**) instead typical approaches using manually curated databases (GO or UniProt). This way, we avoided any literature bias derived from particular analyses carried out on the subcellular localization of a single protein in a specific biological context. We demonstrated that HI-III is the less biased systematic human protein-protein interactome, which made it more suitable for posterior analyses about subcellular localization inference through a protein-protein interaction network.

Shuttling proteins are intended to be those proteins which travel from one subcellular compartment to another for participating in different biological functions. Here, we evaluated how HI-III reflects the connectivity or relationship among different compartments through the enrichment on this kind of proteins. We called **shuttling** or **shared proteins** to those multi-localized proteins which have been annotated to two or more different compartments by Cell Atlas. **Figure 3-R-7** reflected the connectivity among the greatest 20 subcellular compartments due to shuttling or shared proteins enrichment (Fisher's exact test). Main macro-subcellular compartments seemed to be related, while *cytosol* was sharing (as expected) proteins with multiple subcellular spaces. Moreover, we

assessed the same enrichment among compartments for PPIs (**Fig. 3-R-8**), using a randomization of HI-III network as the basis for significance calculation. The agreement between shared proteins and PPI connecting two different subcellular spaces was tested in two different ways: discrete (**Fig. 3-R-9**) and continuous (**Fig. 3-R-10**), both significant and showing a trend of HI-III to connect through PPIs those subcellular compartments very related by shared proteins (Cell Atlas). Interestingly, HI-III connects meta-groups of subcellular compartments properly if we attended to PPIs, while shared proteins slightly differed to these meta-groups (**Fig. 3-R-13**). All these results lead us to conclude that HI-III connects close subcellular compartments involving new information which did not disclose only by considering shared proteins, like original publication of Cell Atlas described (Thul et al. 2017). Finally, based on these results, we briefly validated subcellular localizations inferred by HI-III protein-protein interactome in comparison with Cell Atlas original subcellular localizations.

The agreement between Cell Atlas and HI-III prediction reinforces our main idea that a combination of two systematic and distant approaches like Y2H for protein-protein interactomes and antibody imaging for detecting the subcellular localization could be used to improve both techniques and validates existing data. However, it should not be considered as excluding approaches, since both are systematic approaches and presents different precision, sensitivity and error rates depending on protein characteristics, subcellular compartment or organism. In this way, a specific protein-protein interaction notably reported by Y2H between non-co-localized proteins may be potentially considered as candidate subcellular emplacements for these proteins, because transient protein-protein interactions (**Fig. 3-I-1**) and shuttling proteins are more difficult to detect through antibody imaging.

Interestingly, this collaboration also revisited the ontology of PSI-MI methods from OLS database to create several meta-groups of PSI-MI methods. These meta-groups have been used to collapse similar methods which have been assigned to the same experiment from similar primary databases of protein-protein interaction (**Fig. 3-R-1** and **2**). Thus, we manually curated this ontology due to our experience from APID literature-based interactomes (Alonso-Lopez et al. 2016) and from literature-based interactome used to compare previous version of HuRI project (Rolland et al. 2014). The proposed manual curation and used to create LitBM-17 interactome is included in **Appendix 3**.

We are currently collaborating with Marc Vidal's laboratory to fill gaps related to this integration of subcellular localization information with Human Reference Interactome (HI-III). We aim to properly analyse networks reflecting the cross-talk between two subcellular compartments in HI-II and other human interactomes mentioned. In fact, the significant inference of subcellular prediction given by HI-III would lead us to develop different approaches to improve this prediction. We are investigating different approaches (i.e. edge weighting based on confidence score from HI-III experiments and integration of several human interactomes or profile similarity scores) pursuing this intention.

CHAPTER IV

Co-expression network of the human proteome: integrating tissue-specific and evolutionary timeline information

BRIEF SUMMARY

Throughout this chapter, we will describe an integrative analysis describing the evolutionary age of human protein-coding genes based on transcriptomic data and public databases. This analysis arose from a successful collaboration between Katia de Paiva Lopes from Jose Miguel Ortega's lab and our laboratory, which was published in *BMC Genomics* in 2016 (Lopes et al. 2016).

INTRODUCTION

Transcriptomic technologies have been broadly applied to the study and analysis of a wide variety of organisms, especially eukaryotic transcriptomes (Wang et al. 2009). In the Chapter I, we described the most relevant transcriptomic platforms like microarray, ESTs or RNA-sequencing. The RNA-sequencing (RNA-seq) methodology has greatly improved our ability to measure not only gene expression levels (mRNA), but also other RNA molecules (miRNA, iRNA, non-coding RNAs, etc) and biological processes (splicing). As previously mentioned (**Fig. 1-I-3A**), its growth during last decade has made RNA-seq the most used transcriptomic platform (Lowe et al. 2017).

Attending to human tissues, systematic analyses have been performed based on transcriptomic technologies to characterise their specific gene expression profiles. For example, the FANTOM project consists of transcriptomic profiling of 56 human healthy tissues associated with the functional annotation of mammalian genomes (Consortium et al. 2014). Alternatively, the Genotype-Tissue Expression Consortium or GTEx (Consortium 2013) compiles a vast resource of human transcriptomic data (RNA-sequencing) of 11688 samples for 54 different specific tissues (used in chapter II). Actually, GTEx was used to infer patterns across human tissues or individuals (Mele et al. 2015) and to produce tissue-specific gene co-expression networks (Pierson et al. 2015). Interestingly, the Human Protein Atlas project consists of well-curated RNA-seq samples from 32 human tissues to find the correlation between gene expression and protein presence/absence along secretome, metabolic processes and membrane, druggable or cancer proteomes (Uhlen et al. 2015). Moreover, there are examples of specific studies on tissues or cell-types using RNA-seq, like the characterisation of the placenta transcriptome from 20 healthy women with uncomplicated pregnancies (Saben et al. 2014).

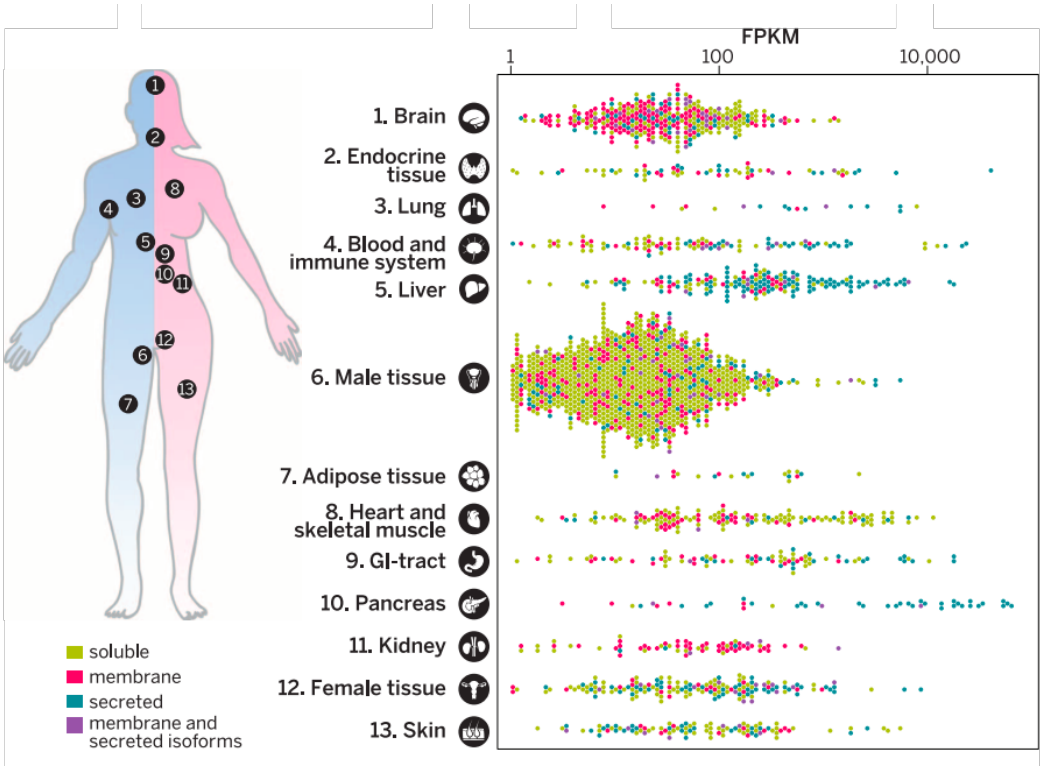


Figure 4-I-1. Figure from Uhlen et al. study presenting the RNA-seq dataset used along this study, which shows the FPKM distributions of tissue-enriched genes for 13 main tissues (Uhlen et al. 2015).

The assembly of comprehensive maps of the human transcriptome is essential for a clear identification of the functional elements of our genome and to reveal the molecular constituents of different cells and tissues (Wang et al. 2009). Despite many transcriptomic studies, little has been reported about the evolutionary determinants of human cell identity, particularly from a joint perspective of protein evolution and gene expression (Sardar et al. 2014). Attending to functional diversity and redundancy of human genome, the evolution of a particular human gene could be informative about the reasons behind its function. Gene age is an important piece of information that can be inferred in different ways and has been used in some genome-scale studies and in some studies on gene families (Capra et al. 2013). Indeed, phylostratigraphy is the common methodology employed to find the origin and emergence of genes (Domazet-Loso et al. 2007, Sestak et al. 2013). Previous phylogenetic studies about human genome revealed relationships with diseases (Domazet-Loso and Tautz 2008), codon usage (Prat et al. 2009), essentiality, interactions (Abrusan 2013), stemness and self-renewal (Hemmrich et al. 2012). Alternative studies have

demonstrated how ancient genes evolve slower (Alba and Castresana 2005), encode longer proteins, present higher expression levels, possess higher intron density and are subject to stronger purifying selection (Wolf et al. 2009, Cai and Petrov 2010). Several of these studies approach the question of gene age in different ways, but most of them are not focused on human genes or do not apply phylostratigraphy using large-scale genomic data.

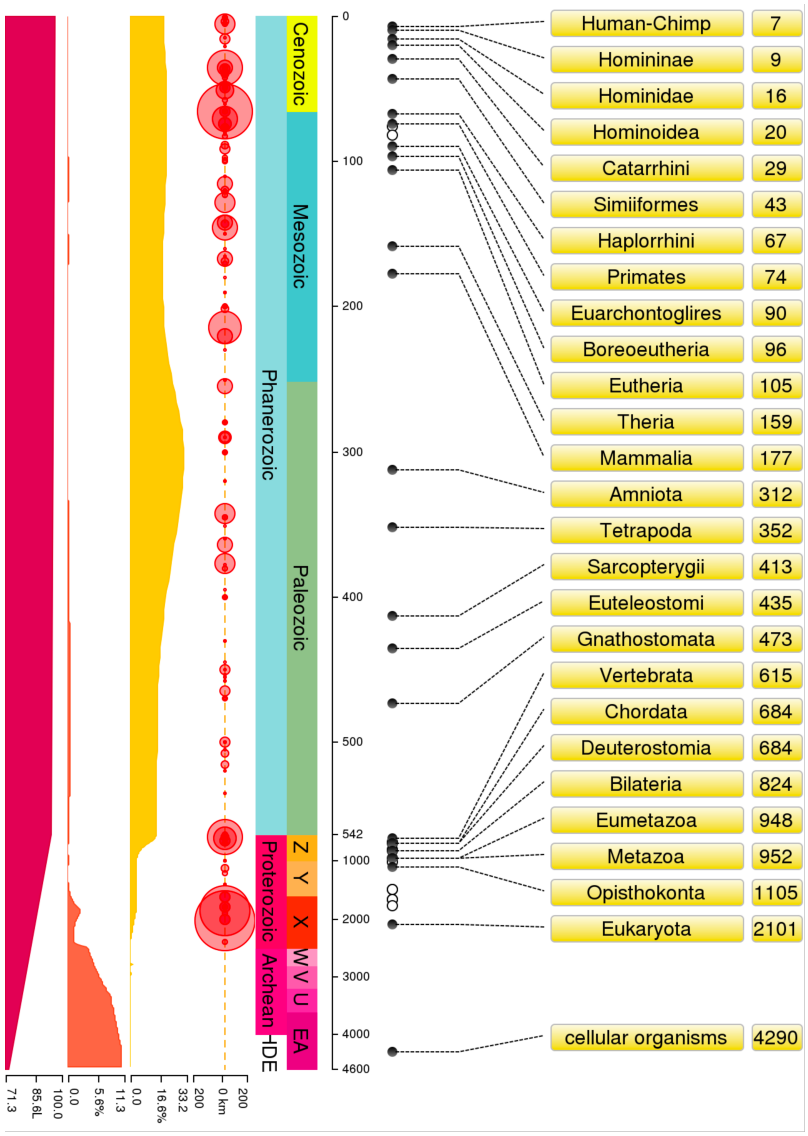


Figure 4-I-2. *Homo sapiens* evolutionary timeline graph obtained from TimeTree resource (www.timetree.org).

Throughout this chapter IV, we addressed the key question about the evolution and age of human genes through the combination of genome-wide data and public databases, aiming to map human genes on the whole evolutionary time-scale. We will manage one of the comprehensive human tissue RNA-seq dataset mentioned above: Human Protein Atlas (Uhlen et al. 2015), which will allow deep expression profiling of protein-coding genes across different human tissues (**Fig. 4-I-1**). Additionally, we will integrate a database of orthologous proteins to find the oldest relatives to each human protein along different species (Altenhoff et al. 2013, Altenhoff et al. 2015). We used taxonomy mapping of these genes to lineage clades from the NCBI Taxonomy database (www.ncbi.nlm.nih.gov/taxonomy); and the time-scale mapping provided by TimeTree resource (www.timetree.org) (Hedges and Kumar 2009).

Clusters of orthologous proteins built along multiple species were demonstrated to be more accurate than simple sequence homology match when we carried out phylogenetic studies (Altenhoff et al. 2013). This involves a conservation along the evolutionary tree instead revealing singular best homologous, providing a simple way to date the origin of different protein modules implicated in specific biological functions and pathways (Donnard et al. 2011). To complete the view of the protein-coding gene phylostratigraphy, we will use the genome-wide expression data to produce a human gene network based on a co-expression analysis of the transcriptomic RNA-seq profiles along multiple tissues, identifying which genes can be considered Housekeeping (HKg) or Tissue-enriched (TEg). We expect the allocation of these gene subsets (HKg and TEg) on the evolutionary time map would show a clear difference in gene age, indicating that housekeeping genes are older.

MATERIAL AND METHODS

1. Gene expression data from human normal tissues

The genome-wide expression dataset used in this work corresponds to a series of RNA-seq analyses performed with Illumina HiSeq 2000 paired end sequencing on cDNA libraries prepared from samples of 122 human individuals from 33 different tissues (ArrayExpress DB: E-MTAB-2836) (Uhlen et al. 2015). The data provided reads for 20,344 genes detected per sample, where 18,545 of these genes showed relevant expression signal corresponding to mean(FPKM) ≥ 1 in all the selected tissues. After normalization and comparative analysis of the expression distributions of the samples from this dataset, we selected a total of 116 samples with two to five biological replicates for the following 32 tissues: adrenal gland (3 replicates), appendices (3), bone marrow (4), brain (3), colon rectum (5), duodenum (2), endometrium (5), oesophagus (3), fallopian tube (5), adipose tissue (3), gallbladder (3), heart (4), kidney (4), liver (3), lung (5), lymph node (5), ovary (3), pancreas (2), placenta (4), prostate (4), rectum (4), salivary gland (3), skeletal muscle (5), skin (3), small intestine (4), smooth muscle (3), spleen (4), stomach (3), testis (5), thyroid (4), tonsil (3) and urinary bladder (2).

2. Expression profiling and co-expression data analyses

RNA-seq expression data from all the tissue samples, taken as normalised FPKM (Fragments Per Kilobase of transcript per Million reads) from (Uhlen et al. 2015), were log2 transformed to obtain the final expression values as: log2(FPKM + 1). Normalised expression distributions of these samples can be seen in **Figure 4-R-1** as density plots and in **Figure 4-R-2** as boxplots. Unsupervised clustering of the samples based on whole gene expression was done applying an agglomerative hierarchical clustering and calculating the distances based on: [1 – Spearman correlation]. This clustering was done for all the 116 samples and just for the 32 tissues using the average expression of the replicates (**Fig. 4-R-3**). Principal Component Analysis (PCA) was performed to alternatively visualize groups

of samples and tissues (**Fig. 4-R-4**).

The co-expression dataset was built calculating the pairwise Spearman correlation coefficient (r) of all the genes (18,545 with mean(FPKM)>1 in all 32 tissues) along the 116 samples and only selecting, as positive gene-pairs, the ones with a correlation coefficient ≥ 0.85 (**Fig. 4-M-1**). Cross-validation of these correlation values was applied by a random selection of two sample replicates from each tissue and recalculating again the Spearman correlation of genes for these random subsets of the data. This sampling was done 100 times, also annotating for each gene-pair the number of times that its r coefficient was ≥ 0.85. Only the gene-pairs validated 100 times in this sampling were selected. A final set of highly correlated gene-pairs was produced including 2298 genes and 20,005 co-expression interactions. This co-expression dataset is provided as Additional file 2 in our website (bioinfow.dep.usal.es/evolutionaryhallmarks), indicating the names of all the gene-pairs and their correlation value. A gene co-expression network derived from the co-expression data was built using Cytoscape (www.cytoscape.org) and we applied the MCODE algorithm to identify clusters inside the network. This algorithm performs an analysis of the topology of the network to find densely connected regions that define modules. The co-expression network built with Cytoscape including all the subnetworks found (with information about the specific proteins in each), as well as the parameters derived from the graph analysis, is provided as Additional file 3 (bioinfow.dep.usal.es/evolutionaryhallmarks).

3. Evolutionary analyses

3.1 Orthologous search for human proteins: Lowest Common Ancestor

For the evolutionary analysis and determination of Lowest Common Ancestor (LCA), we used a database of orthologous proteins: **Orthologous MAtrix** (OMA, http://omabrowser.org/) (Altenhoff et al. 2013, Altenhoff et al. 2015). OMA includes a database and resource with methods for the inference of orthologous among complete genomes. We downloaded the OMA database into a local MySQL database and created Python scripts to search for the Ensembl ID's from our transcriptomic data into this local database and to calculate the LCA into each respective orthologous group. Thus, we obtained a table with the number of protein-coding genes assigned to each clade in the

human taxonomy lineage, as defined by the Taxonomy resource from NCBI (Taxonomy ID 9606 for human, Homo sapiens; database accessed 9 January 2016).

3.2 Mapping of the human taxonomic phyla into evolutionary timeline

Furthermore, we integrated these data with the **TimeTree** of life (www.timetree.org) (Hedges et al. 2015), that includes the tree of living species calibrated to time. The analysis of the number of genes placed along the evolutionary time-scale allowed visualization of the profile of human genes origin for the whole genome (genome-wide) or for specific subsets of genes. In this genes/time profile, we calculated the number of protein-coding genes that had LCA corresponding to each clade (or level) in the taxonomy lineage (31 consecutive levels for human) and we identified certain levels where major changes occur. These are taken as most significant stages and proposed as key evolutionary hallmarks including specific sets of the human protein-coding genes that are identified.

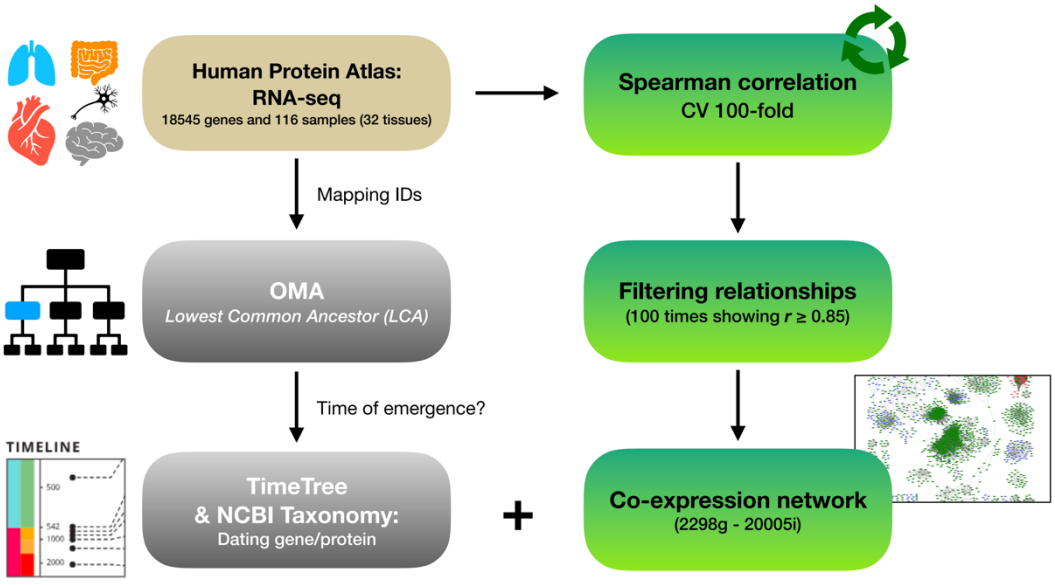


Figure 4-M-3. Workflow of analysis and integration carried out along this study after data normalization and filtering non-expressed genes. We started after filtering non-expressed genes (mean(FPKM) < 1 in all tissues) of RNA-seq dataset from Human Protein Atlas. Then, two parallel analyses were carried out to finally converge in a single co-expression network revealing relationships among differently aged genes.

4. Functional enrichment analysis and identification of gene modules

For the functional enrichment analysis, we used DAVID (david.ncifcrf.gov) (Huang da et al. 2009) and GeneTerm-Linker (gtlinker.cnb.csic.es) (Fontanillo et al. 2011) bioinformatic tools with the list of genes from each evolutionary stage level of the human lineage. In all cases, the enrichment analyses were done using a hypergeometric test and adjusting the p-values for multiple testing with the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995). In the same way, we also investigated the functional enrichment of the subnetworks generated by the clusters and modules found in the analysis of the gene co-expression network.

5. Statistical analyses

All the data analyses and graphics have been produced in the R statistic environment. General functions and statistical tools have been applied over the different data presented. Some specific methods or algorithms are cited along different sections of this chapter.

RESULTS AND DISCUSSION

1. Human global transcriptome profile reveals a clear clustering of similar samples and tissues

As we described above, the RNA-seq data from Human Protein Atlas project (Uhlen et al. 2015) is composed of 116 biological replicates from 32 different human tissues (after samples' filter mentioned in Material and methods). First, we filtered those genes showing low expression values in all different tissues (mean(FPKM) > 1), remaining 18545 genes. Density plots of value distributions before and after filter revealed how values close to zero have been notably reduced, which will allow us to work with expressed genes and avoid problems derived from such accumulation of zero or low gene expression signal (**Fig. 4-R-1**).

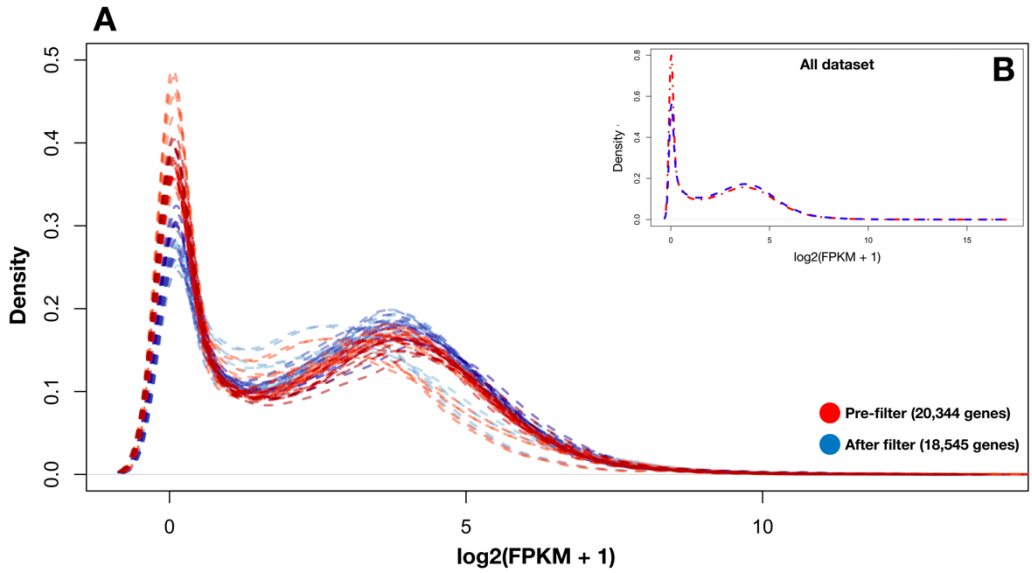


Figure 4-R-1. Density plot of value distributions after normalization of RNA-seq dataset from Human Protein Atlas. **(A)** Density plot per tissue before and after filtering non-expressed genes. **(B)** Density plot using the whole dataset before and after filtering.

Attending to biological replicates, boxplots of value distributions showed in **Figure 4-R-2** point to a slightly variation of FPKM distributions across different tissues and replicates. However, since we will mainly work with non-parametric calculations like Spearman correlations, we considered that value distributions after filtering are enough robust to perform our analysis.

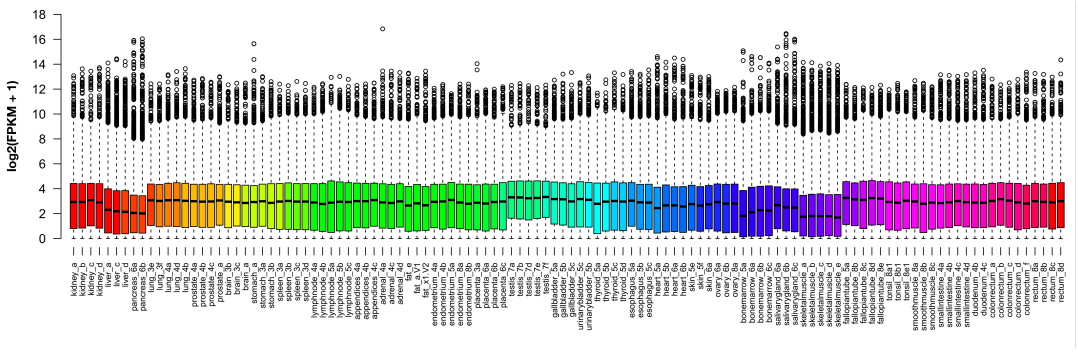


Figure 4-R-2. Boxplots of the expression signal from each one of the RNA-Seq samples studied. These distributions of expression values represented correspond to the log2 of the (FPKM+1) signal for each one of the 116 samples analysed, after filtering non-expressed genes. In total 32 different human tissues are included.

As the first analysis, we conducted an unsupervised clustering analysis based on global expression correlation along 116 samples of 32 human normal tissues and displayed as square heatmap (**Fig. 4-R-3**). All genes after filter were used (18545 genes), generating a dendrogram from hierarchical clustering based on pairwise distances among samples (1 - Spearman correlation). As we can observe, we found a clear relationship among samples from the same tissue, which have been closely placed by this hierarchical clustering (agglomerative method). Moreover, similar tissues were also placed together, like spleen, lymph nodes and tonsils (lymphatic system) or stomach, duodenum, small intestine, rectum and colon (digestive system). Interestingly, the Spearman correlation distribution (top-left panel of **Fig. 4-R-3**) shows a maximum frequency around 0.80-0.85 values while testis tissue seems to be the most different human tissue due to its clear separated branch in the dendrogram.

Additionally, a Principal Component Analysis was performed for the same transcriptomic data. Although the variability explained by main principal component was low (14.006%), those well-separated tissues at hierarchical clustering analysis were placed at the most distant positions from origin when we represent two main principal components (**Fig. 4-R-4**), like testis, brain or bone marrow.

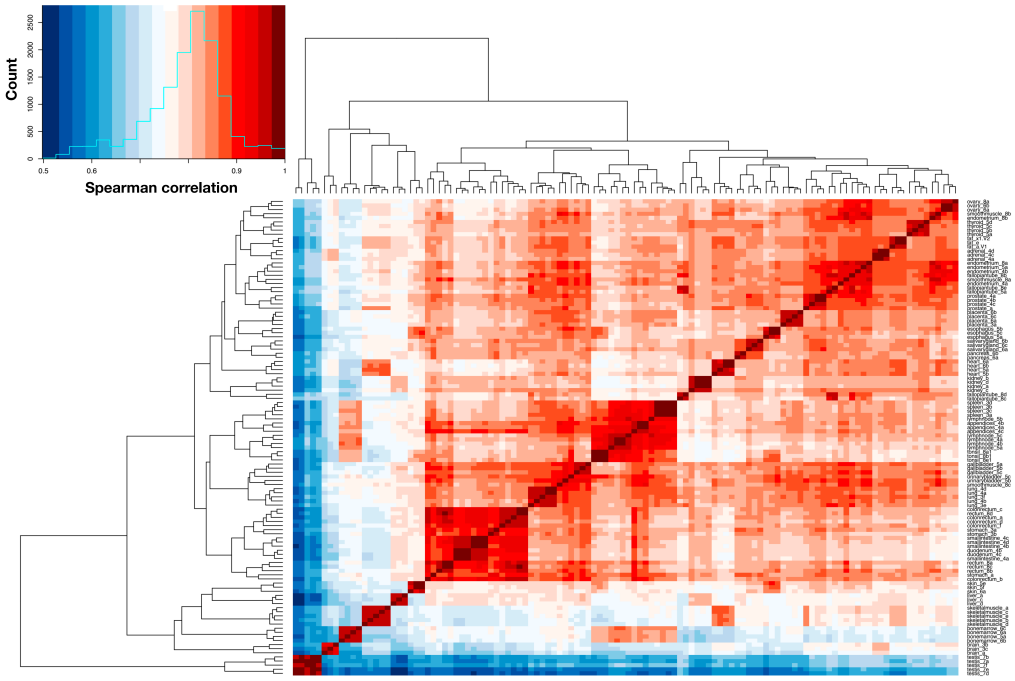


Figure 4-R-3. Heatmap showing square correlation matrix of 116 biological replicates from Human Protein Atlas dataset. Spearman correlation was calculated while 1-Spearman correlation was used as the distance metric for agglomerative hierarchical clustering of samples.

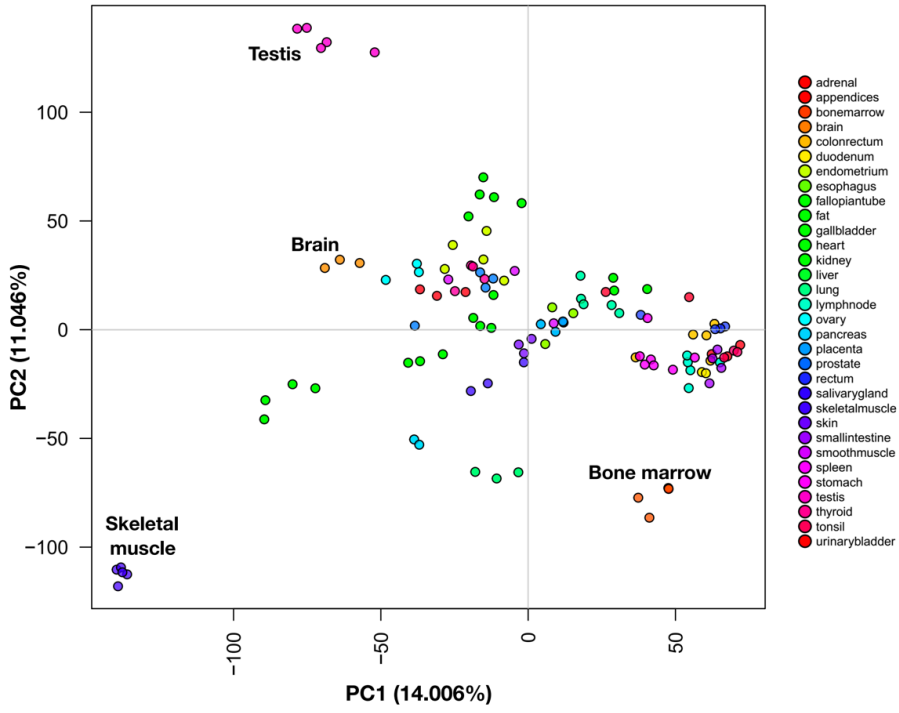


Figure 4-R-4. Principal Component Analysis of 116 samples from Human Protein Atlas dataset, after filtering non-expressed genes.

2. Robust gene expression signal from house-keeping and tissue-enriched genes

Attending to classical biology concepts, we could distinguish between House-keeping (HKg) and Tissue-enriched (TEg) genes. **Figure 4-R-5A** shows the number of genes expressed, with mean(FPKM) equal or higher than 1, per number of tissues. The distribution is asymmetric, showing a great amount of ubiquitously expressed genes along different tissues (8961 genes). Particularly, 7668 out of these 8961 genes are expressed (FPKM equal or higher than 1) in all biological replicates (116 samples). The intersection of these 7668 genes with a curated dataset of 3804 house-keeping genes created by Eisenberg and colleagues (Eisenberg and Levanon 2013) gave a total of 3524 HKg (**Fig. 4-R-5C**), indicating a large overlap of 93 %. Fisher's exact test was calculated to assess the significance of this overlap (OR = 32.09 with 95 % confidence interval, 28.27–36.43; p-value < 0.00001).

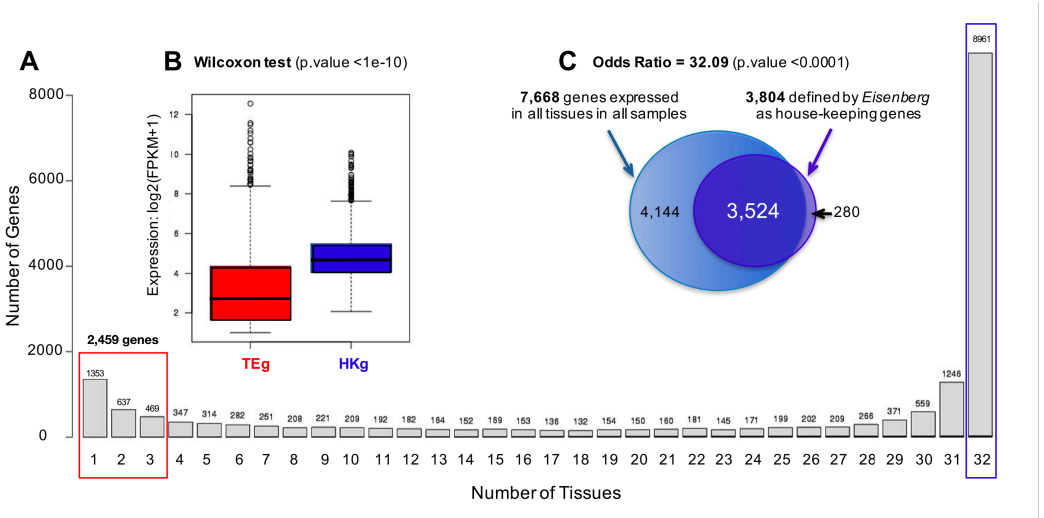


Figure 4-R-5. The number of genes expressed along 32 tissues derived from RNA-seq transcriptomic data. **(A)** Plot showing the number of expressed genes per number of tissues. House-keeping genes as HKg while tissue-enriched genes as TEg. **(B)** Comparison of the expression distributions of HKg versus TEg. **(C)** Venn diagrams showing the intersection of 7668 genes (expressed in all the biological replicates of all tissues) with the dataset of 3804 house-keeping genes obtained from Eisenberg et al. (Eisenberg and Levanon 2013).

The definition of tissue-specific genes is not an easy issue due to the possible differences among intra-tissue variabilities. Specific measurements have been designed in order to mitigate this problem (as explained in chapter II), like Z-score based on median and IQR (interquartile range) designed for GTEx dataset (Sonawane et al. 2017), involving

a posterior heuristic threshold for removing non-specific expression patterns. Here, the original Human Protein Atlas dataset (20344 genes and 116 samples) showed quite notable intra-tissue variability. Indeed, if we had substituted mean by median when the expression filter was applied, we would have obtained 18441 genes showing all tissues with median(FPKM)>1 instead 18545 genes for mean (18416 common genes). Similar to original authors of Human Protein Atlas dataset (Uhlen et al. 2015), we chose a more typical threshold (mean) to avoid False Negatives entering in our filtered dataset. For example, NEUROD2 has been reported as highly brain-specific gene (Chen et al. 2016) while GSTK1 is well known as house-keeping gene (Himmels et al. 2017) (**Fig. 4-R-6**).

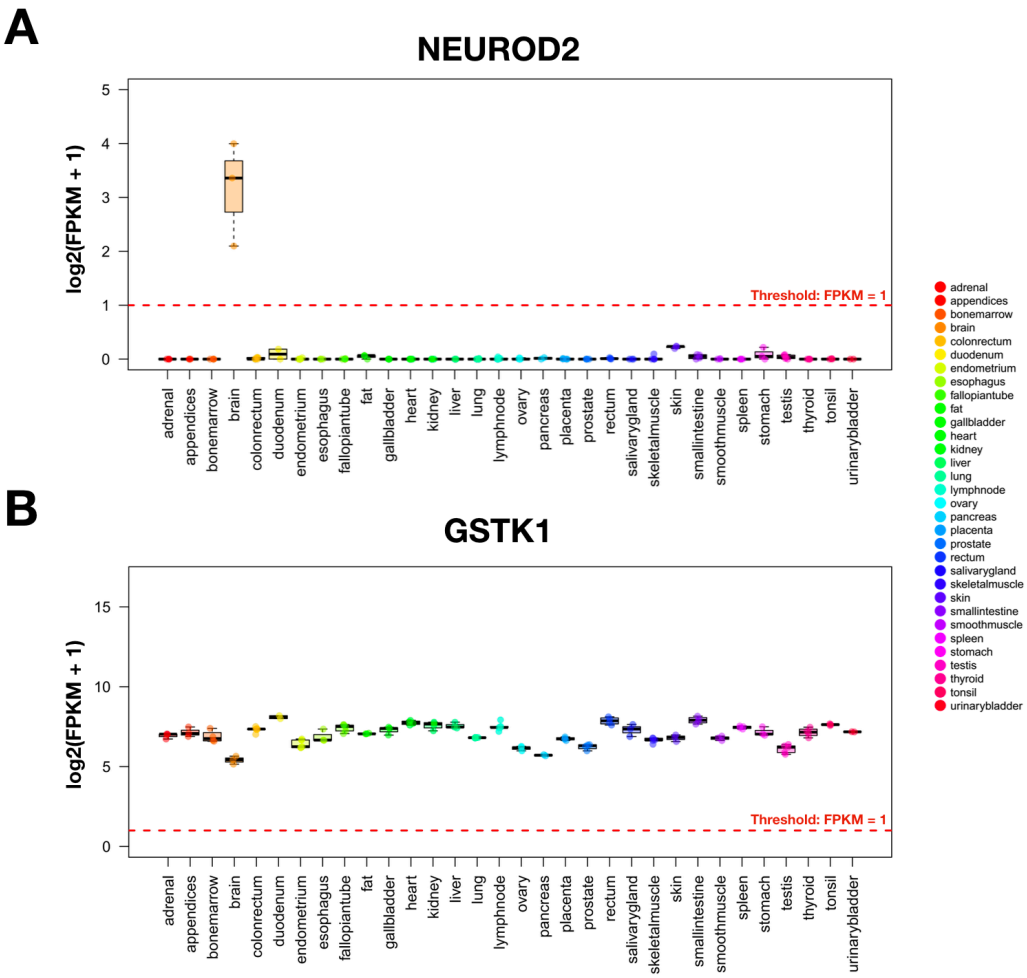


Figure 4-R-6. Examples of gene expression profiles of tissue-enriched **(A)** and housekeeping **(B)** genes from Human Protein Atlas dataset. Mean(FPKM > 1) was considered as threshold, which results in a similar threshold for log scale: log2(FPKM + 1) > 1.

Attending to these considerations, for the tissue-enriched genes analysis we explored the other side of the data in **Figure 4-R-5A** and considered just the genes that were expressed (mean(FPKM) > 1) in only one, two or three tissues (2459 genes). We did not take only one, but also two or three tissues, because some tissues are physiologically very related and in fact presented high correlation between them, for example: colon and rectum; small intestine and duodenum, etc. On the other hand, several tissues showed a notable differential behaviour (testis or brain) in the hierarchical clustering and PCA results with respect to other human tissues. Therefore, we expected an exclusive group of specific expression patterns supporting these evidences (**Fig. 4-R-7B**). Accordingly, **figure 4-R-7A** shows a heatmap of Spearman's correlation across biological replicates based on tissue-specific genes present in two or three tissues (1076 genes). These observations corroborated how related tissues share a common gene expression signal even if only tissue-specific patterns are considered.

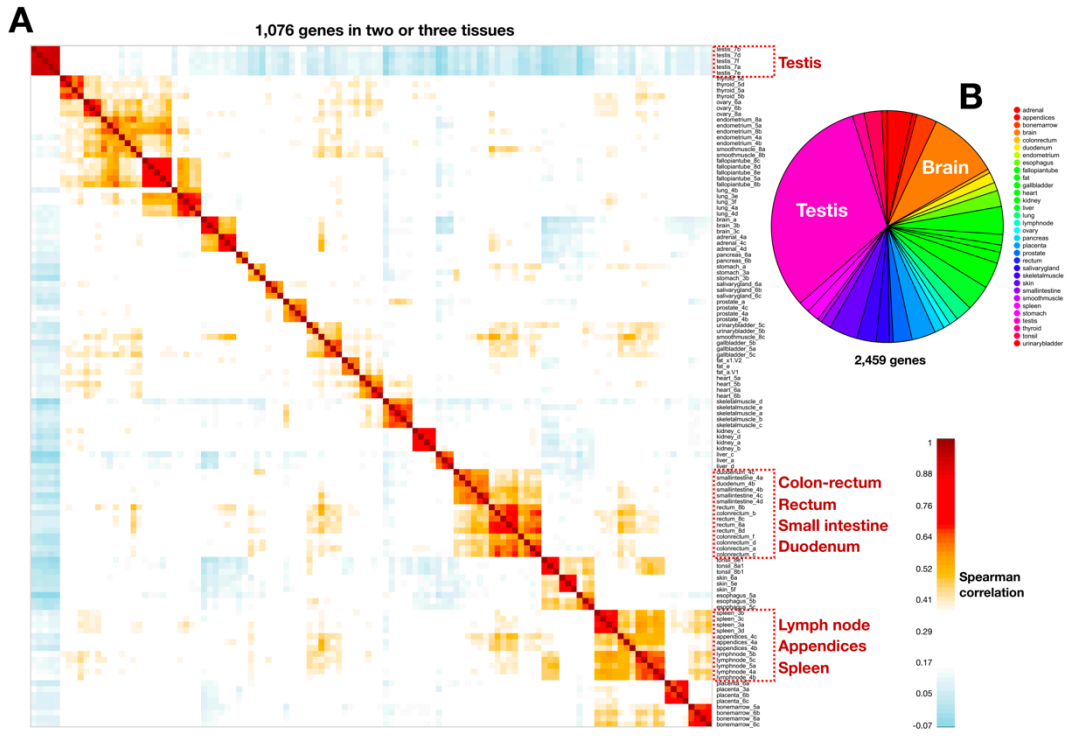


Figure 4-R-7. Analysis of co-occurrence for TEg in our RNA-seq dataset. **(A)** Pairwise Spearman correlation of biological replicates using only TEg appearing in two or three tissues (1076 genes). **(B)** Pie chart showing the frequency of tissues associated with TEg.

Finally, a global comparison of the expression distributions of HKg versus TEg indicated that the Tissue-enriched genes showed significantly lower expression values than the housekeeping genes (**Fig. 4-R-5C**). In order to demonstrate this difference, we conducted two different statistical tests, t-test (assuming normal distribution) and Wilcoxon (rank based test). Both p-values were very low (p-value < 1e-10) while the difference between the mean expressions of TEg and HKg was 1.61 (log2 scale). Interestingly, we can observe the variability of TEg was much larger than the variability of HKg, which could correspond to a tighter regulation of HKg to be considered in further analyses.

3. Human gene hallmarks on the evolutionary time-scale

The evolutionary analysis was carried out through a phylostratigraphic approach for reconstruction of macro-evolutionary trends based on the principle of *founder gene* formation (Domazet-Loso et al. 2007). Typically, these methods first identify the homologues of a given gene and then use the divergence between the two most distant to determine the gene age. Historically, such studies have been using BLAST (Altschul et al. 1990) for homology searches. However, this approach was shown to introduce some biases into the analyses (Moyers and Zhang 2015, Moyers and Zhang 2016).

Another approach is to use orthologous groups to determine the age of a gene. Orthologues are believed to be functionally more similar than paralogues (Koonin 2005) and by definition, they trace back to an ancestral gene that was present in a common ancestor of the compared species (Gabaldon and Koonin 2013). Consequently, the parameters used for clustering orthologous groups affect the age estimations for a gene; for instance, restrictive parameters tend to limit the set of possible progenitors (Capra et al. 2013). Nevertheless, both approaches used for dating the gene origin depend on the correct identification of homologues and/or orthologues, but in the second case, the accurate reconstruction of orthologous families imposes a higher stringency, implying a conservation along the evolutionary clades.

For our analysis, we identified the group of orthologues that corresponded to each of the 18545 genes detected in the transcriptomic study (after filtering non-expressed genes), mapping them to the corresponding human protein-coding genes in the OMA database (mapping of 18545 genes to 17437 proteins). Then, we assigned the Lowest

Common Ancestor (LCA) to each human protein according to its orthologous family. As hinted above, the use of OMA in comparison with BLAST homology approach gives us a more detailed view of gene origin since it uses a more restrictive grouping method. Thereby, the number of genes dated on ancient clades is lower.

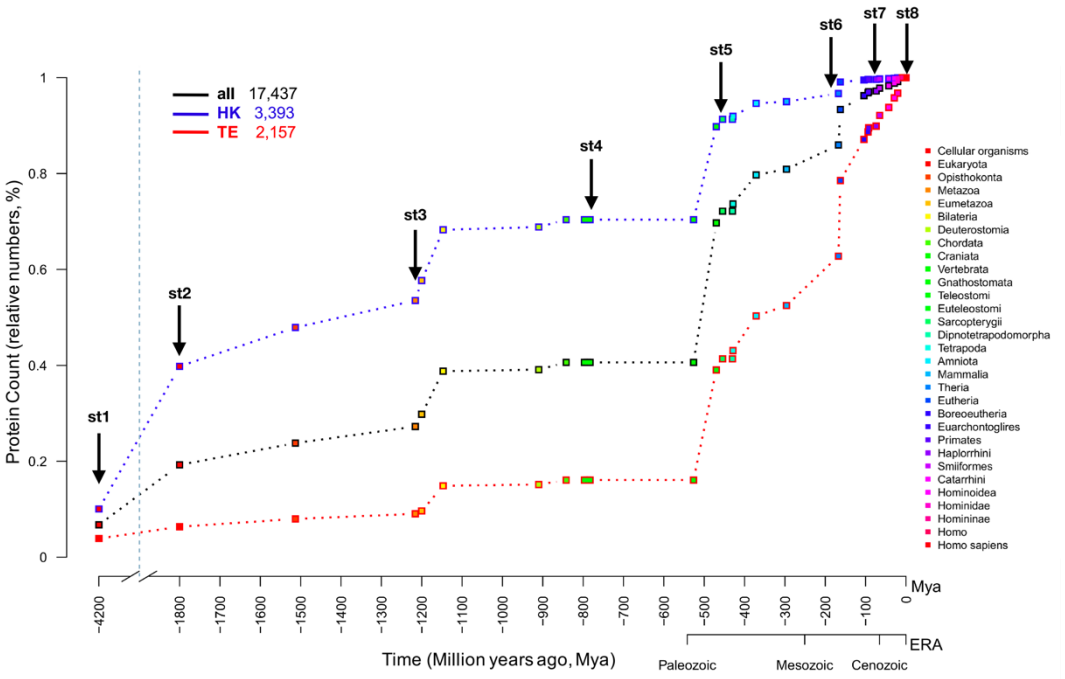


Figure 4-R-8. Evolutionary hallmarks of human protein-coding genes along time-scale. Plot presenting the relative number of human protein-coding genes, which are assigned to each of the 31 taxonomic clades (labelled in colours, legend). For each one of these taxonomy levels, the graph represents the cumulative percentage of protein-coding genes that are dated at such level. The 31 taxonomic clades are placed as dots along the time-scale from the origin to present while arrows point to 8 main *hallmarks*. The black line includes all the 17437 proteins derived from the mapping of expressed genes in OMA. The blue line includes only the HK genes: 3393. The red line includes only the TE genes: 2157.

Once we identified the LCA for each human protein-coding gene, we assigned such protein/gene to the corresponding taxonomy level in the human lineage as defined in NCBI database, which includes 31 taxonomic groups as consecutive clades from the first one, named cellular organisms, to the last one *Homo sapiens*. **Figure 4-R-8** presents these 31 taxonomic clades placed along the time-scale (in million years ago, MYA) from the origin to present.

As we can observe, we represented the cumulative percentage of dated proteins for each one of these taxonomy levels. First, we represented all the genes mapped to OMA

proteins (**Fig. 4-R-8**, black line in the graphic, that includes 17437 proteins); second, the same plot is produced but including only the proteins that correspond to House-keeping genes (**Fig. 4-R-8**, blue line includes 3393 proteins, HKg); third, plot including only the Tissue-enriched genes (**Fig. 4-R-8**, red line includes 2157 proteins, TEg). Interestingly, the analyses of these plots obtained with the phylostratigraphic method revealed the presence of some major differential steps on the emergence of protein-coding genes along the evolutionary time-scale from origin to present. Looking at all the expressed coding genes, we can see the global evolutionary profile of the organism (human), but along this profile, we can identify some more prominent steps in the accumulated relative number of genes along time. For example, a large increase is observed at the start of the curve of HK's trend (blue line), from the first taxonomy level (origin, Cellular organisms) to second (*Eukaryota*) taxonomy level. By contrast, the TE's line (red line) presents the major emergence of genes much later (around the *Mammalia*).

Major STAGES in the evolutionary timescale (hallmarks)			Human genes (all in OMA): 17,437					Human genes (only House-Keeping, HK): 3,393					Human genes (only Tissue-Enriched, TE): 2,157				
Stage level	Species	time Marks (tM)	Gene count (at each tM)	Gene count (cumulative)	Genes (%)	Δ (%) previous-current	AUC	Gene count (at each tM)	Gene count (cumulative)	Genes (%)	Δ (%) previous-current	AUC	Gene count (at each tM)	Gene count (cumulative)	Genes (%)	Δ (%) previous-current	AUC
st1	1st Cellular organisms (Prokaryota)	1	1,178	1,178	0.068	0.07	0.13	341	341	0.101	0.10	0.25	84	84	0.039	0.04	0.05
st2	Cellular org. - Eukaryota	2	2,178	3,356	0.192	0.12	0.22	1,009	1,350	0.398	0.30	0.44	53	137	0.064	0.02	0.07
st3	Eukaryota - Metazoa	4	1,395	4,751	0.272	0.08	0.29	466	1,816	0.535	0.14	0.56	58	195	0.090	0.03	0.09
st4	Metazoa - Vertebrata	10	2,333	7,084	0.406	0.13	0.41	572	2,388	0.704	0.17	0.70	152	347	0.161	0.07	0.16
st5	Vertebrata - Euteleostomi	13	5,070	12,154	0.697	0.29	0.71	658	3,046	0.898	0.19	0.91	495	842	0.390	0.23	0.40
st6	Euteleostomi - Mammalia	18	1,953	14,107	0.809	0.11	0.83	177	3,223	0.950	0.05	0.96	290	1,132	0.525	0.13	0.58
st7	Mammalia - Primates	23	2,821	16,928	0.971	0.16	0.97	157	3,380	0.996	0.05	1.00	799	1,931	0.895	0.37	0.90
st8	Primates - Homo sapiens	31	509	17,437	1.000	0.03	1.00	13	3,393	1.000	0.00	1.00	226	2,157	1.000	0.10	1.00
			17,437			1.00	21.44	3,393			1.00	25.90	2,157			1.00	16.38

Figure 4-R-9. Illustrated table showing data for the 8 evolutionary *hallmarks* of human protein-coding genes, separated in all human genes, HKg and TEg (from left to right).

The complete timeline includes 31 phylogenetic clades (named in the **figure 4-R-8** legend), but by analysing these points it was possible to identify eight major steps or stage levels (what we called *hallmarks*) that appear on the human gene profile along evolutionary time. Moreover, we could assign the number of human protein-coding genes that emerged along each one of these eight stages for the three categories reported: all the expressed protein-coding genes mapped to OMA, the HK and the TE. The eight stage levels identified are: **st1)** Cellular organisms (*Prokaryota*); **st2)** Cellular organisms to *Eukaryota*; **st3)** *Eukaryota* to *Metazoa*; **st4)** *Metazoa* to *Vertebrata*; **st5)** *Vertebrata* to *Euteleostomi*; **st6)** *Euteleostomi* to *Mammalia*; **st7)** *Mammalia* to *Primates*; and **st8)** *Primates* to *Homo sapiens*. All the numbers corresponding to the human protein-coding genes assigned to

each of these eight evolutionary hallmarks are included in the illustrated table in **Fig. 4-R-9**, that indicates how many are in each stage either considering the complete human gene set or just the HK or the TE. All the information about each one of the human protein-coding genes including the assignment to stages is also provided as Additional file 4 (bioinfow.dep.usal.es/evolutionaryhallmarks).

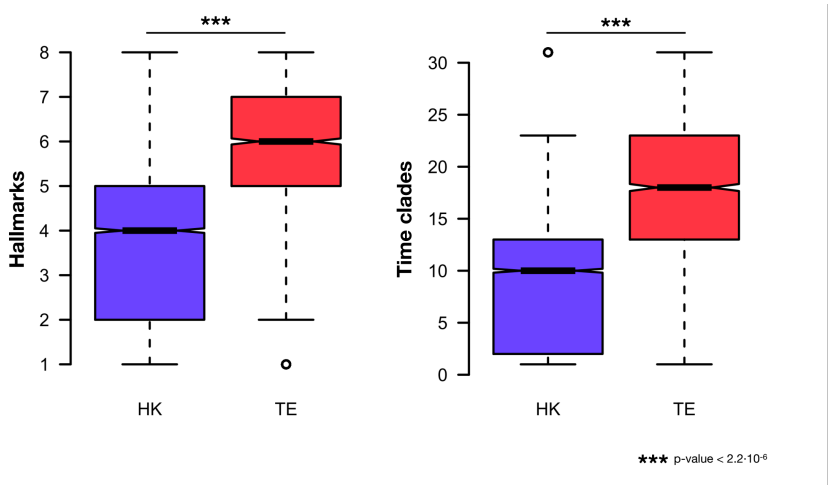


Figure 4-R-10. Boxplot showing distributions of **(left)** *hallmark* assignments and **(right)** *time clade's* assignments for HKg and TEg. Significance was calculated by Wilcoxon test.

The analysis of the hallmarks also reveals that the HK genes are more ancient than TE genes (**Fig. 4-R-8** and **9**). The HK genes present a major increase or expansion in stage 2 (*Prokaryota* to *Eukaryota*), with 1009 genes and a change of ≈ 30 % with respect to the total. By contrast, the TE genes show a major increase in stage 7 (*Mammalia* to *Primates*) with 799 genes and a change of ≈ 37 %. Alternatively, we also conducted a Wilcoxon rank test to compare if there is a significant late-emergence of TE genes in comparison with HK genes (**Fig. 4-R-10**). Whether we took *hallmark's* assignments for both groups of genes as *time clade's* assignments, there was a notable difference between housekeeping and tissue-enriched genes (p-value < $2.2 \cdot 10^{-6}$) which also validates our idea about the late-emergence of tissue-enriched genes.

These observations seem to indicate that house-keeping genes emerged early in evolution and, consequently, are older in age, knowing that they reflect more essential and constitutive functions. This idea relating gene essentiality to older genes was previously reported in several studies, for example on yeast and mammalian genes (Alba and Castresana 2005, Abrusan 2013). By contrast, the observations that human tissue-specific

genes had emerged later in evolution may reveal that human-specific cellular or physiological roles are implemented at molecular level by the appearance of newer functional genes.

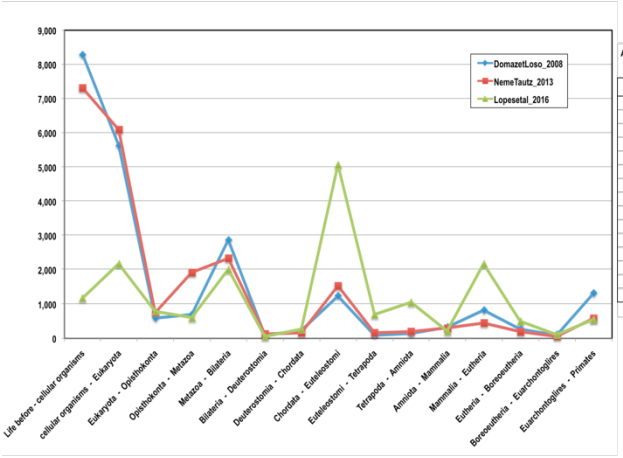
4. Gene age data comparison

As indicated above, there are some studies that use the phylostratigraphic method to explore the age of human genes, but most of these studies use sequence similarity search (with algorithms like BLAST) to look for the oldest homologues to the human (Alba and Castresana 2005, Domazet-Loso and Tautz 2008, Neme and Tautz 2013). To compare the results on human gene age assignment done in this work with alternative available age assignments, we took the published data from Domazet-Lošo (Domazet-Loso and Tautz 2008) and from Neme (Neme and Tautz 2013), and we represented the information about allocation to Lowest Common Ancestor (LCA) of the human genes in phylogenetic clades of the evolutionary tree. The assignments were done using 15 common phylostratum to allow the comparison of the data.

The results of this comparison are included in **Figure 4-R-11** and they show a general similarity but some important differences. The most significant difference corresponds to the fact that both Domazet-Lošo (Domazet-Loso and Tautz 2008) and Neme (Neme and Tautz 2013) placed a very large number of genes on the first stage of the evolutionary time-scale that goes from the origin of life to first cellular organisms (i.e. pre-eukaryota): 8285 of 22,845 (36 %) and 7309 of 22,154 (33 %), respectively. This result denotes a bias that, as we indicated above, can be due to the methodology of using the homology search approach.

In any case, the idea suggested that one third of the human proteome may have emerged in evolutionary time before the origin of eukaryotic cells needs deeper studies and it is not what we observed in our analyses. Another important difference is that the proposed age mapping allocates the largest number of genes, first, to the *Chordata-Vertebrata-Euteleostomi* phylostratums (with 5070 genes) and, second, to the *Mammalia-Eutheria* (with 2172 genes) (**Fig. 4-R-11**). From the evolutionary point of view these results make a lot of sense since the time-scale of life (Hedges and Kumar 2009, Hedges et al. 2015) reveals two large expansions of the species precisely around the vertebrate's time

(between 600 and 400 MYA) and around the time of the mammal's appearance (between 250 and 100 MYA). These expansions are well reflected in our time-scale profile (**Fig. 4-R-8**). Finally, it is important to indicate that the age mapping presented in our study only considers human protein-coding genes that are included in orthologous families (mapping a total of 17437) and, thus, it has a lower coverage over human genes than the other reported studies which include more than 22000 genes in each case (Domazet-Loso and Tautz 2008, Neme and Tautz 2013).



Assignment of Lowest Common Ancestor (LCA) of human genes	DomazetLoso_2008	NemeTautz_2013	Lopesetal_2016
phylogenetic_clades (phylostratum)	number_g_ensembl	number_g_ensembl	number_g_ensembl
Life before - **cellular organisms**	8,285	7,309	1,178
cellular organisms - **Eukaryota**	5,633	6,092	2,178
Eukaryota - **Opisthokonta**	600	750	791
Opisthokonta - **Metazoa**	703	1,918	604
Metazoa - **Bilateria**	2,875	2,337	2,012
Bilateria - **Deuterostomia**	93	128	57
Deuterostomia - **Chordata**	244	162	264
Chordata - **Euteleostomi**	1,247	1,531	5,070
Euteleostomi - **Tetrapoda**	106	161	693
Tetrapoda - **Amniota**	157	197	1,053
Amniota - **Mammalia**	335	292	207
Mammalia - **Eutheria**	830	448	2,172
Eutheria - **Boreoeutheria**	285	189	500
Boreoeutheria - **Euarchontoglires**	124	57	110
Euarchontoglires - **Primates**	1,328	583	548
	22,845	22,154	17,437

Figure 4-R-11. Comparison of different studies on the evolutionary origin of human genes. At left, the plot represents the same data included in the table (right panel) and both show a comparison of the assignment of the human protein-coding genes to the Lowest Common Ancestor (LCA) in phylogenetic clades of the evolutionary tree. The assignments were allocated to 15 phylostratums to allow the comparison of the data. Different datasets including our data correspond to different colours.

5. Functional enrichment of the genes at different evolutionary hallmarks

To improve the interpretation of the genes behind these *hallmarks*, we performed functional enrichment analyses of the sets of protein-coding genes included in each one of eight major stages found in the evolutionary study. The full results of these analyses are provided as tables within Additional file 5, available to download from original publication repository (http://bioinfow.dep.usal.es/evolutionaryhallmarks/). In all the stages, the functional enrichment makes clear biological sense and provides a strong support to the allocation of many biological processes in evolutionary time. We briefly comment and discuss some interesting functions enriched in each stage.

Stage-1, from the origin of life to first cellular organisms.

This stage comprises the genes occurring over two major domains of life: *Archaea* and *Bacteria*. Determining the LCA, our data shows that human has 6.76 % (1178) of the protein-coding genes assigned to Prokaryotic age. *Prokaryotes* are organisms that lack both membrane-bound organelles and nucleus. Functional enrichment analysis showed that this stage involved many basic metabolic processes like glycolysis (GO:0006007, glucose catabolic process), the Krebs cycle (GO:0006099, tricarboxylic acid cycle), and lipid oxidation (GO:0009062, fatty acid catabolic process). The enrichment also shows the appearance of the oldest cellular organelle, the mitochondria, and the oldest macromolecular machine, the ribosome, that are well reported to be dated to Prokaryotic times.

Stage-2, Cellular organisms to Eukaryota.

According to basic literature, the defining feature of eukaryotic cells is that they have membrane-bound organelles, especially the nucleus, which contains the genetic material, and is enclosed by the nuclear envelope. Protists, fungi, animals, and plants all consist of eukaryotic cells. Eukaryotic cells also contain other membrane-bound organelles such as the Golgi apparatus. Eukaryotic organisms can be unicellular or multicellular. The functional enrichment analysis for the 2178 genes that emerged along this stage showed well the formation of the principal complexes expected in *Eukaryotes*. It is noteworthy that the enrichment on nuclear pore proteins, nuclear import proteins, nucleosome and chromatin proteins, as well as many proteins involved DNA and RNA activity: mRNA and rRNA processing, mRNA splicing, DNA unwinding, DNA polymerase, DNA/RNA helicase. This stage also marks in time the appearance and biogenesis of some major molecular complexes: the proteasome (GO:0005839, proteasome core complex), the spliceosome, and the ribosome (at this stage mainly the proteins of the large subunit RPLs, in contrast to the ribosomal proteins of the small subunit RPSs, that were mostly allocated to Prokaryotic age).

Stage-3, Eukaryota to Metazoa.

The third stage comprises organisms from *Opisthokonta* and *Metazoan* clades with 1395 protein-coding genes (27.25 % cumulative). The *Opisthokonts* are a broad group of eukaryotes, including both the animal and fungi kingdoms, sometimes referred to as the

fungi/metazoan group (Parfrey et al. 2006). This stage comprises metazoan, fungal and protistan taxa, and other multicellular taxa (such as plants, or red and brown algae) (Medina et al. 2004). They also include known fungi and/or parasites of plants like *Ascomycota, Basidiomycota, Chytridiomycetes, Glomeromycota, Microsporidia, Urediniomycetes, Ustilaginomycetes* and *Zygomycota* (Adl et al. 2005). According to our functional enrichment analysis, this stage involves different genes responsible for signal transduction like the GTPases. Some of the enriched terms are pyrophosphatase activity, nucleoside-triphosphatase activity, GTP binding, transferring phosphorus-containing groups. All these functions indicate that it may be the time when phosphorus and phospate acquired a key role in protein function and regulation. Other enriched functions, like post-translational protein modification, calcium- binding EF-hand, protein transport and localization also indicate cellular protein regulation.

Stage-4, Metazoa to Vertebrata.

This stage includes organisms from *Eumetazoa, Bilateria, Deuterostomia, Chordata, Craniata*, and *Vertebrata* with 2333 genes (40.63 % cumulative). The main novelties of this stage are the appearance of protein kinase activity, and the presence of growth factors and some specific signalling proteins like WNT. All biochemically characterized members of the WNT superfamily encode enzymes that transfer organic acids, typically fatty acids, onto hydroxyl groups of membrane-embedded targets (Hofmann 2000). Other enriched terms in this stage, like sarcomere and contractile fibber part, may indicate the emergence of the muscular structures present in vertebrates (Neyt et al. 2000).

Stage-5, Vertebrata to Euteleostomi.

This stage represents the largest step in the human lineage according to the number of protein-coding genes assigned (5070), that correspond to a 29 % of the total. The stage comprises organisms from *Gnathostomata* (jawed vertebrates), *Teleostomi* (bony fish and tetrapods) and *Euteleostomi* (bony vertebrate) (Zhu et al. 2013). The enrichment analysis shows a large functional expansion including new biological systems, like the neural and the vascular-circulatory systems, represented in enriched terms like neurogenesis, neuron differentiation, axogenesis, voltage-gated channels, neuromuscular junction development, blood vessel development, vasculature development, mesenchymal cell development and differentiation, etc. Many other genes are assigned to biological regulation and regulation of cellular processes, including cell death and apoptosis. Finally,

the appearance of the large family of homeobox proteins seems to be placed at this stage.

Stage-6, Euteleostomi to Mammalia.

In this stage, there are organisms from *Sarcopterygii* (lobe-finned fishes) (Coates 2009), *Dipnotetrapodomorpha* (new taxon from NCBI comprising lungfishes), *Tetrapoda* (four-legged vertebrates), *Amniota* (comprising the reptiles, birds and mammals that lay their eggs on land or retain the fertilized egg within the mother) and up to *Mammalia* clades. With 1953 genes at this stage, the human lineage achieves 80 % of its gene composition. The most relevant enriched terms are related to the hematologic system, marking the appearance of the leukocytes and the lymphocytes. Previous phylogenetic analyses based on gene expression data also placed the date of many proteins from leukocytes around the time of the mammals' clade (Hughes and Friedman 2009).

Stage-7, Mammalia to Primates.

This stage comprises clades of *Theria, Eutheria, Boreoeutheria, Euarchontoglires,* and *Primates*, representing organisms that give birth to live young without using a shelled egg up to placental mammals (Myers et al. 2006). There are 2821 genes emerged on this stage, adding up to 97.08 % of the cumulative profile in the human gene lineage. A large number of these genes is enriched in the terms *regulation of gene expression* and *transcription*. Other more specific terms are related to the skin (epidermal and epithelial cell differentiation, keratinization) or with the sexual reproductive system (male gamete generation, spermatogenesis and sexual reproduction). This stage also includes a family of cytochrome P450 proteins (that are around 23) and the mammalian defensins (that are 6): DEFA1B, DEFA3, DEFA4, DEFA5, DEFA6, and DEFB4A. Defensins are a family of antimicrobial peptides and vital contributors to host immune response. Being constitutive or inducible expressed genes, they have been shown to contribute to innate host defence via direct bactericidal activity, as well as to adaptive immunity through effector and regulatory functions (Dhople et al. 2006).

Stage-8, Primates to Homo sapiens.

The last stage of human development, with 509 genes, presents a group of quite specific functions played by specific protein families, such as somatotropin hormone, cytochrome P450, GTPase activator activity, defence response to fungus and bacterium provided by histatins. HIS1 and HIS3 (histatin proteins) have been found only in saliva of

humans, macaques and some other primates but not in any other mammals (Sabatini et al. 1993). They are a family of histidine-rich polypeptides that probably function as part of the non-immune host defence system and appeared very late in evolution (Sabatini et al. 1993). Cytochromes P450 constitute a superfamily of proteins that existed in virtually all species from prokaryotes to humans. Most of these proteins in the CYP1, CYP2, CYP3 and CYP4 families encode enzymes involved in the metabolism and elimination of potential toxic compounds like drugs or foreign xenobiotics, and are inducible by various environmental stimuli (Nebert et al. 2013). This last stage includes a small subset of six cytochromes P450 that seem to be very specific of the primates-human clades: CYP2A7, CYP2C9, CYP2D6, CYP2J2, CYP2S1 and CYP3A43. The appearance late in evolution of some of these genes may reflect their functional specificity and it is known that they play a key role in human health (Nebert et al. 2013).

6. Network analysis reveals evolutionary age conservation of co-expressed proteins

The global co-expression analysis of the human protein-coding genes allowed the construction of a network including highly correlated protein pairs. The integration of these data with the data from the evolutionary analysis –that provided the identification of the eight stages along evolutionary time– did allow mapping the age of the genes on the network according to such stages. These results are presented in **Figure 4-R-12** that shows a complex network –like a galaxy– involving 1691 human protein nodes associated by 19615 interactions (Spearman's correlation across human tissues from Human Protein Atlas dataset).

This network corresponds to a subset of the co-expression data, build as indicated above, which included 2298 proteins and 20005 interactions (the full co-expression data is provided in Additional file 2 and the network in Additional file 3 of the original supplementary files: http://bioinfow.dep.usal.es/evolutionaryhallmarks/). The subset is done with only the groups that had at least five linked proteins since we wanted to provide in **Figure 4-R-9** a visible representation of the network with a clear colour mapping of the eight stages. The colours of the stages are also presented in the illustrated table (**Fig. 4-R-6**) to allow better identification of the number and % of proteins at each age hallmark. The

analysis of the network (**Fig. 4-R-12**) done with the algorithm MCODE revealed the existence of 11 major subnetworks, which can be considered as major constellations in the galaxy of relational nodes. The colour legend indicates that there is enrichment in certain colours in each subnetwork. To prove this further, we built a graphic representation for each one of the 11 subnetworks found to show the proportion of proteins assigned to each of the eight evolutionary stages with their corresponding colour code (colours as in **Figure 4-R-9** and **12**).

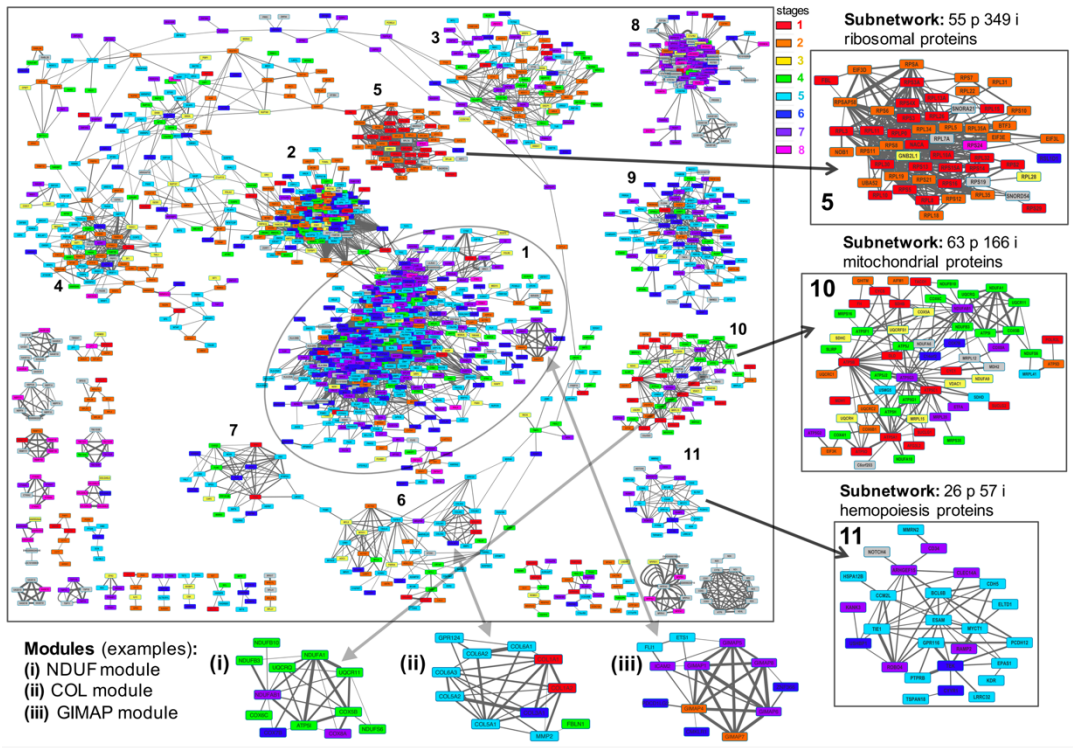


Figure 4-R-12. Human co-expression network mapping the evolutionary age on highly correlated nodes. Representation of the co-expression complex network –like a galaxy– that includes 1691 protein nodes related with 19615 interactions. The colour mapping of the nodes corresponds to the eight stages that were identified in the evolutionary study (as reflected in the labels included at the top right). The network also includes numbers for 11 major subnetworks –clusters of closely related proteins that include more than 20 nodes– considered as major constellations in the galaxy of nodes. Three panels on the right show an enlarged view of three subnetworks corresponding to ribosomal proteins (5), mitochondrial proteins (10) and angiogenesis proteins (11)

This graphic is presented as **Figure 4-R-13**, which shows each subnetwork with its specific colour pattern, indicating that there are always some predominant colours: subnetwork 5 is the oldest with red predominant colours and subnetwork 11 is the newest with blue predominant colours. As a conclusion, these results revealed that in the groups of

highly co-expressed proteins there is a tendency to include proteins of the same evolutionary age.

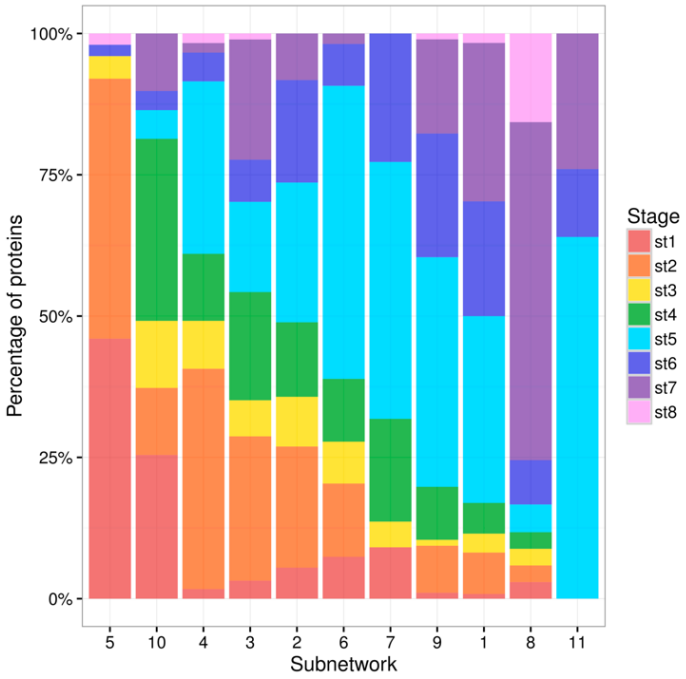


Figure 4-R-13. The relative composition on proteins from different ages in the subnetworks found in the human co-expression network. Graphic plot representing, for each one of the 11 subnetworks found in the co-expression network, the proportion of proteins assigned to each of the 8 evolutionary stages. The stages are marked with their corresponding colour code indicated.

Finally, we did a functional enrichment analysis of the proteins forming the 11 subnetworks which again showed a coherent biological enrichment in specific functions: (subnetwork 1) immune response; (2) cell cycle; (3) cytoskeleton; (4) RNA splicing; (5) ribosome; (6) extracellular matrix; (7) muscle and contraction; (8) gametes and reproductive process; (9) cell junction and cell adhesion; (10) mitochondria and ATP synthesis; (11) angiogenesis and vasculogenesis. More detailed results for this analysis are presented in the illustrated table in **Figure 4-R-14**. Thus, we observed that age-related proteins are predisposed to present expression co-regulation and to have close functional links.

Combining age data, functional data and co-expression data can provide a deeper view about the links and roles of the human protein-coding genes. We observed, for example, that subnetwork 5 (which contains proteins related with ribosome and translation) presented, as expected, an overwhelming majority of ancient genes from the *Prokaryotic* or

Eukaryotic age (stages 1 and 2). On the other hand, subnetworks 1 (immune response, leukocyte/lymphocyte activations), 8 (gametes and reproductive process) and 11 (angiogenesis and vasculogenesis) showed a higher proportion of recent genes dated after *Vertebrata* (**Fig. 4-R-14**). These results agree with studies based on yeast protein physical interaction networks, arguing that proteins preferentially interact with proteins of same age and origin (Capra et al. 2010). Moreover, it was previously shown that co-expression networks can be conserved over the evolutionary history, and these genes tend to be functionally related and provide selective advantages (Stuart et al. 2003). It has been also reported that co-expression networks are found associated with functions like cell adhesion, cell cycle, DNA replication and DNA repair (Monaco et al. 2015), and this is in agreement with functions found enriched in subnetworks of our analyses: subnetwork 2 and 9 (**Fig. 4-R-13**).

Functional enrichment: Gene-Ontology (GO) terms	P-value	Z-score	Combined Score	subnetworks
leukocyte activation (GO:0045321)	5.39E-49	-2.37	248.26	**c1: 517p 7364i**
lymphocyte activation (GO:0046649)	6.01E-42	-2.32	205.61	
activation of immune response (GO:0002253)	1.26E-41	-3.45	303.74	
antigen binding (GO:0003823)	9.93E-10	-4.70	73.40	
cytokine receptor activity (GO:0004896)	2.31E-13	-2.40	54.84	
mitotic cell cycle (GO:0000278)	3.72E-76	-2.30	383.91	**c2: 194p 5919i**
nuclear division (GO:0000280)	2.10E-49	-2.31	244.42	
mitotic nuclear division (GO:0007067)	7.61E-45	-2.27	217.77	
regulation of cell cycle process (GO:0010564)	1.69E-32	-2.45	165.68	
cellular component assembly involved in morphogenesis (GO:0010927)	2.46E-30	-2.34	147.20	**c3: 98p 210i**
microtubule-based process (GO:0007017)	1.64E-15	-2.42	71.37	
microtubule-based movement (GO:0007018)	1.45E-14	-2.29	63.09	
cytoskeleton-dependent intracellular transport (GO:0030705)	3.40E-10	-2.24	39.74	
tubulin binding (GO:0015631)	2.28E-05	-2.43	14.73	
RNA splicing (GO:0008380)	3.02E-29	-2.34	139.84	**c4: 69p 199i**
mRNA processing (GO:0006397)	1.01E-26	-2.39	130.33	
mRNA splicing, via spliceosome (GO:0000398)	2.22E-21	-2.22	95.33	
regulation of RNA splicing (GO:0043484)	3.86E-10	-2.13	37.44	
translational initiation (GO:0006413)	1.98E-79	-2.18	383.77	**c5: 55p 349i**
translation (GO:0006412)	1.18E-70	-2.33	366.97	
ribosomal subunit (GO:0044391)	1.78E-77	-2.11	365.26	
ribosome (GO:0005840)	5.20E-53	-2.23	262.17	
extracellular matrix organization (GO:0030198)	1.49E-19	-2.38	89.10	**c6: 53p 117i**
extracellular structure organization (GO:0043062)	1.57E-19	-2.38	89.09	
collagen metabolic process (GO:0032963)	3.82E-18	-2.17	75.29	
muscle system process (GO:0003012)	2.83E-11	-2.32	42.53	**c7: 22p 59i**
regulation of muscle contraction (GO:0006937)	1.52E-08	-2.26	29.65	
regulation of heart contraction (GO:0008016)	4.92E-08	-2.29	28.64	
regulation of muscle system process (GO:0090257)	5.78E-08	-2.29	28.57	
spermatogenesis (GO:0007283)	2.46E-08	-2.52	32.43	**c8: 149p 4343i**
male gamete generation (GO:0048232)	2.53E-08	-2.52	32.45	
multicellular organismal reproductive process (GO:0048609)	4.55E-08	-2.53	32.19	
gamete generation (GO:0007276)	6.15E-08	-2.51	31.83	
cell-cell junction organization (GO:0045216)	1.79E-06	-2.26	15.55	**c9: 102p 294i**
tight junction assembly (GO:0070830)	2.42E-06	-2.22	15.29	
cell junction organization (GO:0034330)	5.6E-06	-2.28	15.16	
cell-cell junction assembly (GO:0007043)	3.33E-06	-2.15	14.79	
energy coupled proton transport, down electrochemical gradient (GO:0015985)	5.02E-33	-2.89	202.94	**c10: 63p 166i**
ATP synthesis coupled proton transport (GO:0015986)	5.02E-33	-2.89	202.64	
mitochondrial ATP synthesis coupled proton transport (GO:0042776)	4.51E-28	-2.93	175.66	
mitochondrial proton-transporting ATP synthase complex (GO:0005753)	1.27E-31	-2.95	202.82	
angiogenesis (GO:0001525)	2.40E-12	-2.31	47.98	**c11: 26p 57i**
regulation of angiogenesis (GO:0045765)	4.79E-08	-2.30	26.54	
regulation of vasculature development (GO:1901342)	8.32E-08	-2.32	26.39	
regulation of vasculogenesis (GO:2001212)	9.43E-07	-2.68	26.34	

Subnetwork 1: 517 p 7364 i
immune response
leukocyte/lymphocity activation

Subnetwork 2: 194 p 5919 i
cell cycle, cell division

Subnetwork 3: 98 p 210 i
cytoskeleton, microtubules

Subnetwork 4: 69 p 199 i
RNA splicing, mRNA processing

Subnetwork 5: 55 p 349 i
ribosome, translation

Subnetwork 6: 53 p 117 i
extracellular matrix, collagen

Subnetwork 7: 22 p 59 i
muscle system, contraction

Subnetwork 8: 149 p 4343 i
gametes, reproductive process

Subnetwork 9: 102 p 294 i
cell-cell junction, cell adhesion

Subnetwork 10: 63 p 166 i
mitochondria, ATP synthesis

Subnetwork 11: 26 p 57 i
angiogenesis, vasculogenesis

Figure 4-R-14. Human co-expression network: functional enrichment of major subnetworks. Illustrated table showing a summary of the results from the functional enrichment analyses done with the proteins included in each of the 11 subnetworks labelled at the right and included in the network provided in Figure 4-R-8. The number of proteins (p) and interactions (i) that each subnetwork includes are also indicated.

7. Network analysis reveals tissue-specific clusters

Since we successfully mapped evolutionary information over nodes of the co-expression network, we investigated how clusters of replicates-tissues obtained using all genes in our RNA-seq dataset (18545 genes) are conserved by gene relationships within a co-expression network. Thus, we enlarged our previous network (**Fig. 4-R-12**) through filtering all gene correlations ($r \geq 0.75$; repeats = 100) in order to improve the visualization of tissue-specific clusters within the co-expression network.

As we can observe in **Figure 4-R-15**, the new co-expression network is a highly dense network, now composed of 8634 nodes and 131197 edges. We mapped an additional layer of information on nodes about the number of tissues where each gene is expressed (mean(FPKM)>1), corresponding to previous analysis (**Fig. 4-R-5**). We effortlessly identified several clusters of tissue-specific genes within our network, which correspond to main tissue-relationships previously highlighted by replicates' co-expression heatmap using not only all expression dataset (**Fig. 4-R-3**) but also tissue-specific genes (**Fig. 4-R-7**). For example, a highly specific cluster associated with gene expression of testis was found. It was expected because this RNA-seq dataset from Human Protein Atlas revealed a great differential expression signal from testis, also detected in our analysis for clustering samples (**Fig. 4-R-3** and **4**) and in our analysis of tissue-specific genes' frequency (**Fig. 4-R-7**). Doubtless, testis was also reported as one of the most divergent tissues in terms of gene expression and protein concentration (Kosti et al. 2016).

Interestingly, more close groups of tissues appeared as gene co-expression clusters of no tissue-specific genes (orange-like nodes) like digestive system, immune system, bone marrow or brain (**Fig. 4-R-15**). This result supports our previous idea that common restriction of the tissue-specific definition to a unique tissue could be wrong since many tissues are quite similar. Genes/proteins may vary slightly their expression along these related tissues, hindering our classification into tissue-specific or house-keeping when a selective threshold is applied. In fact, it has been well-described how multifunctional proteins interact with related proteins from different biological functions (Chapple et al. 2015), whose gene expression may correspond to these subtle expression differences.

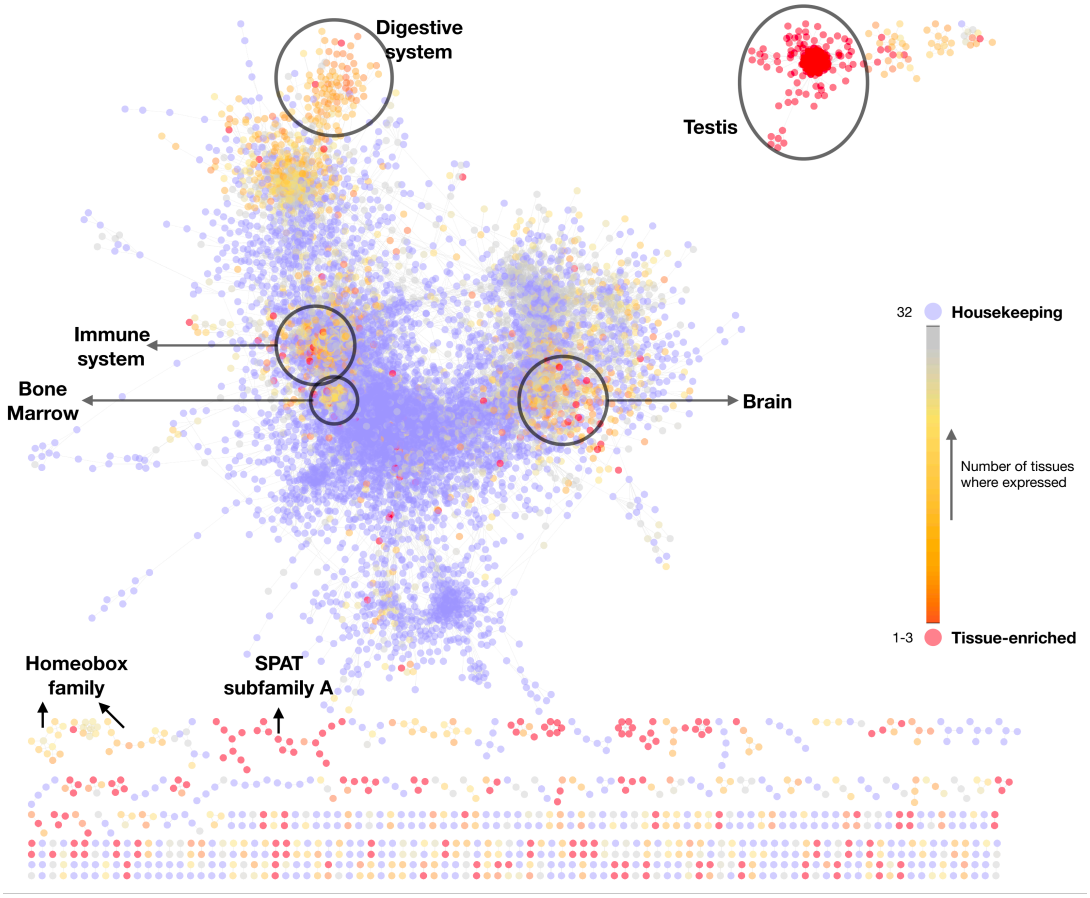


Figure 4-R-15. Large co-expression network of 8634 nodes and 131197 edges, obtained by filtering all Spearman's correlation lower than 0.75 (repeats = 100). A tissue-specific expression layer has been mapped over nodes (right legend), highlighting several tissue-enriched subnetworks. Prefuse force directed layout was used to represent the whole network.

We also found relationships among very similar genes within this large co-expression network, like homeobox family or SPAT subfamily A (**Fig. 4-R-15**). This also suggested to us that a well-managed co-expression network may be an excellent approach to elucidate both global and specific relationships, even across gene families or tissue functions.

CONCLUSIONS

Throughout Chapter IV, we presented a **transcriptomic analysis of the human gene expression profiles along 32 tissues from the Human Protein Atlas project**. This in-depth analysis yielded a global mapping of the activity of most human genes and of the links between them, revealing the expected association of samples from common physiological regions: the digestive system (stomach, duodenum, small intestine, colon and rectum), the hematopoietic and lymphatic system (bone marrow, lymph node, spleen, tonsil and appendix), the muscle (cardiac and skeletal muscle), the brain or the testis.

Interestingly, we conducted an evolutionary study of the human protein-coding genes, placing them in the time-scale of the living species and revealing eight distinct hallmarks along such time-scale, showing that the housekeeping (HK) genes are more ancient than the tissue-enriched (TE) genes. As demonstrated above, the HK genes present significant emergence in stage 2 of the evolutionary profile, while the TE genes have the major appearance in stage 7. The functional enrichment study found coherent groups of terms and annotations assigned to the genes placed at each evolutionary stage. For example, in stage 1 there were many functional terms on essential metabolic processes, like aerobic respiration and mitochondrial activity; and in stage 2 there were enriched functions related to the nucleus and genome regulation, like chromatin and nucleosome assembly, DNA replication or mRNA processing.

Finally, the study of the pair-wise correlation of the gene expression profiles along tissues allowed building human gene co-expression networks and find modules with functional and biological meaning. The mapping of the age of the protein-coding genes on these networks demonstrated the existence of tight links between age-related proteins, while the inclusion of tissue-enriched information denoted how reliable is a gene co-expression network for reflecting both global and specific biological relationships.

GENERAL CONCLUSIONS

Along the four chapters of this PhD dissertation, we have proposed and detailed Bioinformatics algorithms, methods or frameworks to approach major issues in current omic data analyses and interpretation of results. As general conclusions of this PhD, we can conclude the following statements:

1. The proposed method **DECO is able to deal with homo- and heterogeneous omic datasets**, extracting all the relevant intra-variability through an exhaustive differential analysis designed as a resampling of samples (**RDA**), even when classical methods for differential analysis did not find significant signal. Thus, a proper feature's categorization into a four model-types, including *outlier* omic profiles, was proposed to improve the interpretation of the results.

2. **DECO outperforms current methods for *outlier* profile detection** on large and small omic datasets, providing a logical importance at scoring (*Standard Chi-Square*) significant features, decreasing from *complete changes* to *mixed changes*.

3. The ***h-statistic*** **proposed integrates both omic data dispersion and predictor-response information** (NSCA), providing a comprehensive lecture of the existing dependent structures among samples and biological features. We have demonstrated that it greatly enhances the stratification of samples, allows to disclose hidden subgroup of samples respect to the initial categories and highlights significant associations among samples and features.

4. The ***deco*** **R package comprises all steps of DECO method** in a simple and friendly-user protocol, consisting of three R functions: RDA, NSCA and a graphical report (PDF file). A full detailed vignette describing this protocol has been written for an easy handling of the user.

5. The ***cohesiveness*** **statistic has been demonstrated as a simple and non-parametric measurement** to precisely select stable and differential features when phenotypical data is available. We have demonstrated that a simple assessment of the probability of "being close" is similar to the *one VS all* design. It allows to disclose stable patterns along samples even when there are no differences among averages, which may correspond to tissue-specific patterns or differential expression levels.

6. Since it is a non-parametric statistic, ***cohesiveness*** **can be applied to any type of omic data** to assess the ability of any biological feature for discriminating among categories of samples or conditions.

7. As a result of the collaboration with Marc Vidal's laboratory, we proposed a **new categorization of PSI-MI methods into meta-groups to avoid mismatches of manual curation** done by different primary protein-protein interaction databases. Further investigations need be done about this primary issue for the creation of literature-based protein-protein interactomes.

8. The proposed **integration of protein-protein interaction networks with subcellular localization data establishes a good basis for testing new prediction subcellular localization** approaches. Cross-talk among subcellular compartments from HI-III significantly matches with the most connected spaces according to Cell Atlas information, leading to new potential subcellular localizations for those proteins which should move to interact each other (shuttling proteins).

9. The described **integration of human co-expression network (Human Protein Atlas), orthologous identification (OMA through LCA) and evolutionary timeline (TimeTree) provides a successful framework to identify more conserved and correlated group/families** of human protein-coding genes. This type of integration was intended of being extrapolated to other organisms or particular biological pathways and networks.

BIBLIOGRAPHY

Abrusan G (2013). **Integration of new genes into cellular networks, and their structural maturation**. *Genetics* **195**(4): 1407-1417. 10.1534/genetics.113.152256

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, ... et al. (1991). **Complementary DNA sequencing: expressed sequence tags and human genome project**. *Science* **252**(5013): 1651-1656.

Adl SM, Simpson AG, Farmer MA, Andersen RA, Anderson OR, Barta JR, ... Taylor MF (2005). **The new higher level classification of eukaryotes with emphasis on the taxonomy of protists**. *J Eukaryot Microbiol* **52**(5): 399-451. 10.1111/j.1550-7408.2005.00053.x

Aibar S, Abaigar M, Campos-Laborie FJ, Sanchez-Santos JM, Hernandez-Rivas JM and De Las Rivas J (2016). **Identification of expression patterns in the progression of disease stages by integration of transcriptomic data**. *BMC Bioinformatics* **17**(Suppl 15): 432. 10.1186/s12859-016-1290-4

Alba MM and Castresana J (2005). **Inverse relationship between evolutionary rate and age of mammalian genes**. *Mol Biol Evol* **22**(3): 598-606. 10.1093/molbev/msi045

Allott EH, Geradts J, Sun X, Cohen SM, Zirpoli GR, Khoury T, ... Troester MA (2016). **Intratumoral heterogeneity as a source of discordance in breast cancer biomarker classification**. *Breast Cancer Res* **18**(1): 68. 10.1186/s13058-016-0725-1

Almagro Armenteros JJ, Sonderby CK, Sonderby SK, Nielsen H and Winther O (2017). **DeepLoc: prediction of protein subcellular localization using deep learning**. *Bioinformatics* **33**(21): 3387-3395. 10.1093/bioinformatics/btx431

Alonso-Lopez D, Gutierrez MA, Lopes KP, Prieto C, Santamaria R and De Las Rivas J (2016). **APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks**. *Nucleic Acids Res* **44**(W1): W529-535. 10.1093/nar/gkw363

Altaf-Ul-Amin M, Afendi FM, Kiboi SK and Kanaya S (2014). **Systems biology in the context of big data and networks**. *Biomed Res Int* **2014**: 428570. 10.1155/2014/428570

Altenhoff AM, Gil M, Gonnet GH and Dessimoz C (2013). **Inferring hierarchical orthologous groups from orthologous gene pairs**. *PLoS One* **8**(1): e53786. 10.1371/journal.pone.0053786

Altenhoff AM, Skunca N, Glover N, Train CM, Sueki A, Pilizota I, ... Dessimoz C (2015). **The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements**. *Nucleic Acids Res* **43**(Database issue): D240-249. 10.1093/nar/gku1158

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990). **Basic local alignment search tool**. *J Mol Biol* **215**(3): 403-410. 10.1016/S0022-2836(05)80360-2

Anders S and Huber W (2010). **Differential expression analysis for sequence count data**. *Genome Biol* **11**(10): R106. 10.1186/gb-2010-11-10-r106

Ang JC, Mirzal A, Haron H and Hamed HN (2016). **Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection**. *IEEE/ACM Trans Comput Biol Bioinform* **13**(5): 971-989. 10.1109/TCBB.2015.2478454

Ashley EA (2016). **Towards precision medicine**. *Nat Rev Genet* **17**(9): 507-522. 10.1038/nrg.2016.86

Babu GJ (1992). **Subsample and half-sample methods**. *Annals of the Institute of Statistical Mathematics* **44**(4): 703-720. 10.1007/bf00053399

Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T and Hogue CW (2001). **BIND--The Biomolecular Interaction Network Database**. *Nucleic Acids Res* **29**(1): 242-245.

Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, ... Jones SJ (2006). **Analysis of the**

prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**: 246. 10.1186/1471-2164-7-246

Baldi P and Long AD (2001). **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes**. *Bioinformatics* **17**(6): 509-519.

Barabasi AL, Gulbahce N and Loscalzo J (2011). **Network medicine: a network-based approach to human disease**. *Nat Rev Genet* **12**(1): 56-68. 10.1038/nrg2918

Barabasi AL and Oltvai ZN (2004). **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* **5**(2): 101-113. 10.1038/nrg1272

Barbulovic-Nad I, Lucente M, Sun Y, Zhang M, Wheeler AR and Bussmann M (2006). **Bio-microarray fabrication techniques--a review**. *Crit Rev Biotechnol* **26**(4): 237-259. 10.1080/07388550600978358

Basken J, Stuart SA, Kavran AJ, Lee T, Ebmeier CC, Old WM and Ahn NG (2018). **Specificity of Phosphorylation Responses to Mitogen Activated Protein (MAP) Kinase Pathway Inhibitors in Melanoma Cells**. *Mol Cell Proteomics* **17**(4): 550-564. 10.1074/mcp.RA117.000335

Baty F, Jaeger D, Preiswerk F, Schumacher MM and Brutsche MH (2008). **Stability of gene contributions and identification of outliers in multivariate analysis of microarray data**. *BMC Bioinformatics* **9**: 289. 10.1186/1471-2105-9-289

Beckman RA, Schemmann GS and Yeang CH (2012). **Impact of genetic dynamics and single-cell heterogeneity on development of nonstandard personalized medicine strategies for cancer**. *Proc Natl Acad Sci U S A* **109**(36): 14586-14591. 10.1073/pnas.1203559109

Bedard PL, Hansen AR, Ratain MJ and Siu LL (2013). **Tumour heterogeneity in the clinic**. *Nature* **501**(7467): 355-364. 10.1038/nature12627

Beh EJ and Lombardo R (2014). **Correspondence Analysis: Theory, Practice and New Strategies**, Wiley.9781118762905

Beh J and D'Ambra L (2010). **Non-symmetrical correspondence analysis with concatenation and linear constraints**. *Australian & New Zealand Journal of Statistics* **52**(1): 27-44. doi:10.1111/j.1467-842X.2009.00564.x

Benjamini Y and Hochberg Y (1995). **Controlling the false discovery rate: a practical and powerful approach to multiple testing**.0035-9246

Benjamini Y and Hochberg Y (1995). **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1): 289-300.

Berger B, Peng J and Singh M (2013). **Computational solutions for omics data**. *Nat Rev Genet* **14**(5): 333-346. 10.1038/nrg3433

Berggard T, Linse S and James P (2007). **Methods for the detection and analysis of protein-protein interactions**. *Proteomics* **7**(16): 2833-2842. 10.1002/pmic.200700131

Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C, Rudan I, Campbell H, ... Haley CS (2015). **Application of high-dimensional feature selection: evaluation for genomic prediction in man**. *Sci Rep* **5**: 10312. 10.1038/srep10312

Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G and Milanesi L (2016). **Methods for the integration of multi-omics data: mathematical aspects**. *BMC Bioinformatics* **17 Suppl 2**: 15. 10.1186/s12859-015-0857-9

Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, ... Meyerson M (2001). **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses**. *Proc Natl Acad Sci U S A* **98**(24): 13790-13795. 10.1073/pnas.191502998

Bianchini G, Balko JM, Mayer IA, Sanders ME and Gianni L (2016). **Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease**. *Nat Rev Clin Oncol* **13**(11): 674-690. 10.1038/nrclinonc.2016.66

Billingsley P (1995). **Probability and measure**. *A Wiley-Interscience Publication, John Wiley*.

Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R and Jensen LJ (2014). **COMPARTMENTS: unification and visualization of protein subcellular localization evidence**. *Database (Oxford)* **2014**: bau012. 10.1093/database/bau012

Bjorklund AK, Light S, Hedin L and Elofsson A (2008). **Quantitative assessment of the structural bias in protein-protein interaction**

assays. *Proteomics* **8**(22): 4657-4667. 10.1002/pmic.200800150

Bogachev MI, Kayumov AR, Markelov OA and Bunde A (2016). **Statistical prediction of protein structural, localization and functional properties by the analysis of its fragment mass distributions after proteolytic cleavage**. *Sci Rep* **6**: 22286. 10.1038/srep22286

Bolón-Canedo V, Sánchez-Maroño N and Alonso-Betanzos A (2015). **Recent advances and emerging challenges of feature selection in the context of big data**. *Knowledge-Based Systems* **86**: 33-45. doi.org/10.1016/j.knosys.2015.05.014

Bradley PS and Mangasarian OL (1998). **Feature selection via concave minimization and support vector machines**. *ICML*.

Breiman L (2001). **Random forests**. *Machine learning* **45**(1): 5-32.

Breitling R, Armengaud P, Amtmann A and Herzyk P (2004). **Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments**. *FEBS Lett* **573**(1-3): 83-92. 10.1016/j.febslet.2004.07.055

Bruckner A, Polge C, Lentze N, Auerbach D and Schlattner U (2009). **Yeast two-hybrid, a powerful tool for systems biology**. *Int J Mol Sci* **10**(6): 2763-2788. 10.3390/ijms10062763

Bullard JH, Purdom E, Hansen KD and Dudoit S (2010). **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments**. *BMC Bioinformatics* **11**: 94. 10.1186/1471-2105-11-94

Cai JJ and Petrov DA (2010). **Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes**. *Genome Biol Evol* **2**: 393-409. 10.1093/gbe/evq019

Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, ... Stuart JM (2013). **The Cancer Genome Atlas Pan-Cancer analysis project**. *Nat Genet* **45**(10): 1113-1120. 10.1038/ng.2764

Capra JA, Pollard KS and Singh M (2010). **Novel genes exhibit distinct patterns of function acquisition and network integration**. *Genome Biol* **11**(12): R127. 10.1186/gb-2010-11-12-r127

Capra JA, Stolzer M, Durand D and Pollard KS (2013). **How old is my gene?** *Trends in genetics : TIG* **29**(11): 10.1016/j.tig.2013.1007.1001. 10.1016/j.tig.2013.07.001

Chapple CE, Robisson B, Spinelli L, Guien C, Becker E and Brun C (2015). **Extreme multifunctional proteins identified from a human protein interaction network**. *Nat Commun* **6**: 7412. 10.1038/ncomms8412

Chen F, Moran JT, Zhang Y, Ates KM, Yu D, Schrader LA, ... Hall BJ (2016). **The transcription factor NeuroD2 coordinates synaptic innervation and cell intrinsic properties to control excitability of cortical pyramidal neurons**. *J Physiol* **594**(13): 3729-3744. 10.1113/JP271953

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, ... Snyder M (2012). **Personal omics profiling reveals dynamic molecular and medical phenotypes**. *Cell* **148**(6): 1293-1307. 10.1016/j.cell.2012.02.009

Chi SM and Nam D (2012). **WegoLoc: accurate prediction of protein subcellular localization using weighted Gene Ontology terms**. *Bioinformatics* **28**(7): 1028-1030. 10.1093/bioinformatics/bts062

Chrominski K and Tkacz M (2015). **Comparison of High-Level Microarray Analysis Methods in the Context of Result Consistency**. *PLoS One* **10**(6): e0128845. 10.1371/journal.pone.0128845

Ciavolino E, D'Ambra A, Venuleo C and Vernai M (2017). **Non-symmetrical correspondence analysis to evaluate how age influences the addiction discourses**. *Statistica & Applicazioni* **15**(1): 3-18.

Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, ... Perou CM (2015). **Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer**. *Cell* **163**(2): 506-519. 10.1016/j.cell.2015.09.033

Coates MI (2009). **Palaeontology: beyond the age of fishes**. *Nature* **458**(7237): 413-414. 10.1038/458413a

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, ... Mortazavi A (2016). **A survey of best practices for RNA-seq data analysis**. *Genome Biol* **17**: 13. 10.1186/s13059-016-0881-8

Consortium F, the RP, Clst, Forrest AR, Kawaji H, Rehli M, ... Hayashizaki Y (2014). **A promoter-**

level mammalian expression atlas. *Nature* **507**(7493): 462-470. 10.1038/nature13182

Consortium GT (2013). **The Genotype-Tissue Expression (GTEx) project**. *Nat Genet* **45**(6): 580-585. 10.1038/ng.2653

Cyll K, Ersvaer E, Vlatkovic L, Pradhan M, Kildal W, Avranden Kjaer M, ... Danielsen HE (2017). **Tumour heterogeneity poses a significant challenge to cancer biomarker research**. *Br J Cancer* **117**(3): 367-375. 10.1038/bjc.2017.171

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, ... Meng F (2005). **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucleic Acids Res* **33**(20): e175. 10.1093/nar/gni179

Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, ... Quake SR (2015). **A survey of human brain transcriptome diversity at the single cell level**. *Proceedings of the National Academy of Sciences* **112**(23): 7285-7290.

De Las Rivas J and Fontanillo C (2010). **Protein-protein interactions essentials: key concepts to building and analyzing interactome networks**. *PLoS Comput Biol* **6**(6): e1000807. 10.1371/journal.pcbi.1000807

De Palma M and Hanahan D (2012). **The biology of personalized cancer medicine: facing individual complexities underlying hallmark capabilities**. *Mol Oncol* **6**(2): 111-127. 10.1016/j.molonc.2012.01.011

de Ronde JJ, Rigaill G, Rottenberg S, Rodenhuis S and Wessels LF (2013). **Identifying subgroup markers in heterogeneous populations**. *Nucleic Acids Res* **41**(21): e200. 10.1093/nar/gkt845

Dhople V, Krukemeyer A and Ramamoorthy A (2006). **The human beta-defensin-3, an antibacterial peptide with multiple biological functions**. *Biochim Biophys Acta* **1758**(9): 1499-1512. 10.1016/j.bbamem.2006.07.007

Diday E (1984). **Data analysis and informatics, III: proceedings of the Third International Symposium on Data Analysis and Informatics**, North-Holland.9780444875556

Dimitrakopoulos C, Hindupur SK, Hafliger L, Behr J, Montazeri H, Hall MN and Beerenwinkel N (2018). **Network-based integration of multi-omics data for prioritizing cancer genes**. *Bioinformatics*. 10.1093/bioinformatics/bty148

Dobbin K and Simon R (2005). **Sample size determination in microarray experiments for class comparison and prognostic classification**. *Biostatistics* **6**(1): 27-38. 10.1093/biostatistics/kxh015

Domazet-Loso T, Brajkovic J and Tautz D (2007). **A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages**. *Trends Genet* **23**(11): 533-539. 10.1016/j.tig.2007.08.014

Domazet-Loso T and Tautz D (2008). **An ancient evolutionary origin of genes associated with human genetic diseases**. *Mol Biol Evol* **25**(12): 2699-2707. 10.1093/molbev/msn214

Donnard E, Barbosa-Silva A, Guedes RL, Fernandes GR, Velloso H, Kohn MJ, ... Ortega JM (2011). **Preimplantation development regulatory pathway construction through a text-mining approach**. *BMC Genomics* **12 Suppl 4**: S3. 10.1186/1471-2164-12-S4-S3

Dunn KW, Kamocka MM and McDonald JH (2011). **A practical guide to evaluating colocalization in biological microscopy**. *Am J Physiol Cell Physiol* **300**(4): C723-742. 10.1152/ajpcell.00462.2010

Ebrahim A, Brunk E, Tan J, O'Brien EJ, Kim D, Szubin R, ... Palsson BO (2016). **Multi-omic data integration enables discovery of hidden biological regularities**. *Nat Commun* **7**: 13091. 10.1038/ncomms13091

Eisenberg E and Levanon EY (2013). **Human housekeeping genes, revisited**. *Trends Genet* **29**(10): 569-574. 10.1016/j.tig.2013.05.010

Fields S and Song O (1989). **A novel genetic system to detect protein-protein interactions**. *Nature* **340**(6230): 245-246. 10.1038/340245a0

Fisher RA (1925). **Statistical methods for research workers**. *Edinburgh, London,*, Oliver and Boyd

Fontanillo C, Nogales-Cadenas R, Pascual-Montano A and De las Rivas J (2011). **Functional analysis beyond enrichment: non-redundant reciprocal linkage of genes and biological terms**. *PLoS One* **6**(9): e24289. 10.1371/journal.pone.0024289

Gabaldon T and Koonin EV (2013). **Functional and evolutionary implications of gene orthology**. *Nat Rev Genet* **14**(5): 360-366. 10.1038/nrg3456

Gillies RJ, Verduzco D and Gatenby RA (2012). **Evolutionary dynamics of carcinogenesis and why targeted therapy does not work**. *Nat Rev Cancer* **12**(7): 487-493. 10.1038/nrc3298

Goeman JJ, van de Geer SA, de Kort F and van Houwelingen HC (2004). **A global test for groups of genes: testing association with a clinical outcome**. *Bioinformatics* **20**(1): 93-99.

Goodman LA and Kruskal WH (1959). **Measures of Association for Cross Classifications. II: Further Discussion and References**. *Journal of the American Statistical Association* **54**(285): 123-163. 10.1080/01621459.1959.10501503

Gur-Dedeoglu B, Konu O, Kir S, Ozturk AR, Bozkurt B, Ergul G and Yulug IG (2008). **A resampling-based meta-analysis for detection of differential gene expression in breast cancer**. *BMC Cancer* **8**: 396. 10.1186/1471-2407-8-396

Hakes L, Robertson DL, Oliver SG and Lovell SC (2007). **Protein interactions from complexes: a structural perspective**. *Comp Funct Genomics*: 49356. 10.1155/2007/49356

Hardcastle TJ and Kelly KA (2010). **baySeq: empirical Bayesian methods for identifying differential expression in sequence count data**. *BMC Bioinformatics* **11**: 422. 10.1186/1471-2105-11-422

Hasan MA, Ahmad S and Molla MK (2017). **Protein subcellular localization prediction using multiple kernel learning based support vector machine**. *Mol Biosyst* **13**(4): 785-795. 10.1039/c6mb00860g

Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, ... Symmans WF (2011). **A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer**. *JAMA* **305**(18): 1873-1881. 10.1001/jama.2011.593

Hawrylycz M, Miller JA, Menon V, Feng D, Dolbeare T, Guillozet-Bongaarts AL, ... Lein E (2015). **Canonical genetic signatures of the adult human brain**. *Nat Neurosci* **18**(12): 1832-1844. 10.1038/nn.4171

Hedges SB and Kumar S (2009). **Discovering the timetree of life**. *The TimeTree of life*. S. B. Hedges and S. Kumar. New York, Oxford University Press.

Hedges SB, Marin J, Suleski M, Paymer M and Kumar S (2015). **Tree of life reveals clock-like speciation and diversification**. *Mol Biol Evol* **32**(4): 835-845. 10.1093/molbev/msv037

Heimberg G, Bhatnagar R, El-Samad H and Thomson M (2016). **Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing**. *Cell Syst* **2**(4): 239-250. 10.1016/j.cels.2016.04.001

Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y, ... Mann M (2015). **A human interactome in three quantitative dimensions organized by stoichiometries and abundances**. *Cell* **163**(3): 712-723. 10.1016/j.cell.2015.09.053

Hemmrich G, Khalturin K, Boehm AM, Puchert M, Anton-Erxleben F, Wittlieb J, ... Bosch TC (2012). **Molecular signatures of the three stem cell lineages in hydra and the emergence of stem cell function at the base of multicellularity**. *Mol Biol Evol* **29**(11): 3267-3280. 10.1093/molbev/mss134

Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, ... Apweiler R (2004). **IntAct: an open source molecular interaction database**. *Nucleic Acids Res* **32**(Database issue): D452-455. 10.1093/nar/gkh052

Hesterberg T, Monaghan S, S Moore D, Clipson A, Epstein R, H Freeman W and New York C (2005). **Bootstrap Methods and Permutation Tests**

Hiissa J, Elo LL, Huhtinen K, Perheentupa A, Poutanen M and Aittokallio T (2009). **Resampling reveals sample-level differential expression in clinical genome-wide studies**. *OMICS* **13**(5): 381-396. 10.1089/omi.2009.0027

Himmels P, Paredes I, Adler H, Karakatsani A, Luck R, Marti HH, ... Ruiz de Almodovar C (2017). **Motor neurons control blood vessel patterning in the developing spinal cord**. *Nat Commun* **8**: 14583. 10.1038/ncomms14583

Hira ZM and Gillies DF (2015). **A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data**. *Adv Bioinformatics* **2015**: 198363. 10.1155/2015/198363

Hofmann K (2000). **A superfamily of membrane-bound O-acyltransferases with implications for wnt signaling**. *Trends Biochem Sci* **25**(3): 111-112.

Hogenbirk MA, Heideman MR, de Rink I, Velds A, Kerkhoven RM, Wessels LF and Jacobs H (2016). **Defining chromosomal translocation risks in**

cancer. *Proc Natl Acad Sci U S A* **113**(26): E3649-3656. 10.1073/pnas.1602025113

Holm S (1979). **A Simple Sequentially Rejective Multiple Test Procedure**. *Scandinavian Journal of Statistics* **6**(2): 65-70.

Huang da W, Sherman BT and Lempicki RA (2009). **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* **4**(1): 44-57. 10.1038/nprot.2008.211

Huang H, Li X, Guo Y, Zhang Y, Deng X, Chen L, ... Ao L (2016). **Identifying reproducible cancer-associated highly expressed genes with important functional significances using multiple datasets**. *Sci Rep* **6**: 36227. 10.1038/srep36227

Huang S, Chaudhary K and Garmire LX (2017). **More Is Better: Recent Progress in Multi-Omics Data Integration Methods**. *Front Genet* **8**: 84. 10.3389/fgene.2017.00084

Huang Y, Zhang JL, Yu XL, Xu TS, Wang ZB and Cheng XC (2013). **Molecular functions of small regulatory noncoding RNA**. *Biochemistry (Mosc)* **78**(3): 221-230. 10.1134/S0006297913030024

Hubbell E, Liu WM and Mei R (2002). **Robust estimators for expression analysis**. *Bioinformatics* **18**(12): 1585-1592.

Huber W, von Heydebreck A, Sultmann H, Poustka A and Vingron M (2002). **Variance stabilization applied to microarray data calibration and to the quantification of differential expression**. *Bioinformatics* **18 Suppl 1**: S96-104.

Hubner NC, Bird AW, Cox J, Splettstoesser B, Bandilla P, Poser I, ... Mann M (2010). **Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions**. *J Cell Biol* **189**(4): 739-754. 10.1083/jcb.200911091

Hughes AL and Friedman R (2009). **A phylogenetic approach to gene expression data: evidence for the evolutionary origin of mammalian leukocyte phenotypes**. *Evol Dev* **11**(4): 382-390. 10.1111/j.1525-142X.2009.00345.x

Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, ... Harper JW (2017). **Architecture of the human interactome defines protein communities and disease networks**. *Nature* **545**(7655): 505-509. 10.1038/nature22366

Iorio F, Bernardo-Faura M, Gobbi A, Cokelaer T, Jurman G and Saez-Rodriguez J (2016). **Efficient randomization of biological networks while preserving functional characterization of individual nodes**. *BMC Bioinformatics* **17**(1): 542. 10.1186/s12859-016-1402-1

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP (2003). **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* **4**(2): 249-264. 10.1093/biostatistics/4.2.249

Ivanov SV, Panaccione A, Nonaka D, Prasad ML, Boyd KL, Brown B, ... Yarbrough WG (2013). **Diagnostic SOX10 gene signatures in salivary adenoid cystic and breast basal-like carcinomas**. *Br J Cancer* **109**(2): 444-451. 10.1038/bjc.2013.326

Jafari M, Ansari-Pour N, Azimzadeh S and Mirzaie M (2017). **A logic-based dynamic modeling approach to explicate the evolution of the central dogma of molecular biology**. *PLoS One* **12**(12): e0189922. 10.1371/journal.pone.0189922

Jeannin P, Chaze T, Giai Gianetto Q, Matondo M, Gout O, Gessain A and Afonso PV (2018). **Proteomic analysis of plasma extracellular vesicles reveals mitochondrial stress upon HTLV-1 infection**. *Sci Rep* **8**(1): 5170. 10.1038/s41598-018-23505-0

Jiang YZ, Liu YR, Xu XE, Jin X, Hu X, Yu KD and Shao ZM (2016). **Transcriptome Analysis of Triple-Negative Breast Cancer Reveals an Integrated mRNA-lncRNA Signature with Predictive and Prognostic Value**. *Cancer Res* **76**(8): 2105-2114. 10.1158/0008-5472.CAN-15-3284

Johnson KC, Houseman EA, King JE and Christensen BC (2017). **Normal breast tissue DNA methylation differences at regulatory elements are associated with the cancer risk factor age**. *Breast Cancer Res* **19**(1): 81. 10.1186/s13058-017-0873-y

Jupp S, Burdett T, Leroy C and Parkinson HE (2015). **A new Ontology Lookup Service at EMBL-EBI**. *SWAT4LS*.

Kammers K, Cole RN, Tiengwe C and Ruczinski I (2015). **Detecting Significant Changes in Protein Abundance**. *EuPA Open Proteom* **7**: 11-19. 10.1016/j.euprot.2015.02.002

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, ... Sedmak G (2011). **Spatio-temporal**

transcriptome of the human brain. *Nature* **478**(7370): 483.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, ... Sestan N (2011). **Spatio-temporal transcriptome of the human brain**. *Nature* **478**(7370): 483-489. 10.1038/nature10523

Karrila S, Lee JH and Tucker-Kellogg G (2011). **A comparison of methods for data-driven cancer outlier discovery, and an application scheme to semisupervised predictive biomarker discovery**. *Cancer Inform* **10**: 109-120. 10.4137/CIN.S6868

Kauraniemi P and Kallioniemi A (2006). **Activation of multiple cancer-associated genes at the ERBB2 amplicon in breast cancer**. *Endocr Relat Cancer* **13**(1): 39-49. 10.1677/erc.1.01147

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, ... Pandey A (2009). **Human Protein Reference Database--2009 update**. *Nucleic Acids Res* **37**(Database issue): D767-772. 10.1093/nar/gkn892

Khondoker MR, Bachmann TT, Mewissen M, Dickinson P, Dobrzelecki B, Campbell CJ, ... Ghazal P (2010). **Multi-factorial analysis of class prediction error: estimating optimal number of biomarkers for various classification rules**. *J Bioinform Comput Biol* **8**(6): 945-965.

Kim YJ, Sung M, Oh E, Vrancken MV, Song JY, Jung K and Choi YL (2018). **Engrailed 1 overexpression as a potential prognostic marker in quintuple-negative breast cancer**. *Cancer Biol Ther* **19**(4): 335-345. 10.1080/15384047.2018.1423913

Koch I (2014). **Analysis of multivariate and high-dimensional data**. *Cambridge*, Cambridge University Press.9780521887939 (hardback)

Koonin EV (2005). **Orthologs, paralogs, and evolutionary genomics**. *Annu Rev Genet* **39**: 309-338. 10.1146/annurev.genet.39.073003.114725

Korkola JE, DeVries S, Fridlyand J, Hwang ES, Estep AL, Chen YY, ... Waldman FM (2003). **Differentiation of lobular versus ductal breast carcinomas by expression microarray analysis**. *Cancer Res* **63**(21): 7167-7175.

Kosti I, Jain N, Aran D, Butte AJ and Sirota M (2016). **Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues**. *Sci Rep* **6**: 24799. 10.1038/srep24799

Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV and Fotiadis DI (2015). **Machine learning applications in cancer prognosis and prediction**. *Comput Struct Biotechnol J* **13**: 8-17. 10.1016/j.csbj.2014.11.005

Kuzniar A, Laffeber C, Eppink B, Bezstarosti K, Dekkers D, Woelders H, ... Kanaar R (2017). **Semi-quantitative proteomics of mammalian cells upon short-term exposure to non-ionizing electromagnetic fields**. *PLoS One* **12**(2): e0170762. 10.1371/journal.pone.0170762

La D, Kong M, Hoffman W, Choi YI and Kihara D (2013). **Predicting permanent and transient protein-protein interfaces**. *Proteins* **81**(5): 805-818. 10.1002/prot.24235

Law CW, Chen Y, Shi W and Smyth GK (2014). **voom: Precision weights unlock linear model analysis tools for RNA-seq read counts**. *Genome Biol* **15**(2): R29. 10.1186/gb-2014-15-2-r29

Lee S, Rahnenfuhrer J, Lang M, De Preter K, Mestdagh P, Koster J, ... Schramm A (2014). **Robust selection of cancer survival signatures from high-throughput genomic data using two-fold subsampling**. *PLoS One* **9**(10): e108818. 10.1371/journal.pone.0108818

Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, ... Kendziorski C (2013). **EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments**. *Bioinformatics* **29**(8): 1035-1043. 10.1093/bioinformatics/btt087

Lenoir T and Giannella E (2006). **The emergence and diffusion of DNA microarray technology**. *J Biomed Discov Collab* **1**: 11. 10.1186/1747-5333-1-11

Lenz G, Wright G, Dave SS, Xiao W, Powell J, Zhao H, ... Lymphoma/Leukemia Molecular Profiling P (2008). **Stromal gene signatures in large-B-cell lymphomas**. *N Engl J Med* **359**(22): 2313-2323. 10.1056/NEJMoa0802885

Lenz M, Muller FJ, Zenke M and Schuppert A (2016). **Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data**. *Sci Rep* **6**: 25696. 10.1038/srep25696

Levitin HM, Yuan J and Sims PA (2018). **Single-Cell Transcriptomic Analysis of Tumor Heterogeneity**. *Trends Cancer* **4**(4): 264-268. 10.1016/j.trecan.2018.02.003

Li C and Wong WH (2001). **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection**. *Proc Natl Acad Sci U S A* **98**(1): 31-36. 10.1073/pnas.011404098

Li J and Tibshirani R (2013). **Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data**. *Stat Methods Med Res* **22**(5): 519-536. 10.1177/0962280211428386

Li L, Chaudhuri A, Chant J and Tang Z (2007). **PADGE: analysis of heterogeneous patterns of differential gene expression**. *Physiol Genomics* **32**(1): 154-159. 10.1152/physiolgenomics.00259.2006

Lian H (2008). **MOST: detecting cancer differential gene expression**. *Biostatistics* **9**(3): 411-418. 10.1093/biostatistics/kxm042

List M, Alcaraz N, Dissing-Hansen M, Ditzel HJ, Mollenhauer J and Baumbach J (2016). **KeyPathwayMinerWeb: online multi-omics network enrichment**. *Nucleic Acids Res* **44**(W1): W98-W104. 10.1093/nar/gkw373

Liu H and Setiono R (1996). **A probabilistic approach to feature selection-a filter solution**. *ICML*, Citeseer.

Liu H and Yu L (2005). **Toward integrating feature selection algorithms for classification and clustering**. *IEEE Transactions on knowledge and data engineering* **17**(4): 491-502.

Liu W-m, Mei R, Di X, Ryder TB, Hubbell E, Dee S, ... Baid J (2002). **Analysis of high density expression microarrays with signed-rank call algorithms**. *Bioinformatics* **18**(12): 1593-1599.

Liu YR, Jiang YZ, Xu XE, Yu KD, Jin X, Hu X, ... Shao ZM (2016). **Comprehensive transcriptome analysis identifies novel molecular subtypes and subtype-specific RNAs of triple-negative breast cancer**. *Breast Cancer Res* **18**(1): 33. 10.1186/s13058-016-0690-8

Liu Z and Hu J (2016). **Mislocalization-related disease gene discovery using gene expression based computational protein localization prediction**. *Methods* **93**: 119-127. 10.1016/j.ymeth.2015.09.022

Lopes KP, Campos-Laborie FJ, Vialle RA, Ortega JM and De Las Rivas J (2016). **Evolutionary hallmarks of the human proteome: chasing the age and coregulation of protein-coding genes**. *BMC Genomics* **17**(Suppl 8): 725. 10.1186/s12864-016-3062-y

Lowe R, Shirley N, Bleackley M, Dolan S and Shafee T (2017). **Transcriptomics technologies**. *PLoS Comput Biol* **13**(5): e1005457. 10.1371/journal.pcbi.1005457

Luck K, Sheynkman GM, Zhang I and Vidal M (2017). **Proteome-Scale Human Interactomics**. *Trends Biochem Sci* **42**(5): 342-354. 10.1016/j.tibs.2017.02.006

MacDonald JW and Ghosh D (2006). **COPA--cancer outlier profile analysis**. *Bioinformatics* **22**(23): 2950-2951. 10.1093/bioinformatics/btl433

Mardia KV, Kent JT and Bibby JM (1979). **Multivariate analysis**. *London ; New York*, Academic Press.0124712509

0124712525 (pbk.)

Margolin AA, Ong SE, Schenone M, Gould R, Schreiber SL, Carr SA and Golub TR (2009). **Empirical Bayes analysis of quantitative proteomics experiments**. *PLoS One* **4**(10): e7454. 10.1371/journal.pone.0007454

Martínez-Canales S, Cifuentes F, López De Rodas Gregorio M, Serrano-Oviedo L, Galán-Moya EM, Amir E, ... Ocaña A (2017). **Transcriptomic immunologic signature associated with favorable clinical outcome in basal-like breast tumors**. *PLOS ONE* **12**(5): e0175128. 10.1371/journal.pone.0175128

Martorell-Marugan J, Gonzalez-Rumayor V and Carmona-Saez P (2018). **mCSEA: Detecting subtle differentially methylated regions**. *bioRxiv*. 10.1101/293381

Mastriani E, Zhai R and Zhu S (2018). **Microarray-Based MicroRNA Expression Data Analysis with Bioconductor**. *Methods Mol Biol* **1751**: 127-138. 10.1007/978-1-4939-7710-9_9

Medina M, Collins AG, Taylor JW, Valentine JW, Lipps JH, Amaral-Zettler L and Sogin ML (2004). **Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa**. *International Journal of Astrobiology* **2**(3): 203-211. 10.1017/S1473550403001551

Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, ... Guigo R (2015). **Human genomics. The human transcriptome across tissues and individuals**. *Science* **348**(6235): 660-665. 10.1126/science.aaa0355

Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J and Barabasi AL (2015). **Disease networks. Uncovering disease-disease relationships through the incomplete interactome**. *Science* **347**(6224): 1257601. 10.1126/science.1257601

Meng C, Kuster B, Culhane AC and Gholami AM (2014). **A multivariate approach to the integration of multi-omics datasets**. *BMC Bioinformatics* **15**: 162. 10.1186/1471-2105-15-162

Meyer D, Dimitriadou E, Hornik K, Weingessel A and Leisch F (2014). **e1071: Misc Functions of the Department of Statistics (e1071), TU Wien**. *R package version 1.6-4*.

Molinaro AM, Simon R and Pfeiffer RM (2005). **Prediction error estimation: a comparison of resampling methods**. *Bioinformatics* **21**(15): 3301-3307. 10.1093/bioinformatics/bti499

Monaco G, van Dam S, Casal Novo Ribeiro JL, Larbi A and de Magalhaes JP (2015). **A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels**. *BMC Evol Biol* **15**: 259. 10.1186/s12862-015-0534-7

Mooney C, Wang YH and Pollastri G (2011). **SCLpred: protein subcellular localization prediction by N-to-1 neural networks**. *Bioinformatics* **27**(20): 2812-2819. 10.1093/bioinformatics/btr494

Moyers BA and Zhang J (2015). **Phylostratigraphic bias creates spurious patterns of genome evolution**. *Mol Biol Evol* **32**(1): 258-267. 10.1093/molbev/msu286

Moyers BA and Zhang J (2016). **Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution**. *Mol Biol Evol* **33**(5): 1245-1256. 10.1093/molbev/msw008

Mpindi JP, Sara H, Haapa-Paananen S, Kilpinen S, Pisto T, Bucher E, ... Kallioniemi O (2011). **GTI: a novel algorithm for identifying outlier gene expression profiles from integrated microarray datasets**. *PLoS One* **6**(2): e17259. 10.1371/journal.pone.0017259

Myers P, Espinosa R, Parr CS, Jones T, Hammond GS and Dewey TA. (2006). "The Animal Diversity Web." University of Michigan. from http://animaldiversity.org/.

Nabavi S (2016). **Identifying candidate drivers of drug response in heterogeneous cancer by mining high throughput genomics data**. *BMC Genomics* **17**(1): 638. 10.1186/s12864-016-2942-5

Nebert DW, Wikvall K and Miller WL (2013). **Human cytochromes P450 in health and disease**. *Philos Trans R Soc Lond B Biol Sci* **368**(1612): 20120431. 10.1098/rstb.2012.0431

Neme R and Tautz D (2013). **Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution**. *BMC Genomics* **14**: 117. 10.1186/1471-2164-14-117

Neyt C, Jagla K, Thisse C, Thisse B, Haines L and Currie PD (2000). **Evolutionary origins of vertebrate appendicular muscle**. *Nature* **408**(6808): 82-86. 10.1038/35040549

Nigam K, McCallum AK, Thrun S and Mitchell T (2000). **Text classification from labeled and unlabeled documents using EM**. *Machine learning* **39**(2-3): 103-134.

Noh H, Shoemaker JE and Gunawan R (2018). **Network perturbation analysis of gene transcriptional profiles reveals protein targets and mechanism of action of drugs and influenza A viral infection**. *Nucleic Acids Res* **46**(6): e34. 10.1093/nar/gkx1314

Noto K, Majidi S, Edlow AG, Wick HC, Bianchi DW and Slonim DK (2015). **CSAX: Characterizing Systematic Anomalies in eXpression Data**. *J Comput Biol* **22**(5): 402-413. 10.1089/cmb.2014.0155

Nueda MJ, Sebastian P, Tarazona S, Garcia-Garcia F, Dopazo J, Ferrer A and Conesa A (2009). **Functional assessment of time course microarray data**. *BMC Bioinformatics* **10 Suppl 6**: S9. 10.1186/1471-2105-10-S6-S9

O'Connor CM, Adams JU and Fairman J (2010). **Essentials of cell biology**. *Cambridge, MA: NPG Education* **1**.

Ozsolak F and Milos PM (2011). **RNA sequencing: advances, challenges and opportunities**. *Nat Rev Genet* **12**(2): 87-98. 10.1038/nrg2934

Pagel O, Loroch S, Sickmann A and Zahedi RP (2015). **Current strategies and findings in clinically relevant post-translational modification-specific proteomics**. *Expert Rev Proteomics* **12**(3): 235-253. 10.1586/14789450.2015.1042867

Parfrey LW, Barbero E, Lasser E, Dunthorn M, Bhattacharya D, Patterson DJ and Katz LA (2006). **Evaluating support for the current classification of eukaryotic diversity**. *PLoS Genet* **2**(12): e220. 10.1371/journal.pgen.0020220

Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, ... Bernard PS (2009). **Supervised risk predictor of breast cancer based on intrinsic subtypes**. *J Clin Oncol* **27**(8): 1160-1167. 10.1200/JCO.2008.18.1370

Peng J, Hui W and Shang X (2018). **Measuring phenotype-phenotype similarity through the interactome**. *BMC Bioinformatics* **19**(5): 114. 10.1186/s12859-018-2102-9

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, ... Botstein D (2000). **Molecular portraits of human breast tumours**. *Nature* **406**(6797): 747-752. 10.1038/35021093

Pierson E, Consortium GT, Koller D, Battle A, Mostafavi S, Ardlie KG, ... Dermitzakis ET (2015). **Sharing and Specificity of Co-expression Networks across 35 Human Tissues**. *PLoS Comput Biol* **11**(5): e1004220. 10.1371/journal.pcbi.1004220

Pounds S and Cheng C (2006). **Robust estimation of the false discovery rate**. *Bioinformatics* **22**(16): 1979-1987. 10.1093/bioinformatics/btl328

Prat Y, Fromer M, Linial N and Linial M (2009). **Codon usage is associated with the evolutionary age of genes in metazoan genomes**. *BMC Evol Biol* **9**: 285. 10.1186/1471-2148-9-285

Prieto C and De Las Rivas J (2006). **APID: Agile Protein Interaction DataAnalyzer**. *Nucleic Acids Res* **34**(Web Server issue): W298-302. 10.1093/nar/gkl128

Prieto C, Rivas MJ, Sanchez JM, Lopez-Fidalgo J and De Las Rivas J (2006). **Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes**. *Bioinformatics* **22**(9): 1103-1110. 10.1093/bioinformatics/btl053

Qian M, Wang DC, Chen H and Cheng Y (2017). **Detection of single cell heterogeneity in cancer**. *Semin Cell Dev Biol* **64**: 143-149. 10.1016/j.semcdb.2016.09.003

Rahman J, Mondal NI, Islam KB and Hasan AM (2016). **Feature Fusion Based SVM Classifier for Protein Subcellular Localization Prediction**. *J Integr Bioinform* **13**(1): 23-33. 10.1515/jib-2016-288

Rao VS, Srinivas K, Sujini GN and Kumar GN (2014). **Protein-protein interaction detection: methods and analysis**. *Int J Proteomics* **2014**: 147648. 10.1155/2014/147648

Risueno A, Fontanillo C, Dinger ME and De Las Rivas J (2010). **GATExplorer: genomic and transcriptomic explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs**. *BMC Bioinformatics* **11**: 221. 10.1186/1471-2105-11-221

Robinson MD, McCarthy DJ and Smyth GK (2010). **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* **26**(1): 139-140. 10.1093/bioinformatics/btp616

Robnik-Šikonja M and Kononenko I (2003). **Theoretical and empirical analysis of ReliefF and RReliefF**. *Machine learning* **53**(1-2): 23-69.

Roden DL, Sewell GW, Lobley A, Levine AP, Smith AM and Segal AW (2014). **ZODET: software for the identification, analysis and visualisation of outlier genes in microarray expression data**. *PLoS One* **9**(1): e81123. 10.1371/journal.pone.0081123

Rodriguez-Gonzalez FG, Mustafa DA, Mostert B and Sieuwerts AM (2013). **The challenge of gene expression profiling in heterogeneous clinical samples**. *Methods* **59**(1): 47-58. 10.1016/j.ymeth.2012.05.005

Rolland T, Tasan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, ... Vidal M (2014). **A proteome-scale map of the human interactome network**. *Cell* **159**(5): 1212-1226. 10.1016/j.cell.2014.10.050

Ronan T, Qi Z and Naegle KM (2016). **Avoiding common pitfalls when clustering biological data**. *Science Signaling* **9**(432): re6-re6. 10.1126/scisignal.aad1932

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, ... Vidal M (2005). **Towards a proteome-scale map of the human protein-protein interaction network**. *Nature* **437**(7062): 1173-1178. 10.1038/nature04209

Rubben A and Araujo A (2017). **Cancer heterogeneity: converting a limitation into a source of biologic information**. *J Transl Med* **15**(1): 190. 10.1186/s12967-017-1290-9

Sabatini LM, Ota T and Azen EA (1993). **Nucleotide sequence analysis of the human salivary protein genes HIS1 and HIS2, and**

evolution of the STATH/HIS gene family. *Mol Biol Evol* **10**(3): 497-511. 10.1093/oxfordjournals.molbev.a040022

Saben J, Zhong Y, McKelvey S, Dajani NK, Andres A, Badger TM, ... Shankar K (2014). **A comprehensive analysis of the human placenta transcriptome**. *Placenta* **35**(2): 125-131. 10.1016/j.placenta.2013.11.007

Saeys Y, Inza I and Larranaga P (2007). **A review of feature selection techniques in bioinformatics**. *Bioinformatics* **23**(19): 2507-2517. 10.1093/bioinformatics/btm344

Saito Y, Sugimoto C, Mituyama T and Wakao H (2017). **Epigenetic silencing of V(D)J recombination is a major determinant for selective differentiation of mucosal-associated invariant t cells from induced pluripotent stem cells**. *PLoS One* **12**(3): e0174699. 10.1371/journal.pone.0174699

Sandberg R and Larsson O (2007). **Improved precision and accuracy for microarrays using updated probe set definitions**. *BMC Bioinformatics* **8**: 48. 10.1186/1471-2105-8-48

Sanguansat P (2012). **Principal Component Analysis**, IntechOpen.978-953-51-0195-6

Sardar AJ, Oates ME, Fang H, Forrest AR, Kawaji H, Consortium F, ... Rackham OJ (2014). **The evolution of human cells in terms of protein innovation**. *Mol Biol Evol* **31**(6): 1364-1374. 10.1093/molbev/mst139

Sasahira T, Kirita T, Nishiguchi Y, Kurihara M, Nakashima C, Bosserhoff AK and Kuniyasu H (2016). **A comprehensive expression analysis of the MIA gene family in malignancies: MIA gene family members are novel, useful markers of esophageal, lung, and cervical squamous cell carcinoma**. *Oncotarget* **7**(21): 31137-31152. 10.18632/oncotarget.9082

Schena M, Shalon D, Davis RW and Brown PO (1995). **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* **270**(5235): 467-470.

Seebacher J and Gavin AC (2011). **SnapShot: Protein-protein interaction networks**. *Cell* **144**(6): 1000, 1000 e1001. 10.1016/j.cell.2011.02.025

Seiler J, Breinig M, Caudron-Herger M, Polycarpou-Schwarz M, Boutros M and Diederichs S (2017). **The lncRNA VELUCT strongly regulates viability of lung cancer cells despite its extremely low abundance**. *Nucleic Acids Research* **45**(9): 5458-5469. 10.1093/nar/gkx076

Sestak MS, Bozicevic V, Bakaric R, Dunjko V and Domazet-Loso T (2013). **Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems**. *Front Zool* **10**(1): 18. 10.1186/1742-9994-10-18

Sevimoglu T and Arga KY (2014). **The role of protein interaction networks in systems biomedicine**. *Comput Struct Biotechnol J* **11**(18): 22-27. 10.1016/j.csbj.2014.08.008

Sewell JA and Fuxman Bass JI (2017). **Cellular network perturbations by disease-associated variants**. *Curr Opin Syst Biol* **3**: 60-66. 10.1016/j.coisb.2017.04.009

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, ... Ideker T (2003). **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* **13**(11): 2498-2504. 10.1101/gr.1239303

Shaw GT, Shih ES, Chen CH and Hwang MJ (2011). **Preservation of ranking order in the expression of human Housekeeping genes**. *PLoS One* **6**(12): e29314. 10.1371/journal.pone.0029314

Shen YQ and Burger G (2010). **TESTLoc: protein subcellular localization prediction from EST data**. *BMC Bioinformatics* **11**: 563. 10.1186/1471-2105-11-563

Shin CJ, Wong S, Davis MJ and Ragan MA (2009). **Protein-protein interaction as a predictor of subcellular location**. *BMC Syst Biol* **3**: 28. 10.1186/1752-0509-3-28

Singh RK and Sivabalakrishnan M (2015). **Feature Selection of Gene Expression Data for Cancer Classification: A Review**. *Procedia Computer Science* **50**: 52-57. doi.org/10.1016/j.procs.2015.04.060

Smyth GK (2004). **Linear models and empirical bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* **3**: Article3. 10.2202/1544-6115.1027

Sonawane AR, Platig J, Fagny M, Chen CY, Paulson JN, Lopes-Ramos CM, ... Kuijjer ML (2017). **Understanding Tissue-Specific Gene Regulation**. *Cell Rep* **21**(4): 1077-1088. 10.1016/j.celrep.2017.10.001

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A and Tyers M (2006). **BioGRID: a general repository for interaction datasets**. *Nucleic Acids Res* **34**(Database issue): D535-539. 10.1093/nar/gkj109

Stefan M, Zhang W, Concepcion E, Yi Z and Tomer Y (2014). **DNA methylation profiles in type 1 diabetes twins point to strong epigenetic effects on etiology**. *J Autoimmun* **50**: 33-37. 10.1016/j.jaut.2013.10.001

Storey JD (2002). **A direct approach to false discovery rates**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(3): 479-498.

Stouffer SA and Hovland CI (1949). **Studies in Social Psychology in World War II**, Princeton: Princeton University Press

Stretch C, Khan S, Asgarian N, Eisner R, Vaisipour S, Damaraju S, ... Baracos VE (2013). **Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature**. *PLoS One* **8**(6): e65380. 10.1371/journal.pone.0065380

Stuart JM, Segal E, Koller D and Kim SK (2003). **A gene-coexpression network for global discovery of conserved genetic modules**. *Science* **302**(5643): 249-255. 10.1126/science.1087447

Svetnik V, Liaw A, Tong C and Wang T (2004). **Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules**. *International Workshop on Multiple Classifier Systems*, Springer.

Szekely GJ, Rizzo ML and Bakirov NK (2007). **Measuring and testing dependence by correlation of distances**. *Ann. Statist.* **35**(6): 2769-2794. 10.1214/009053607000000505

Tan CS, Ting WS, Mohamad MS, Chan WH, Deris S and Shah ZA (2014). **A review of feature extraction software for microarray gene expression data**. *Biomed Res Int* **2014**: 213656. 10.1155/2014/213656

Tanamai W, Chen C, Siavoshi S, Cerussi A, Hsiang D, Butler J and Tromberg B (2009). **Diffuse optical spectroscopy measurements of healing in breast tissue after core biopsy: case study**. *J Biomed Opt* **14**(1): 014024. 10.1117/1.3028012

Tang J, Alelyani S and Liu H (2014). **Feature selection for classification: A review**. *Data classification: Algorithms and applications*: 37.

Thomou T, Mori MA, Dreyfuss JM, Konishi M, Sakaguchi M, Wolfrum C, ... Kahn CR (2017). **Adipose-derived circulating miRNAs regulate gene expression in other tissues**. *Nature* **542**(7642): 450-455. 10.1038/nature21365

Thul PJ, Akesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, ... Lundberg E (2017). **A subcellular map of the human proteome**. *Science* **356**(6340). 10.1126/science.aal3321

Tibshirani R and Hastie T (2007). **Outlier sums for differential gene expression analysis**. *Biostatistics* **8**(1): 2-8. 10.1093/biostatistics/kxl005

Ting L, Cowley MJ, Hoon SL, Guilhaus M, Raftery MJ and Cavicchioli R (2009). **Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling**. *Mol Cell Proteomics* **8**(10): 2227-2242. 10.1074/mcp.M800462-MCP200

Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, ... Chinnaiyan AM (2005). **Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer**. *Science* **310**(5748): 644-648. 10.1126/science.1117679

Tusher VG, Tibshirani R and Chu G (2001). **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci U S A* **98**(9): 5116-5121. 10.1073/pnas.091062498

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, ... Ponten F (2015). **Proteomics. Tissue-based map of the human proteome**. *Science* **347**(6220): 1260419. 10.1126/science.1260419

Uhlen M, Zhang C, Lee S, Sjostedt E, Fagerberg L, Bidkhori G, ... Ponten F (2017). **A pathology atlas of the human cancer transcriptome**. *Science* **357**(6352). 10.1126/science.aan2507

Vega-Hernandez MC, Patino-Alonso MC, Cabello R, Galindo-Villardon MP and Fernandez-Berrocal P (2017). **Perceived Emotional Intelligence and Learning Strategies in Spanish University Students: A New Perspective from a Canonical Non-symmetrical Correspondence Analysis**. *Front Psychol* **8**: 1888. 10.3389/fpsyg.2017.01888

Veres DV, Gyurko DM, Thaler B, Szalay KZ, Fazekas D, Korcsmaros T and Csermely P (2015). **ComPPI: a cellular compartment-specific database for protein-protein interaction network analysis**. *Nucleic Acids Res* **43**(Database issue): D485-493. 10.1093/nar/gku1007

Vidal M (2009). **A unifying view of 21st century systems biology**. *FEBS Lett* **583**(24): 3891-3894. 10.1016/j.febslet.2009.11.024

Vidal M, Cusick ME and Barabasi AL (2011). **Interactome networks and human disease**. *Cell* **144**(6): 986-998. 10.1016/j.cell.2011.02.016

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P (2002). **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* **417**(6887): 399-403. 10.1038/nature750

Wan C, Borgeson B, Phanse S, Tu F, Drew K, Clark G, ... Emili A (2015). **Panorama of ancient metazoan macromolecular complexes**. *Nature* **525**(7569): 339-344. 10.1038/nature14877

Wan S, Mak MW and Kung SY (2015). **mPLR-Loc: an adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction**. *Anal Biochem* **473**: 14-27. 10.1016/j.ab.2014.10.014

Wan YW, Allen GI and Liu Z (2016). **TCGA2STAT: simple TCGA data access for integrated statistical analysis in R**. *Bioinformatics* **32**(6): 952-954. 10.1093/bioinformatics/btv677

Wang C, Taciroglu A, Maetschke SR, Nelson CC, Ragan MA and Davis MJ (2012). **mCOPA: analysis of heterogeneous features in cancer expression data**. *J Clin Bioinforma* **2**(1): 22. 10.1186/2043-9113-2-22

Wang L and Fu X (2006). **Data mining with computational intelligence**, Springer Science & Business Media.3540288031

Wang L, Wang Y and Chang Q (2016). **Feature selection methods for big data bioinformatics: A survey from the search perspective**. *Methods* **111**: 21-31. 10.1016/j.ymeth.2016.08.014

Wang Y and Rekaya R (2010). **LSOSS: Detection of Cancer Outlier Differential Gene Expression**. *Biomark Insights* **5**: 69-78.

Wang Z, Gerstein M and Snyder M (2009). **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* **10**(1): 57-63. 10.1038/nrg2484

Werner HM, Mills GB and Ram PT (2014). **Cancer Systems Biology: a peek into the future of patient care?** *Nat Rev Clin Oncol* **11**(3): 167-176. 10.1038/nrclinonc.2014.6

Wockner LF, Noble EP, Lawford BR, Young RM, Morris CP, Whitehall VL and Voisey J (2014). **Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients**. *Transl Psychiatry* **4**: e339. 10.1038/tp.2013.111

Wolf YI, Novichkov PS, Karev GP, Koonin EV and Lipman DJ (2009). **The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages**. *Proc Natl Acad Sci U S A* **106**(18): 7273-7280. 10.1073/pnas.0901808106

Wu B (2007). **Cancer outlier differential gene expression detection**. *Biostatistics* **8**(3): 566-575. 10.1093/biostatistics/kxl029

Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F and Spencer F (2004). **A Model-Based Background Adjustment for Oligonucleotide Expression Arrays**. *Journal of the American Statistical Association* **99**(468): 909-917. 10.1198/016214504000000683

Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM and Eisenberg D (2000). **DIP: the database of interacting proteins**. *Nucleic Acids Res* **28**(1): 289-291.

Xu X, Lu L, He P and Chen L (2013). **Protein localization prediction using random walks on graphs**. *BMC Bioinformatics* **14 Suppl 8**: S4. 10.1186/1471-2105-14-S8-S4

Xue M, Liu H, Zhang L, Chang H, Liu Y, Du S, ... Wang P (2017). **Computational identification of mutually exclusive transcriptional drivers dysregulating metastatic microRNAs in prostate cancer**. *Nat Commun* **8**: 14917. 10.1038/ncomms14917

Yang Z and Yang Z (2013). **Prediction of heterogeneous differential genes by detecting outliers to a Gaussian tight cluster**. *BMC Bioinformatics* **14**: 81. 10.1186/1471-2105-14-81

Yi S, Lin S, Li Y, Zhao W, Mills GB and Sahni N (2017). **Functional variomics and network perturbation: connecting genotype to phenotype in cancer**. *Nat Rev Genet* **18**(7): 395-410. 10.1038/nrg.2017.8

Yip CH and Rhodes A (2014). **Estrogen and progesterone receptors in breast cancer**. *Future Oncol* **10**(14): 2293-2301. 10.2217/fon.14.110

Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, ... Vidal M (2008). **High-quality binary protein interaction map of the**

yeast interactome network. *Science* **322**(5898): 104-110. 10.1126/science.1158684

Yu L and Liu H (2004). **Efficient feature selection via analysis of relevance and redundancy**. *Journal of machine learning research* **5**(Oct): 1205-1224.

Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M and Cesareni G (2002). **MINT: a Molecular INTeraction database**. *FEBS Lett* **513**(1): 135-140.

Zhao Q, Wang D, Chen Y and Qu X (2015). **Multisite protein subcellular localization prediction based on entropy density**. *Biomed Mater Eng* **26 Suppl 1**: S2003-2009. 10.3233/BME-151504

Zhu M, Yu X, Ahlberg PE, Choo B, Lu J, Qiao T, ... Zhu Y (2013). **A Silurian placoderm with osteichthyan-like marginal jaw bones**. *Nature* **502**(7470): 188-193. 10.1038/nature12617

APPENDIX 1

R vignette describing the use and required inputs to run DECO algorithm and DECO R package. This vignette is also available in the original R package, called *deco*, and in the website of Javier De Las Rivas' laboratory (http://bioinfow.dep.usal.es/deco/).

DECO

DEcomposing heterogeneous Cohorts using Omic profiling data

F. J. Campos-Laborie, J. M. Sanchez-Santos and J. De Las Rivas
Bioinformatics and Functional Genomics Group
Cancer Research Centre (CiC-IBMCC, USAL/CSIC/IBSAL)
Salamanca (Spain)

Abstract

Here we present a tutorial to use **DECO**, a method to explore and find differences in heterogeneous large datasets usually produced in biological or biomedical omic-wide studies. The method makes a comprehensive analysis of multidimensional datasets (usually consisting on a collection of samples where hundreds or thousands of features have been measured with a large-scale high-throughput technology, for example, a genomic or proteomic technique). The method finds the differences in the profiles of the features along the samples and identifies the associations between them, showing the features that best mark a given class or category as well as possible sample outliers that do not follow the same pattern of the majority of the corresponding cohort. The method can be used in a supervised or unsupervised mode, it allows the discovery of multiple classes or categories and is quite adequate for patients stratification.

Contents

1. Variability and heterogeneity in high-dimensional data

Individual diversity and variability is one of the most complex issues to deal within high-dimensional studies of large populations, as the ones currently performed in biomedical analyses using omic technologies. DECO is a method that combines two main computational procedures: **(i)** a **Recursive Differential Analysis (RDA)** that performs combinatorial sampling without replacement to select multiple sample subsets followed by differential analysis; and **(ii)** a **Non-Symmetrical Correspondence Analysis (NSCA)** of differential events that allow the characterization and assigment of features and samples in a common multidimensional space, combining in a single statistic parameterization both the feature-sample changes detected and a predictor-response information.

The statistical procedure followed in both parts of the method are detailed in the original publication [1], but this brief **vignette** explains how to use **DECO** to analyze multidimensional datasets that may include heterogeneous samples. The aim is to improve characterization and stratification of complex sample series, mostly focusing on large patient cohorts, where the existence of outlier or mislabeled samples is quite possible.

In this way, **DECO** can reveal exclusive associations between features and samples based in specific differential signal and provide a better way for the stratifycation of populations using multidimensional large-scale data. The method is applied to data derived from different **omic technologies**, for example: genome-wide expression data obtained with microarrays or with RNA-seq (either for genes, miRNAs, ncRNAs, etc).

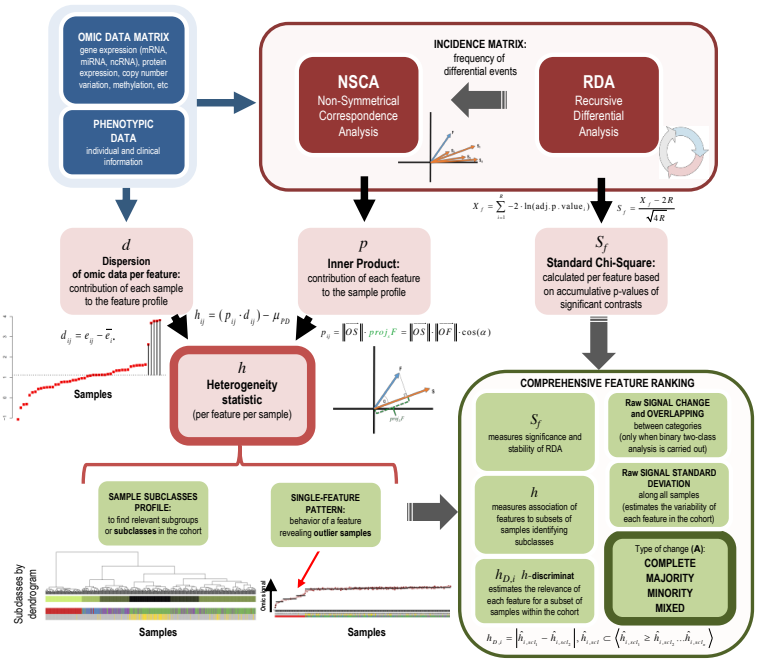


Figure 1: Workflow of DECO algorithm.

2. Installation

The **deco** R source package can be downloaded from R CRAN repository or from our lab website (http://bioinfow.dep.usal.es/deco/). It can be installed by downloading and executing in UNIX-like terminal:

```
R CMD INSTALL /path/to/deco_1.0.tar.gz
```

or first choosing your nearest CRAN mirror using `chooseCRANmirror()` and executing in R:

```
install.packages("deco", type="source", dependencies = TRUE)
```

If any problem with dependencies are reported, user can download R packages directly from Bioconductor repository and repeat previous instruction to install **deco** R package:

```
# Bioconductor dependencies
source("https://bioconductor.org/biocLite.R")
biocLite(c("AnnotationDbi", "foreign", "limma", "Biobase",
         "gdata", "cluster", "gplots", "RColorBrewer",
         "snowfall", "scatterplot3d", "made4", "locfit",
         "lisp", "ade4", "sfsmisc"))

# Loading package in R
library(deco)
```

It can be also done using `devtools` R package to install dependencies from local directory (in this case the user has to decompress .tar.gz previously):

```
# Loading devtools R package...
library(devtools)

# Installing dependencies. Path to directory containing decompressed R package
install_deps("/path/to/deco/",dependencies="logical")

# Loading package in R
library(deco)
```

This R package contains a experimental dataset as example and all functions needed to run an analysis.

3. Experimental data

At presernt, the method directly supports two types of data from transcriptomic technologies: microarrays and RNA-seq platforms. In case of microarrays, robust normalization of raw signal is needed for correct application of *eBayes* method (done for example with the **RMA** method or with **normalizeBetweenArrays** method from LIMMA package [2]). Notwithstanding, RNA-seq read counts matrix (genes or transcripts as rows and samples as columns) can be the input; and in this case then user should apply voom normalization method [2]. Below, we show two different examples of both types of datasets obtained with such platforms.

3.1 Microarrays dataset: study on lymphoma subtypes

Here, a normalized microarray gene expression matrix from clinical samples is used as example taken from Scarfo et al.[3]. **Anaplastic Large Cell Lymphoma** (**ALCL**) is an heterogeneous disease with two well differentiated forms based on ALK gene expression: *ALK(-)* and *ALK(+)*. The dataset, obtained in GSE65823 from GEO database, corresponds to genome-wide expression profiles of human T-cell samples hybridizated on *Affymetrix* HGU133Plus2.0 platform, which were mapped to ENSEMBL genes with *genemapperhgu133plus2cdf*

CDF package from GATExplorer [4]. The mapping from *Affymetrix* probes to genes can be also done using BrainArray CDF packages.

The main interest to include this dataset, to be analysed with DECO, is because using this sample set Scarfo et al. [3] identified of **a subset of patients within the ALK(-) class** discovering high ectopic expression of several gene markers. To do so, the authors applied one of the most used methods for detection of outliers and heterogeneous behavior, that is COPA [5]. Further comparisons between **COPA** and other related methods can be found in our paper about DECO [1]. The phenotypic information about this sample set provided by GEO database were included in an *ExpressionSet* object. This R object called `ALCLdata` could be directly loaded as follows:

```
data(ALCLdata)
# to see the ExpressionSet object
ALCLdata
```

```
# to see the phenotypic information
pData(ALCLdata)
```

Classes vector to run a *supervised* analysis (explained in following section) to compare both ALCL classes: *positive*ALK and *negative*ALK.

```
classes.ALCL <- pData(ALCLdata)[,"Alk.positivity"]
names(classes.ALCL) <- sampleNames(ALCLdata)
```

3.2 RNA-seq dataset: study on breast cancer subtypes

Here, we show a RNAseq dataset analysed using DECO. The dataset was downloaded from The Cancer Genome Atlas (TCGA). It is composed by 878 clinical samples from patients with different subtypes of Breast Cancer [6], that include the standard classes (given by markers ESR1, PGR and HER2) and two classes associated to the cell-type, called: **Invasive Ductal Carcinoma** (IDC) and **Invasive Lobular Carcinoma** (ILC). The genes of the dataset are mapped to HGNC symbol IDs. The dataset can be loaded directly in R or downlad from the TCGA data portal:

```
# Load required R package to download data.
library(TCGA2STAT)
# Download complete Breast Cancer RNAseq counts matrix from TCGA database.
BRCA_count <- getTCGA(disease = "BRCA", data.type = "RNASeq",
                      type = "count", clinical=TRUE)

# Different slots included in the data (count matrix, clinical data and a merged matrix).
names(BRCA_count)
[1] "dat"        "clinical"   "merged.dat"

# Apply 'voom' normalization from LIMMA R package to read count matrix
# It would return a log2_counts_per_million (logCPM) inside 'E' slot.
BRCA_count$dat <- limma::voom(BRCA_count$dat, normalize = "quantile")$E
```

Then, user can run *voom* normalization method provided in LIMMA R package to calculates matrix of **logCPMs**. Further information about **voom** normalization and its properties can be found in LIMMA R package [2]. The normalized matrix is then analysed using the RDA method of DECO.

DECO is also able to analyse other RNAseq data types (RPKMs, FPKMs or TPMs values). The data are usually log scaled. Here, we shown an example using data type RPKMs from TCGA database.

```
# Load required R package to download data.
library(TCGA2STAT)
```

```
# Download complete Breast Cancer RNAseq RPKMs matrix from TCGA database.
BRCA_rpkm <- getTCGA(disease = "BRCA", data.type = "RNASeq2", clinical=TRUE)

# Different slots included in the data (count matrix, clinical data and a merged matrix).
names(BRCA_rpkm)
[1] "dat"        "clinical"   "merged.dat"

# Conversion to log2 scale, required to apply LIMMA.
BRCA_rpkm$dat <- log2(BRCA_rpkm$dat + 1)
```

3.3 Use of other *omic* platforms

Together with RNA-seq or microarray platforms, DECO algorithm can be applied to datasets obtained with other **omic platforms** (as far as a correct normalization of the data per sample can be achieved).

In order to provide an example of other platform, we show an example of a **miRNAs dataset** from same TCGA database used above:

```
library(TCGA2STAT)
# Download complete Breast Cancer miRNAseq rpmmms matrix from TCGA database.
BRCA_mirna <- getTCGA(disease = "BRCA", data.type = "miRNASeq",
                      type = "rpmmm", clinical=TRUE)

# Different slots included in the data.
names(BRCA_mirna)
[1] "dat"      "clinical"   "merged.dat"

# Convert 'rpmmm' to log2 scale.
BRCA_mirna$dat <- log2(BRCA_mirna$dat + 1)
```

Additionally, more information about different data platforms available for direct download to R environment can be queried on `TCGA2STAT` R package vignette.

4. RDA, Recursive Differential Analysis: *Standard.Chi.Square*

We proposed a recursive subsampling strategy which selects subsets of samples (from the different classes) and compares all against all (in an exhaustive search). When the number of combinations is very large a random selection of all possible subsets is done. In order to obtain the best possible results, three parameters should be taken into consideration before running the analysis: **(i)** the subsampling size called `r` (DECO method calculates an optimal size of subsampling subsets if the user does not define it); **(ii)** number of subsets or combinations, called `iterations`, to compare in this subsampling step; and **(iii)** adjusted.p.value threshold for the differential tests or contrasts, called `q.val` and computed using eBayes from LIMMA [2].

Aiming to summarize all positive differential events (DE) for each feature (combinations with a lower *adj.p.value* than threshold), Fisher's combined probability test is applied to each final feature vector of *adj.p.values* to obtain a **Standard.Chi.Square**, which will is not affected by type of analysis (*supervised* or *unsupervised*) because it only takes into account number of positive DE events.

4.1 Supervised analysis

Depending on *classes* input vector, a *supervised* analysis compares just two types of samples (i.e. healthy donors *versus* patients in a typical biomedical study). The `decoRDA()` RDA function will adjust the optimal subsampling size r (that the user can modify) to explore all DE signal, and UP and DOWN events will be taken into account for posterior **NSCA**. Here, an example of *decoRDA* function using `ALCLdata` dataset:

```
# if gene annotation will be required (annot = TRUE or rm.xy = TRUE)
library(org.Hs.eg.db)
# example of SUPERVISED design with Affymetrix microarrays data
sub <- decoRDA(data = exprs(ALCLdata), classes = classes.ALCL, q.val = 0.01,
                    rm.xy = T, r = NULL, cpus = 6, parallel = T, temp.path = getwd(),
                    control = "pos", annot = T, id.type = "ENSEMBL",
                    pack.db = "org.Hs.eg.db")
```

This **RDA** procedure generates an `incidenceMatrix` which counts differential events per gene (feature) per sample. Thus, this matrix would contain just differential genes as rows and samples as columns with one differential event at least.

```
dim(sub$incidenceMatrix)
```

The `incidenceMatrix` produced after the RDA, can reveal the important changes that mark an entire subclass (grey boxes in **Figure 1**), as well as specific signal changes that mark a subclass of samples (red boxes in **Figure 1**). As we can see in a simple example (**Figure 1**), both *Gene 1* and *Gene 2* seem to mark two subclasses (or subtypes) inside each compared class, while *Gene 3* and *Gene 4* reflect the behaviour of control and case classes respectively. Following the RDA step, the NSCA step analyses the numbers of the `incidenceMatrix`. The NSCA analysis is also done splitting UP and DOWN changes when the algorithm is run in *supervised* mode.

4.2 Unsupervised analysis

If *classes* input vector is empty, a *unsupervised* analysis is run comparing all against all samples taking different subsets (each combination of samples is unique) and looking for UP events. Then, those samples which show any differential change with statistical significance will be counted. In order to clarify final results of NSCA analysis, it is important to underline that just UP regulated events will be assigned to samples, while both UP and DOWN regulation events are counted in the *supervised* analyisis explain above.

```
# if gene annotation will be required (annot = TRUE or rm.xy = TRUE)
library(org.Hs.eg.db)
```

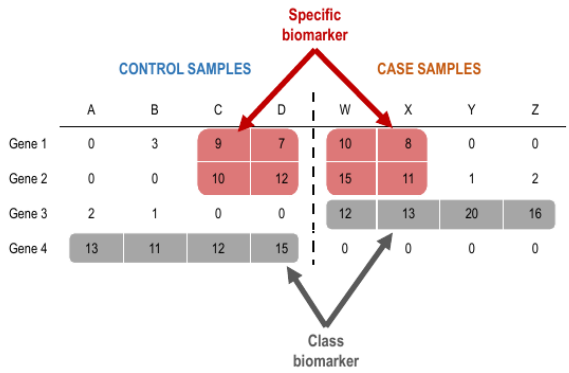


Figure 2: Example of an "incidence matrix" obtained after RDA for a SUPERVISED analysis.

```
# example of UNSUPERVISED design with RNA-seq data (log2[RPKM])
sub <- decoRDA(data = BRCA_rpkm$dat, q.val = 0.01, r = NULL, cpus = 6,
               parallel = T, temp.path = getwd(), annot = T, rm.xy = T,
               id.type = "SYMBOL", pack.db = "org.Hs.eg.db")
```

RDA procedure generates in this case an `incidenceMatrix` which counts just UP events per gene per sample.

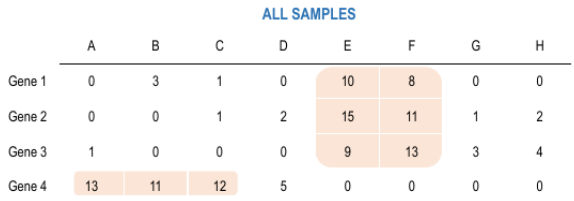


Figure 3: Example of an "incidence matrix" obtained after RDA for a UNSUPERVISED analysis

4.3 Multiclass analysis

Together with *supervised* or *unsupervised* analyses, the method can be run for *multiclass comparison*, taking subsets of samples from several classes identified *a priori* and forcing them to be compared. Then, we would count differential events per feature per sample but there will not be mix between different classes. Here, we show an example of a breast cancer dataset (from Ciriello et al. [6], log2(RPKM+1) scaled) that uses the well-defined PAM50 classes:

```
# Loading phenoData for TCGA paper (Cell 2012)
data(phenoData.BRCA.TCGA)
# Selecting samples with complete phenoData.
BRCA_rpkm$dat <- BRCA_rpkm$dat[, colnames(BRCA_rpkm$dat) %in%
                               phenoData.BRCA.TCGA[,"mRNA"]]
# Selecting IDC and ILC samples.
BRCA_rpkm$dat <- BRCA_rpkm$dat[, rownames(phenoData.BRCA.TCGA[
```

```
  !phenoData.BRCA.TCGA[,"Final.Pathology"] %in% c("Other","Mixed.IDC.ILC"),])]
# Creating a multiclass vector with PAM50 subclasses.
classes.PAM <- phenoData.BRCA.TCGA[colnames(BRCA_rpkm$dat), "PAM50"]
names(classes.PAM) <- rownames(phenoData.BRCA.TCGA[colnames(BRCA_rpkm$dat),])
# example of MULTICLASS design with RNA-seq data (log2[RPKM])
sub <- decoRDA(data = BRCA_rpkm$dat, classes = classes.PAM, q.val = 0.01,
               r = NULL, cpus = 6, parallel = T,
               temp.path = getwd(), annot = T, rm.xy = T,
               id.type = "SYMBOL", pack.db = "org.Hs.eg.db")
```

4.4 Running the RDA function: `decoRDA()`

This vignette presented some examples of `decoRDA()` subsampling function for *supervised* and *unsupervised* analyses (if user has two classes of samples or not). Now, details about all the input parameters which control the RDA procedure are indicated:

- `data` input corresponds to our expression matrix with features as rows and samples as columns.
- `q.val` is the threshold imposed to the *adjusted.p.value* from *LIMMA* method in each iteration.
- `r` is the resampling size.
- `temp.path` defines a location in your computer where `decoRDA()` would save temporary results.
- `classes` is a character vector or factor indicating to which class each sample belongs.
- `control` is a character indicating which label has to be set as control class in a *supervised* analysis.
- `rm.xy` is a logical indicating if X or Y chromosome placed genes/proteins/features should be removed before run RDA (requires `id.type` and `annot` inputs).

All the rest of parameters are used to annotate features or to establish a parallel computation of processes, so they are explained within a longer and more detailed vignette included in the DECO R package.

5. NSCA, Non-Symmetrical Correspondence Analysis: *h* statistic

Once the frequency matrix of DE events or *incidenceMatrix* has been produced, DECO follows applying a NSCA [7] procedure. **NSCA** allows analyse all dependencies and covariances between differential features and samples placing them in the same relational space. Further information can be found in a more detailed vignette included in the DECO R package and also in our original publication [1].

As a measure of this significant association, NSCA function returns a *inner product* matrix relating feature-sample dependencies in the differential context. After the *inner product* matrix is generated, samples with similar profiles (using all the genes that gave DE events over a threshold: `pos.rep`) are grouped together using a hierarchical clustering based on Pearson correlation distances between samples: $dist_{ij} = 1 - corr(p_i, p_j)$.

Additionally, all different *agglomeration methods* to creates a dendrogram (see further information in `hclust` R function) are assessed looking for the method that shows highest *cophenetic correlation* [8]. Thus, we identify the best clustering procedure to make subclasses, choosing an optimal number of subclasses depending on the best *Hubber's Pearson* γ cutting this dendrogram.

5.1 Running the NSCA function: `decoNSCA()`

Here, we show an example of how user can run the second step of **DECO**:

```
# It can be applied to any subsampling design (SUPERVISED or UNSUPERVISED) and
# any transcriptomic platform (microarray, RNAseq, methylation, etc)
deco_results_ma <- decoNSCA(sub = sub, v = 80, method = "ward.D",
                            k.control = 3, k.case = 3, samp.perc = 0.05, rep.thr = 5,
                            parallel = T, cpus = 6, rm.complete = F, overlap = T)
```

Several important **input parameters** of this `decoNSCA()` function can be set up by user. Further information could be found in `?decoNSCA` help page. Finally, this function will return an output R object `deco` that is described in the following section.

6. Description of output results

```
slotNames(deco_results_ma)
[1] "featureTable"    "NSCAcluster"     "incidenceMatrix" "classes"
"pos.iter"        "control"         "q.val"          "subsampling.call"
"deco.call"
```

After running NSCA function *decoNSCA*, the method produces an R object of *deco* class. The main slots with relevant information inside this object are:

- `featureTable` is the main output table with the **feature statistics** and rankings.
- `NSCAcluster` contains the NSCA information and sample subclasses. It will be duplicated if a *supervised* analysis is run.
- `incidenceMatrix` is the **Absolute frequency matrix** with DE events per sample used in the NSCA.
- Vector of *classes* with labels per sample. For *unsupervised* analysis it will be **NA**.
- Label set as *control*.
- `q.val` is the adjusted.p.val threshold previously defined.
- `subsampling.call` and `deco.call` correspond to both `decoRDA` and `decoNSCA` function calls.

Feature ranking and statistics

The main output table with relevant feature information from RDA, NSCA and subclasses searching corresponds to *featureTable*.

```
dim(deco_results_ma@featureTable)
# Statistics of top-10 features
deco_results_ma@featureTable[1:10,]
```

The most relevant statistic derived from RDA technique is the *Standard.Chi.Square*. The amount of *differential events* or *Repeats* that each gene (each feature) appears differentially changed among classes or samples is also very important, and it is summarized in *Standard.Chi.Square* since this parameter weights the significance of the DE. Genes with similar *Repeats* values which correspond to lower *adj.p.value* resemble higher *Standard.Chi.Square* values, meanwhile genes with higher *adj.p.value*, or near *q.val* threshold imposed by user, give lower *Standard.Chi.Square* values.

IDs	Standard.Chi.Square	Repeats	adj.p.values	h.Range	Dendrogram.group
DE feature 1	250	100	~ 0.01	3.26	2
DE feature 2	150	100	~ 0.05	12.65	5

Moreover, for *supervised* analysis the *exprsUpDw* character indicates if case class shows UP or DOWN regulation of each feature. In some cases, several genes could follow deregulation in both classes for some subgroup of samples, which we called change-type **MIXED**. This kind of change pattern could explain some hidden characteristic of the samples and allows finding outliers: a subgroup of samples that only change in a subset of genes. In this cases there are not differences between the mean or median for the whole classes, and so classical methods like **SAM** or **LIMMA** do not find these patterns.

After RDA and NSCA analysis, the statistics referred to sample subclasses found is used to rank DE features properly. In this way, h statistic obtained per feature is used to determine how each feature discriminates each subclass found. As we mentioned above, this statistic combines both the DE changes and the predictor-response relationship between features and samples, so it refers to feature's discriminant ability. Furthermore, `Dendrogram.group` helps to identify to which pattern belongs each feature and each sample within the h statistic heatmap (`decoReport()` PDF report).

Sample subclasses membership

To see how samples are grouped into different *subclasses* within class:

```
# If SUPERVISED analysis
sampleSubclass <- rbind(deco_results_ma@NSCAcluster$Control$samplesSubclass,
                        deco_results_ma@NSCAcluster$Case$samplesSubclass)
# If UNSUPERVISED analysis
sampleSubclass <- deco_results_ma@NSCAcluster$All$samplesSubclass
## Sample subclass membership
head(sampleSubclass)
                  Subclass
GSM1607016 "pos Subclass 1"
GSM1607017 "pos Subclass 1"
GSM1607018 "pos Subclass 1"
GSM1607019 "pos Subclass 1"
GSM1607021 "pos Subclass 1"
GSM1607024 "pos Subclass 2"
```

Additionally, we can print a brief summary of DECO analysis using `summary` or `print` native R functions.

```
## Example of summary of a 'deco' R object (ALCL supervised/binary example)

summary(deco_results_ma)
# Decomposing Heterogeneous Cohorts from Omic profiling: DECO
# Summary:
# Analysis design: Supervised
# Classes compared:
# neg pos
#  20  11
#            RDA.q.value Minimum.repeats Percentage.of.affected.samples NSCA.variability
# Thresholds        0.01          10.00                            5.00               86
# Number of features out of thresholds: 297
# Feature profile table:
# Complete Majority Minority    Mixed
#       12       87      197        1
# Number of samples affected: 31
# Number of positive RDA comparisons: 1999
# Number of total RDA comparisons: 10000
```

An **extended report** (as PDF file) including more detailed information of the analysis and several plots illustrating all the results (as the bi-clustering approach to `h` statistic matrix) can be also produced with the

`decoReport()` R function. Information about the extended report is included in longer and more detailed vignette in the DECO R package.

7. Output reports

7.1 Generating a PDF report

DECO R package implements an additional function to help users to view and analyse the output results. It contains a detailed representation of main results (subclasses found, main biomarkers, h statistic heatmap, best feature profiles, feature's overlapping signal...). Here, we briefly describe how to run `decoReport()` R function:

```
### microarray results
decoReport(deco_results_ma, sub, pdf.file = "Report.pdf",
           info.sample = pData(ALCLdata)[,c("Type","Blood.paper"), drop = F],
           cex.names = 0.3, print.annot = T)
### rnaseq results
decoReport(deco_results_rnaseq, sub, pdf.file = "Report.pdf",
           info.sample = clinical.data.tcga[, 1:2, drop=F],
           cex.names = 0.3, print.annot = T)
```

A main result of DECO analysis is the h *statistic matrix* derived from both combination of RDA and NSCA information. In this way, `decoReport()` generates the **heatmap** representation of this h *statistic matrix* that includes a double correlation analysis between samples and between features and two derived clustering dendrograms. In this way the **heatmap** reveals subclasses of samples and feature patterns.

Further information about all plots included in the PDF report (`decoReport()`) can be found using *help*.

Heatmap (based on double correlation and clustering of h matrix) including features found.

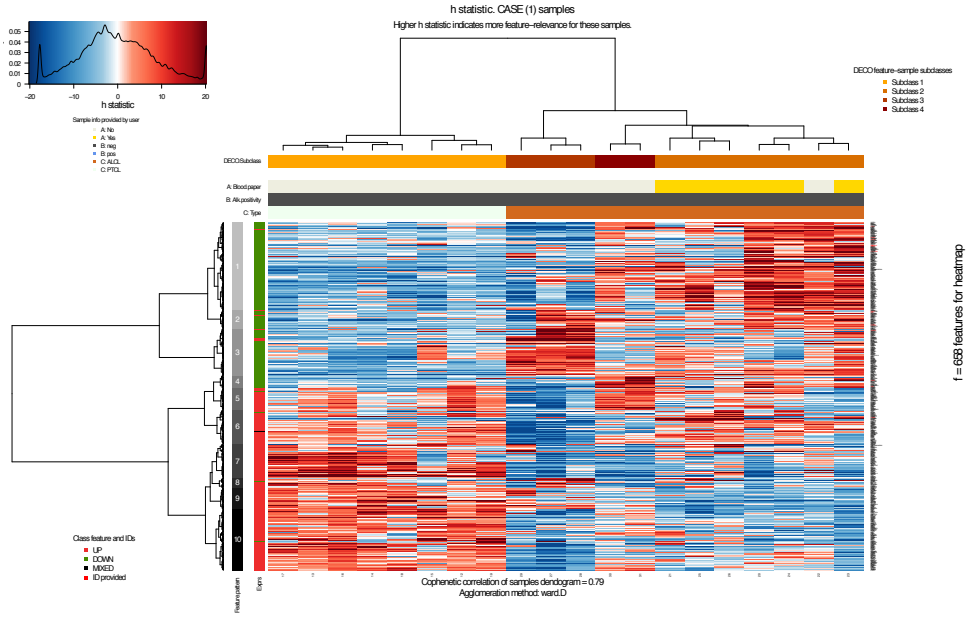


Figure 4: Heatmap of h statistic values from DECO, included in PDF generated by decoReport() function.

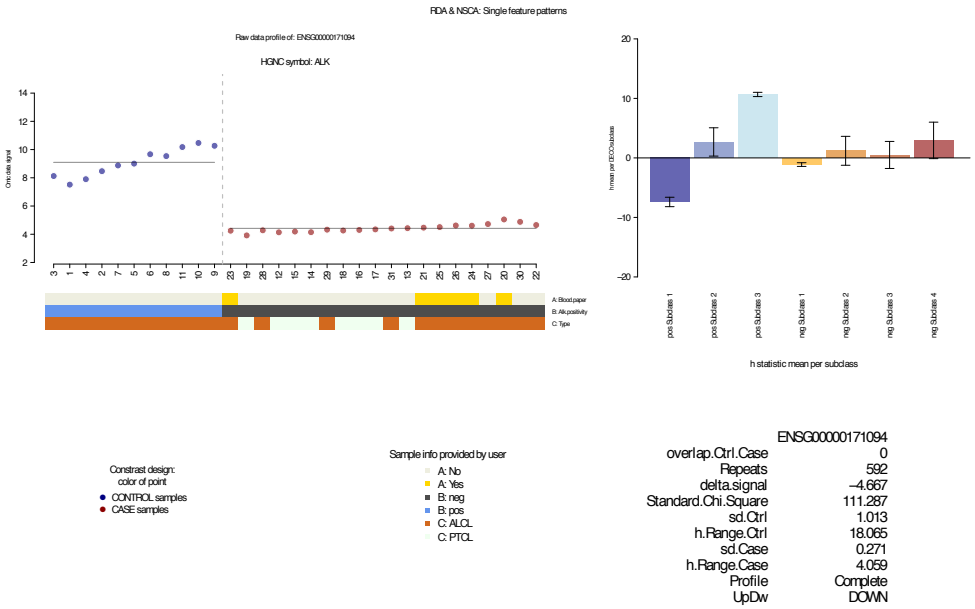
7.2 Generating plots and profiles for a specific feature (i.e. gene profiling)

Additionally, `plotDECOProfile()` R function provides a way to visualize a single feature profile.

Here, we show the examples for two genes discovered by Scarfo et al. [3] in the analysis of the ALCL samples: **(i) ALK**, that is the key gene-marker used by doctors to separate the two major subtypes of **Anaplastic Large Cell Lymphomas** (in the analysis done with DECO, this gene shows a change-profile *Complete* which supports the value of the gene to separate the ALCL samples); **(ii) ERBB4**, that was reported by authors as biomarker of a new subclass found inside the *negative* ALCL samples (in the analysis done with DECO, this gene shows a change-profile *Minority* indicating the existence of a subset of *negative* ALCL samples that are separated from the rest).

```
### ALK gene profile
plotDECOProfile(deco = deco, id = "ENSG00000171094",
                data = data, pdf.file = "ALK_ALCLdata.pdf",
                cex.samples = 2, info.sample = pData(ALCLdata)[,c(9,8,10)])
```

ALK gene: profile section

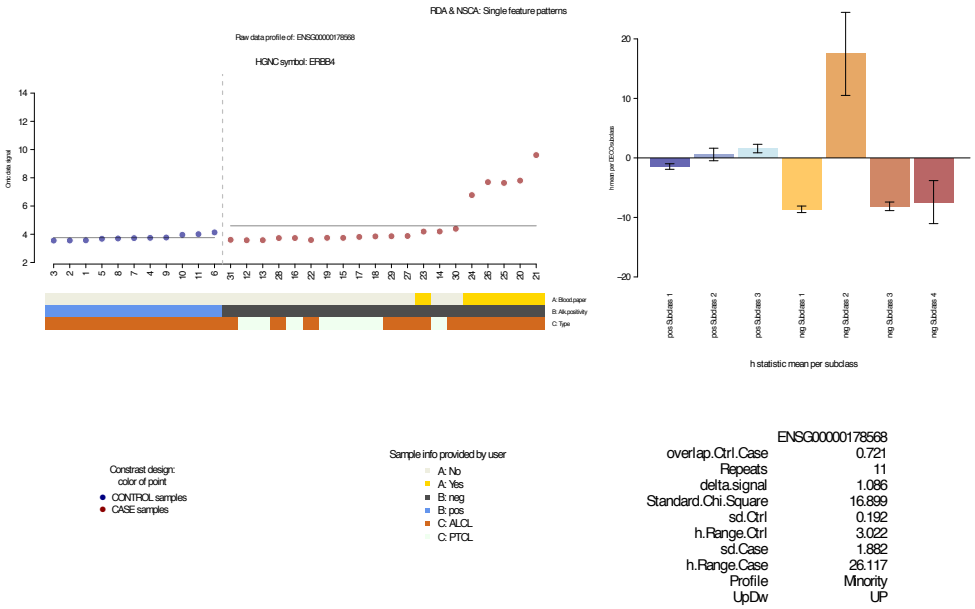


	ENSG00000171094
overlap.Ctrl.Case	0
Repeats	592
delta.signal	−4.667
Standard.Chi.Square	111.287
sd.Ctrl	1.013
h.Range.Ctrl	18.065
sd.Case	0.271
h.Range.Case	4.059
Profile	Complete
UpDw	DOWN

Figure 5: Search for the specific patterns for a feature within the profiles obtained with DECO. The figures correspond to ALK gene and include a plot of its raw expression along samples and a plot of the h statistic of this gene per subclass. This gene shows a change type COMPLETE. The h statistics per subclass found are large for the controls ("positive" ALCLs), and constant and close to 0 for the "negative" ALCLs.

```
### ERBB4 gene profile
plotDECOProfile(deco = deco, id = "ENSG00000178568",
                data = data, pdf.file = "ERBB4_ALCLdata.pdf",
                cex.samples = 2, info.sample = pData(ALCLdata)[,c(9,8,10)])
```

ERBB4 gene: profile section



	ENSG00000178568
overlap.Ctrl.Case	0.721
Repeats	11
delta.signal	1.086
Standard.Chi.Square	16.899
sd.Ctrl	0.192
h.Range.Ctrl	3.022
sd.Case	1.882
h.Range.Case	26.117
Profile	Minority
UpDw	UP

Figure 6: Search for the specific patterns for a feature within the omic profile derived from the RDA-NSCA results. The figures correspond to ERBB4 gene and include a plot of its raw expression along samples and a plot of the h statistic of this gene per subclasses. This gene shows a profile of change type MINORITY, that reveals a different behaviour for a subset of samples inside the "negative" ALCLs. The h statistics per subclass found in this case do not change for the controls (blue boxes corresponding to "positive" ALCL samples), but change a lot within the "negative" ALCL samples, indicating that is a clear marker of this group (segregating a subtype inside that corresponds to negative subclass 2) (see also Figure 3).

8. References

1: Campos-Laborie FJ, Risueño A, Roson-Burgo B, Droste C, Fontanillo C, Ortiz-Estevez M, Trotter MW, Sánchez-Santos JM and De Las Rivas J (2018). **Decomposing heterogeneous population cohorts for patient stratification and discovery of subclass biomarkers using omic data profiling.** Article submitted.

2: Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK (2015). **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res.*, 43:e47. doi:10.1093/nar/gkv007.

3: Scarfo I, Pellegrino E, Mereu E, Kwee I, Agnelli L, Bergaggio E, Garaffo G, Vitale N, Caputo M, Machiorlatti R, Circosta P, Abate F, Barreca A, Novero D, Mathew S, Rinaldi A, Tiacci E, Serra S, Deaglio S, Neri A, Falini B, Rabadan R, Bertoni F, Inghirami G, Piva R; European T-Cell Lymphoma Study Group (2016). **Identification of a new subclass of ALK-negative ALCL expressing aberrant levels of ERBB4 transcripts.** *Blood*, 127:221-232. doi:10.1182/blood-2014-12-614503.

4: Risueño A, Fontanillo C, Dinger ME, De Las Rivas J (2010). **GATExplorer: genomic and transcriptomic explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs.** *BMC Bioinformatics*, 11:221. doi:10.1186/1471-2105-11-221.

5: MacDonald JW and Ghosh D (2006). **COPA–Cancer Outlier Profile Analysis.** *Bioinformatics*, 22:2950-2951. doi:10.1093/bioinformatics/btl433

6: Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, McLellan M, Yau C, Kandoth C, Bowlby R, Shen H, Hayat S, Fieldhouse R, Lester SC, Tse GM, Factor RE, Collins LC, Allison KH, Chen YY, Jensen K, Johnson NB, Oesterreich S, Mills GB, Cherniack AD, Robertson G, Benz C, Sander C, Laird PW, Hoadley KA, King TA; TCGA Research Network., Perou CM (2015). **Comprehensive molecular portraits of invasive lobular breast cancer**. *Cell*, 163:505-519. doi:10.1016/j.cell.2015.09.033.

7: Lauro N and D'Ambra L (1984). **L'analyse non symetrique des correspondances.** In: *Data Analysis and Informatics III*, (Diday E, Jambu M, Lebart L, Pages J and Tomassone R Eds.). North Holland, Amsterdam, p.433-446.

8: Sokal RR and Rohlf FJ (1962). **The comparison of dendrograms by objective methods.** *Taxon*, 11, 33.

APPENDIX 2

R script to iterative calculate all ***cohesiveness*** statistics per feature within an omic dataset, based on two or more categories of samples.

```
################################################################################
################################################################################
#### Cohesiveness R script
####
# Author: Fco. Jose Campos-Laborie (CiC-IBMCC, USAL-CSIC)
# Salamanca, Spain.
# Date: July 16th, 2018
##
## Cohesiveness statistic is intended as a feature selection method
## to assess the probability of "being close" of any group of samples
## within a feature.

## R dependencies
library(metap)

#### INPUT:
# mx = matrix with rows as features and columns as samples
# cl = named vector with group of samples as values and names of samples as
# names of 'cl'.
# comb.p.val = "fisher" for Fisher's combined probability test or "z.score" for
# Stouffer's method.
# method = "complete" to consider all elements and "trimmed" to consider elements
# after removing outliers from both tails.
# trim = portion of outlier elements to remove for each category (from 0 to 1).


cohesiveness <- function(mx, cl, comb.p.val = "fisher",
                         method = "complete", trim = 0.01)
{
  require(metap)

  if(!method %in% c("complete","trimmed"))
    stop("'method' must be 'complete' or 'trimmed'")
  # ordering the matrix
  mx <- mx[,names(cl)]
  # counter for biological features
  counter <- 0

  # wrapper for all biological features
  res <- t(apply(mx, 1, function(x) {
    counter <<- counter + 1
    cat("\rFeature: ",counter)
    # ranking the feature
    x1 <- rank(-x, ties.method = "random")
    names(x1) <- names(x)

    res <- sapply(unique(as.character(cl)), function(y) {
      x2 <- x1
      cl2 <- cl
      ## Calculating gaps between elements of each group
```

```
      d <- diff(sort(x2[cl2 == y]))
      ## Average, 'r' and 'n'
      if(method == "complete"){
        avrg.samp <- mean(d)
        r <- length(x2[cl2 == y])
        n <- length(x2)
      }
      if(method == "trimmed"){
        avrg.samp <- mean(d, trim = trim)
        r <- length(x2[cl2 == y])
        n <- length(x2)
        r <- r - ceiling(trim*r*2)
        n <- n - ceiling(trim*r*2)
      }
      ## Calculating theoretical average and sd
      avrg <- (n+1)/(r+1)
      variance <- ((n+1)*(n-r)*r)/(((r+1)^2)*(r+2))
      std <- sqrt(variance) ## standard deviation
      ## Z-score
      z <- (avrg.samp - avrg)/(std/sqrt(r))

      return(c(z, pnorm(z)))})
  }))
  colnames(res) <- c(rbind(paste(unique(as.character(cl)),"cohesiveness",sep = "_"),
                           paste(unique(as.character(cl)),"p.value",sep = "_")))

  if(comb.p.val == "fisher"){
    ### Fisher's combination
    summ <- apply(res[,seq(2,dim(res)[2],2)], 1, function(x) -2*sum(log(x)))
    p.val <- 1 - pchisq(summ, df = dim(res)[2])
  }
  else if(comb.p.val == "z.score"){
    ### Stouffer's Z-score
    summ <- t(apply(res[,seq(2,dim(res)[2],2)], 1, function(x)
unlist(sumz(x)[c("z","p")])))
    p.val <- summ[,"p"]
    summ <- summ[,"z"]
  }
  ## Multiple correction of p-values
  fdr <- p.adjust(p.val, method = "fdr")

  res <- data.frame(res, summ, p.val, fdr)
  return(res)
}
```

APPENDIX 3

Manual categorization of frequent PSI-MI methods into meta-groups to produce a non-redundant literature-based interactome. Deleted PSI-MI methods correspond to non-reliable methods.

MI_ID	Name of the PSI-MI method	PSI-MI TYPE	Meta-group PSI-MI (proposed)	If the PPIs come "from the same PubMed (PMID)" we unify to the meta-group:
8	array technology	binary	8	MI:0008 array technology
81	peptide array	binary	8	MI:0008 array technology
89	protein array	binary	8	MI:0008 array technology
95	proteinchip(r) on a surface-enhanced laser desorption/ionization	binary	8	MI:0008 array technology
678	antibody array	binary	8	MI:0008 array technology
921	surface plasmon resonance array	binary	8	MI:0008 array technology
18	two hybrid	binary	18	MI:0018 two-hybrid
397	two hybrid array	binary	18	MI:0018 two-hybrid
398	two hybrid pooling approach	binary	18	MI:0018 two-hybrid
399	two hybrid fragment pooling approach	binary	18	MI:0018 two-hybrid
655	lambda repressor two hybrid	binary	18	MI:0018 two-hybrid
726	reverse two hybrid	binary	18	MI:0018 two-hybrid
727	lexa b52 complementation	binary	18	MI:0018 two-hybrid
728	gal4 vp16 complementation	binary	18	MI:0018 two-hybrid
1112	two hybrid prey pooling approach	binary	18	MI:0018 two-hybrid
1113	two hybrid bait and prey pooling approach MI:1113	binary	18	MI:0018 two-hybrid
1203	split luciferase complementation	binary	18	MI:0018 two-hybrid
2215	barcode fusion genetics two hybrid	binary	18	MI:0018 two-hybrid
112	ubiquitin reconstruction	binary	18	MI:0018 two-hybrid
232	transcriptional complementation assay	binary	18	MI:0018 two-hybrid
369	lex-a dimerization assay	binary	18	MI:0018 two-hybrid
1356	validated two hybrid	binary	18	*MI:0018 two-hybrid*
30	cross-linking study	binary	30	*MI:0030 cross-linking*
31	protein cross-linking with a bifunctional reagent	binary	30	MI:0030 cross-linking

34	display technology	binary	**34**	MI:0034 display technology
48	filamentous phage display	binary	**34**	MI:0034 display technology
66	lambda phage display	binary	**34**	MI:0034 display technology
84	phage display	binary	**34**	MI:0034 display technology
108	t7 phage display	binary	**34**	MI:0034 display technology
115	yeast display	binary	**34**	MI:0034 display technology
47	far western blotting	binary	**47**	
12	bioluminescence resonance energy transfer	binary	**51**	MI:0051 fluorescence technology
55	fluorescent resonance energy transfer	binary	**51**	MI:0051 fluorescence technology
905	amplified luminescent proximity homogeneous assay	binary	**51**	MI:0051 fluorescence technology
1016	fluorescence recovery after photobleaching	binary	**51**	MI:0051 fluorescence technology
52	fluorescence correlation spectroscopy	binary	**52**	MI:0052 fluorescence spectroscopy
53	fluorescence polarization spectroscopy	binary	**52**	MI:0052 fluorescence spectroscopy
65	isothermal titration calorimetry	binary	**65**	
77	nuclear magnetic resonance	binary	**77**	
10	beta galactosidase complementation	binary	**90**	MI:0090 protein complementation assay
11	beta lactamase complementation	binary	**90**	MI:0090 protein complementation assay
90	protein complementation assay	binary	**90**	MI:0090 protein complementation assay
111	dihydrofolate reductase reconstruction	binary	**90**	MI:0090 protein complementation assay
231	mammalian protein protein interaction trap	binary	**90**	MI:0090 protein complementation assay
370	tox-r dimerization assay	binary	**90**	MI:0090 protein complementation assay
809	bimolecular fluorescence complementation	binary	**90**	MI:0090 protein complementation assay
1037	split renilla luciferase complementation	binary	**90**	MI:0090 protein complementation assay
1204	split firefly luciferase complementation	binary	**90**	MI:0090 protein complementation assay
1235	thermal shift binding	binary	**90**	MI:0090 protein complementation assay
99	scintillation proximity assay	binary	**99**	MI:0099 scintillation proximity assay
425	kinase scintillation proximity assay	binary	**99**	MI:0099 scintillation proximity assay
107	surface plasmon resonance	binary	**107**	
114	x-ray crystallography	binary	**114**	
411	enzyme linked immunosorbent assay	binary	**411**	
417	footprinting	binary	**417**	MI:0417 footprinting
605	enzymatic footprinting	binary	**417**	MI:0417 footprinting
814	protease accessibility laddering	binary	**417**	MI:0417 footprinting
440	saturation binding	binary	**440**	
729	luminescence based mammalian interactome mapping	binary	**729**	
813	proximity enzyme linked immunosorbent assay	binary	**813**	
888	small angle neutron scattering	binary	**888**	

1	interaction detection method MI:0001	**deleted**	**1**	
21	colocalization by fluorescent probes cloning	**deleted**	**21**	
22	colocalization by immunostaining	**deleted**	**22**	
23	colocalization/visualisation technologies	**deleted**	**23**	
403	colocalization	**deleted**	**25**	
25	copurification	**deleted**	**105**	
10023	co-fractionation	**deleted**	**256**	
105	**structure based prediction**	**deleted**	**260**	
260	inhibitor small molecules	**deleted**	**330**	
330	**molecular source**	**deleted**	**339**	
339	**undetermined sequence position**	**deleted**	**363**	
492	**in vitro**	**deleted**	**364**	
493	**in vivo**	**deleted**	**403**	
10018	protein-peptide	**deleted**	**418**	
10020	affinity capture-RNA	**deleted**	**492**	
10021	protein-RNA	**deleted**	**493**	
256	RNA interference	**deleted**	**686**	
1017	RNA immunoprecipitation	**deleted**	**1017**	
363	**inferred by author**	**deleted**	**10018**	
364	**inferred by curator**	**deleted**	**10020**	
418	**genetic**	**deleted**	**10021**	
686	**unspecified method**	**deleted**	**10023**	
4	affinity chromatography technology	indirect	**4**	
6	anti bait coimmunoprecipitation	indirect	**6**	
7	anti tag coimmunoprecipitation	indirect	**7**	
13	biophysical	indirect	**13**	
16	circular dichroism	indirect	**16**	
17	classical fluorescence spectroscopy	indirect	**17**	
19	coimmunoprecipitation	indirect	**19**	
20	transmission electron microscopy	indirect	**20**	
27	cosedimentation	indirect	**27**	
28	cosedimentation in solution	indirect	**28**	
29	cosedimentation through density gradient	indirect	**29**	
38	dynamic light scattering	indirect	**38**	
40	electron microscopy	indirect	**40**	
42	electron paramagnetic resonance	indirect	**42**	
45	experimental interaction detection	indirect	**45**	

49	filter binding	indirect	**49**	
51	fluorescence technology	indirect	**51**	
54	fluorescence-activated cell sorting	indirect	**54**	
67	light scattering	indirect	**59**	
69	mass spectrometry studies of complexes	indirect	**61**	
71	molecular sieving	indirect	**67**	
91	chromatography technology	indirect	**69**	
96	pull down	indirect	**71**	
97	reverse ras recruitment system	indirect	**91**	
104	static light scattering	indirect	**96**	
226	ion exchange chromatography	indirect	**97**	
227	reverse phase chromatography	indirect	**104**	
254	genetic interference	indirect	**226**	
257	antisense rna	indirect	**227**	
276	blue native page	indirect	**254**	
400	affinity technology	indirect	**257**	
401	biochemical	indirect	**276**	
402	chromatin immunoprecipitation assays	indirect	**400**	
404	comigration in non denaturing gel electrophoresis	indirect	**401**	
405	competition binding	indirect	**402**	
406	deacetylase assay	indirect	**404**	
410	electron tomography	indirect	**405**	
412	electrophoretic mobility supershift assay	indirect	**406**	
413	electrophoretic mobility shift assay	indirect	**410**	
415	enzymatic study	indirect	**412**	
416	fluorescence microscopy	indirect	**413**	
419	gtpase assay	indirect	**415**	
420	kinase homogeneous time resolved fluorescence	indirect	**416**	
423	in-gel kinase assay	indirect	**419**	
424	protein kinase assay	indirect	**420**	
426	light microscopy	indirect	**423**	
428	imaging techniques	indirect	**424**	
434	phosphatase assay	indirect	**426**	
435	protease assay	indirect	**428**	
437	protein tri hybrid	indirect	**434**	
510	homogeneous time resolved fluorescence	indirect	**435**	
512	zymography	indirect	**437**	
515	methyltransferase assay	indirect	**510**	

516	methyltransferase radiometric assay	indirect	**512**	
588	3 hybrid method	indirect	**515**	
663	confocal microscopy	indirect	**516**	
676	tandem affinity purification	indirect	**588**	
807	comigration in gel electrophoresis	indirect	**663**	
808	comigration in sds page	indirect	**676**	
825	x-ray fiber diffraction	indirect	**807**	
826	x ray scattering	indirect	**808**	
841	phosphotransferase assay	indirect	**825**	
858	immunodepleted coimmunoprecipitation	indirect	**826**	
859	intermolecular force	indirect	**841**	
870	demethylase assay	indirect	**858**	
872	atomic force microscopy	indirect	**859**	
880	atpase assay	indirect	**870**	
889	acetylase assay	indirect	**872**	
892	solid phase assay	indirect	**880**	
893	neutron diffraction	indirect	**889**	
894	electron diffraction	indirect	**892**	
920	ribonuclease assay	indirect	**893**	
943	detection by mass spectrometry	indirect	**894**	
944	mass spectrometry study of hydrogen/deuterium exchange	indirect	**920**	
947	bead aggregation assay	indirect	**943**	
949	gdp/gtp exchange assay	indirect	**944**	
953	polymerization	indirect	**947**	
963	interactome parallel affinity capture	indirect	**949**	
964	infrared spectroscopy	indirect	**953**	
968	biosensor	indirect	**963**	
969	bio-layer interferometry	indirect	**964**	
982	electrophoretic mobility-based method	indirect	**968**	
990	cleavage assay	indirect	**969**	
991	lipoprotein cleavage assay	indirect	**979**	
997	ubiquitinase assay	indirect	**982**	
1000	hydroxylase assay	indirect	**984**	
1010	neddylase assay	indirect	**990**	
1019	protein phosphatase assay	indirect	**991**	
1022	field flow fractionation	indirect	**997**	
1024	scanning electron microscopy	indirect	**998**	

1038	silicon nanowire field-effect transistor	indirect	**1000**	
1086	equilibrium dialysis	indirect	**1005**	
1103	solution state nmr	indirect	**1008**	
1104	solid state nmr	indirect	**1010**	
1147	ampylation assay	indirect	**1019**	
1246	ion mobility mass spectrometry of complexes	indirect	**1022**	
1247	mst(micro-scale thermophoresis)	indirect	**1024**	
1311	differential scanning calorimetry	indirect	**1038**	
1313	proximity labelling technology	indirect	**1086**	
1314	proximity-dependent biotin identification	indirect	**1103**	
2189	avexis(avidity-based extracellular interaction screening)	indirect	**1104**	
979	oxidoreductase assay MI:0979	indirect	**1138**	
984	deamination assay MI:0984	indirect	**1147**	
998	deubiquitinase assay MI:0998	indirect	**1246**	
1005	adp ribosylase assay MI:1005	indirect	**1247**	
1008	sumoylase assay MI:1008	indirect	**1311**	
1138	decarboxylation assay MI:1138	indirect	**1313**	
1354	lipase assay MI:1354	indirect	**1314**	
59	gst pull down	indirect	**1354**	
61	his pull down	indirect	**2189**	

LIST OF PUBLICATIONS

Several scientific studies have been published in high-quality journals during this PhD project. Two of these publications are directly related with Chapter I and IV of this PhD dissertation, while other Chapters II and III are still in development. All these original publications are attached in the CD, included at the back of this Dissertation.

Scientific record: https://scholar.google.es/citations?hl=es&user=ZzuHB4cAAAAJ

ORCID: http://orcid.org/0000-0002-1213-0457

Main publications associated with this PhD:

Campos-Laborie FJ, Risueño A, Ortiz-Estévez M, Rosón-Burgo B, Droste C, Fontanillo C, Loos R, Sánchez-Santos JM, Trotter MW and De Las Rivas J. (2018) **DECO: decompose heterogeneous population cohorts for patient stratification and discovery of subclass biomarkers using omic data profiling**. *Bioinformatics*. Under revision.

Lopes KP, Campos-Laborie FJ, Vialle RA, Ortega JM and De Las Rivas J (2016). **Evolutionary hallmarks of the human proteome: chasing the age and coregulation of protein-coding genes.** *BMC Genomics* 17(Suppl 8): 725. PMID: 27801289.

Publications where algorithm DECO (Chapter I) was used:

Del Rey M, Benito R, Fontanillo C, Campos-Laborie FJ, Janusz K, Velasco-Hernandez T, Abaigar M, Hernandez M, Cuello R, Borrego D, Martin-Zanda D, De Las Rivas J, Mills KI, Hernandez-Rivas JM (2015). **Deregulation of genes related to iron and mitochondrial metabolism in refractory anemia with ring sideroblasts.** PLoS One 10(5): e0126555. PMID: 25955609.

Zandueta C, Ormazabal C, Perurena N, Martinez-Canarias S, Zalacain M, Julian MS, Grigoriadis AE, Valencia K, Campos-Laborie FJ, De Las Rivas J, Vicent S, Patino-Garcia A, Lecanda F (2016). **Matrix-Gla protein promotes osteosarcoma lung metastasis and associates with poor prognosis**. *J Pathol* 239(4): 438-449. PMID: 27172275

Additional publications:

Aibar S, Abaigar M, Campos-Laborie FJ, Sanchez-Santos JM, Hernandez-Rivas JM and De Las Rivas J (2016). **Identification of expression patterns in the progression of disease stages by integration of transcriptomic data.** *BMC Bioinformatics* 17(Suppl 15): 432. PMID: 28185568.

Aibar S, Fontanillo C, Droste C, Roson-Burgo B, Campos-Laborie FJ, Hernandez-Rivas JM and De Las Rivas J (2015). **Analyse multiple disease subtypes and build associated gene networks using genome-wide expression profiles**. *BMC Genomics* 16(Suppl 5): S3. PMID: 26040557.

De Las Rivas J, Bonavides-Martinez C and Campos-Laborie FJ (2017). **Bioinformatics in Latin America and SolBio impact, a tale of spin-off and expansion around genomes and protein structures**. *Brief Bioinform.* 18(6):1091. PMID: 28968628.

GOBIERNO
DE ESPAÑA

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES