Methodological Advances in the Functional Profiling of Genomic Studies

David Montaner González

Departamento de Bioquímica y Biología Molecular
Universidad de Valencia
Programa de Doctorado en Biotecnología

Tésis dirigida por Joaquín Dopazo

Mayo 2013

To my Dad.

Acknowledgments

I would like to thank Joaquín Dopazo for supervising this thesis and for giving me the chance to join the Bioinformatics Department at CIPF where I developed it.

I also want to acknowledge all my colleagues from the Department, for their continuous support and for creating a wonderful research atmosphere, specially to Ignacio Medina.

I am grateful to Francisco Estruch from the University of Valencia for its advice while organizing this work.

I praise Blanca and Diego and all my family for their understanding during the time I stole from them to write.

And finally, I should give thanks to all parents, the greatest funding body of the Spanish science.

Abstract

In this thesis we present bioinformatic tools and algorithms for the analysis of genomic data such as those generated by microarray devices or next generation sequencing techniques. Particularly, we develop new approaches to gene set analysis. The described procedures should be useful in practice to tackle complex biological experiments, but hopefully will also be methodologically relevant, as they introduce new ways of conceptualizing genomic functional profiling.

Our very flexible approach allows for the inclusion of not just one kind of genomic measurement but many. It makes possible, for instance, to analyze expression measurement and genomic variation data at a time. This multidimensional gene set analysis approach is able to unravel genomic interactions that coordinately regulate functional blocks.

We also indicate how to use data available in public repositories to asses gene importance within gene sets. Such importance can be included into our algorithms as a weight, improving performance of the analysis. But, more interestingly, it models functional blocks as non discrete entities, featuring a new concept of fuzzy gene set.

Contents

List of Figures

Glossary

Babelomics: An integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling methods.
http://www.babelomics.org

CIPF: Centro de Investigación Príncipe Felipe. The research center in Valencia where I developed this thesis.
http://www.cipf.es
http://bioinfo.cipf.es (The web of the Bioinformatics Department)

FatiGO: A bioinformatic tool for *functional enrichment analysis*. Is currently Pratt of the Babelomics suite.

FatiScan: A bioinformatic tool for *gene set analysis* (GSA). Is currently part of the Babelomics suite.

Functional Enrichment Analysis: One of the most basic functional profiling methodologies. It is carried out in two steps, in the first one some genes are *selected* according so some biological property; in the second step, the database information is explored just in the selected genes. The method is also called "Over-Representation Analysis" by some authors.

Functional Profiling: The interpretation genomic experimental results in therms of the information already available in biological databases.

Gene Set: Also referred to in this work as *functional block*. A group of genes that is supposed to perform a biological function. Usually these genes will be tagged with the same label in some biological database.

GEO: Gene Expression Omnibus. A public data repository for microarray and sequence data hold by the NCBI.
http://www.ncbi.nlm.nih.gov/geo

GEPAS: Gene Expression Pattern Analysis Suit. As a web-based tool for the analysis of genomic data.
http://www.gepas.org

GO: Gene Ontology. A controlled vocabulary of terms for describing gene product characteristics.
http://www.geneontology.org

GSA: Gene Set Analysis. A general terminology to refer to functional profiling methods that analyze all genes available in the dataset, without any prior filtering or gene *selection* step.

GSEA: Gene Set Enrichment Analysis. The original *gene set method* method used in Mootha et al. 2003 and fully developed in Subramanian et al. 2005.

Interactome: The network defined by all the known interactions between pairs of proteins.

InterPro: An integrated database of predictive protein signatures.
http://www.ebi.ac.uk/interpro

KEGG: Kyoto Encyclopedia of Genes and Genomes. A database for molecular-level information of the biological systems such as the cell.
http://www.genome.jp/kegg

Microarray: Devices that allow for the measurement of genomic characteristics like gene expression levels, SNP variants or copy number alterations.

NCBI: The National (USA) Center for Biotechnology Information.
http://www.ncbi.nlm.nih.gov

NGS: Next Generation Sequencing, also known as High throughput sequencing. A technique for measuring genomic characteristics.

R: R is a free software environment for statistical computing and graphics.
http://www.r-project.org

REACTOME: An open-source, open access, manually curated and peer-reviewed pathway database.
http://www.reactome.org

SNP: Single Nucleotide Polymorphism.

Chapter 1

Introduction

This thesis summarizes some conceptual, methodological and computational advances I developed in the field of genomic data analysis. The main objective of my work was to improve what, in recent years, as been known as *gene set enrichment analysis*, *gene set analysis*, *functional annotation analysis* or *functional profiling*. That is, procedures for the interpretation of experimental genomic data which take advantage of the knowledge already available in biological databases.

The work here exposed was mostly developed in three scientific publications: **Montaner** et al. 2006, **Montaner** et al. 2009 and **Montaner** and Dopazo 2010. These three papers are embedded within chapters of the thesis, presenting the motivation for the research besides the achieved solutions. The remaining sections of the text aim to clarify the context in which the articles where written. I hope, these intermediate chapters will help the reader understanding the improvements introduced as a unique and coherent piece of work.

This first introductory chapter advances the work carried out and explains the key concepts which link together the three publications above mentioned. Section 1.1 presents general concepts in the analysis of genomic data and substantiates the need for *functional profiling* methodologies. Section 1.2 describes some of the most widely used genomic databases and shows how the available data are modeled and conceptualized in terms of *gene sets*. Section 1.3 introduces the general goal of *gene set analysis* methodologies. Section 1.4 presents the statistical framework of the *logistic regression* models; these models provide the analytical methodologies upon which my *functional profiling* algorithms rely on. Section 1.5 shows how weights can be included into the *logistic regression model*. In this same section I explain how the weights can account for relevant genomic information and the advantages of introducing them into the analysis. In section 1.6 the framework of the *logistic models* is extended to several dimensions; it is explained how multidimensional *gene set tests* can be carried out and why this may be suitable in some experimental contexts.

Chapter 2 displays the article **Montaner** et al. 2006. This paper presented tools and methodologies I developed for the analysis of genomic data. Chapter 4 includes a copy of the paper **Montaner** et al. 2009. This paper presented a novel way of conceptualizing *gene sets* as *non-discrete* entities. It also showed how to take advantages of this new paradigm using the weighting schema of the *logistic regression models*. Chapter 5 reproduces the publication **Montaner** and Dopazo 2010. In This work we developed the first methodology which can handle several kinds of experimental genomic data and analyze them in a multidimensional context.

Chapter 3 clarifies the rationale that took me from the general context of genomic studies presented in the first publication to the more specific algorithms developed in the two subsequent articles.

1.1 Genomic data analysis

Just in a decade, genomic scale measurement devices have completely changed biological research and clinical practice. DNA microarrays (Lockhart et al., 1996; Schena et al., 1996), have rapidly evolved from homemade gadgets to highly accurate biomedical instruments. Manufactured arrays like those of Affymetrix, Agilent or Illumina have lowered down prices and expanded the usage of their technologies to all areas of genomic research. Gene expression, copy number alterations or single nucleotide polymorphism (SNP) can now be measured at a genomic scale, providing adequately reproducible results (Shi et al., 2006). But microarray devices have not just been restricted to basic biological research. They have been successfully used in clinical diagnosis (Glas et al., 2006), and it has been proven that accurate and reproducible predictive models can be built up based on genomic measurements (Shi et al., 2010).

More recently, *high throughput sequencing* (Church, 2006; Hall, 2007), also known as *next generation sequencing* (NGS), has incredibly extended the potentiality of genomic research. Discovering new transcripts or isoforms, finding not previously reported genomic variants and pointing out novel miRNAs is now possible because the "universe" of genomic features that can be explored by means of NGS technologies is "not closed" or pre-defined, as it is in the context of microarrays.

Thus, in this very short period of time, we have had the chance of addressing biological complexity to an extent that was not even dreamed before. As a counterpart, biological research has become a discipline highly reliant on computational and statistical methodologies. The large amount of data generated by high throughput technologies requires not only highly sophisticated computational infrastructures, but also, optimized software and customized protocols. The variety of the information included in such datasets has also challenged experiment designers, as the possibility of screening samples at a genomic level induces new ways of conceptualizing biology and establishing hypotheses. Such cutting edge scenario entails the necessity of developing new statistical methods and analytical tools. The application of the same statistical test to thousands of genes extremely increases the probability of false positives, driving us to draw erroneous conclusions if care is not taken (Dopazo, 2009). Thus, specialized methods for p-value correction, such as those developed in Benjamini and Hochberg 1995 or Carvajal-Rodriguez and de Una-Alvarez 2011, needed to be included as a compulsory step in all genomic analyses. Missing data, scanning artifacts or very low signals required pre-processing and normalization routines particularly shaped for genomic technologies; see Troyanskaya et al. 2001, Irizarry et al. 2003 and Lin et al. 2008 respectively. Noisy but highly correlated data compel us to undertake gene selection steps before the information can be accurately used in clinical class prediction. Then, estimation of classification errors must take into account this gene selection step in order to avoid biases that do not rise in other machine learning applications; see Medina et al. 2007 for details.

In such novel and complex context GEPAS was developed. The acronym stands for: Gene Expression Pattern Analysis Suit. It is a web application mainly devised for the analysis of genomic data. It includes several modules used for analyzing microarrays, nevertheless, most of its tools are general purpose and may be applied in many other "omic" contexts.

The first release of GEPAS, described in Herrero et al. 2003, was basically a compendium of *unsupervised* classification algorithms such as *aggregative hierarchical* methods, *self organizing maps*, or the *self organizing tree algorithm*. All this *clustering* algorithms aimed to find gene groups with similar expression patterns across samples. At that time, biologists were thoroughly engaged in describing gene functionality and gene clustering was an invaluable tool for such endeavor. That first version of the software also contained very basic utilities for data preprocessing, class comparison and class prediction, thereby establishing the working lines we developed in subsequent years, see figure 1.1 on page 5. Another remarkable characteristic of the Herrero et al. 2003 approach was to point out the suitability of counting on tools for the interpretation of experimental results in the light of already established biological knowledge. In that sense, GEPAS allowed for the redirection of the "relevant" genes to a different tool called FatiGO, (Al-Shahrour et al., 2004, 2007b). FatiGO lets the user testing for the "enrichment" of Gene Ontology (GO) terms within a list of genes previously selected. This was the first functional profiling module included in our analysis pipeline and, indeed, one of the most successful tools used in bioinformatics.

The two following versions of the software added new functionalities

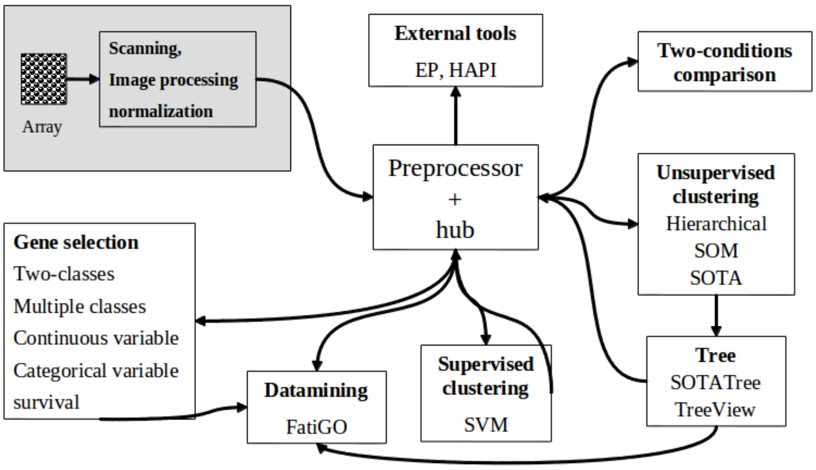


Figure 1.1: GEPAS pipeline in Herrero et al. 2003.

or included new methods in the already available ones. Herrero et al. 2004 included options for the normalization of spotted microarrays, see figure 1.2 on page 5.

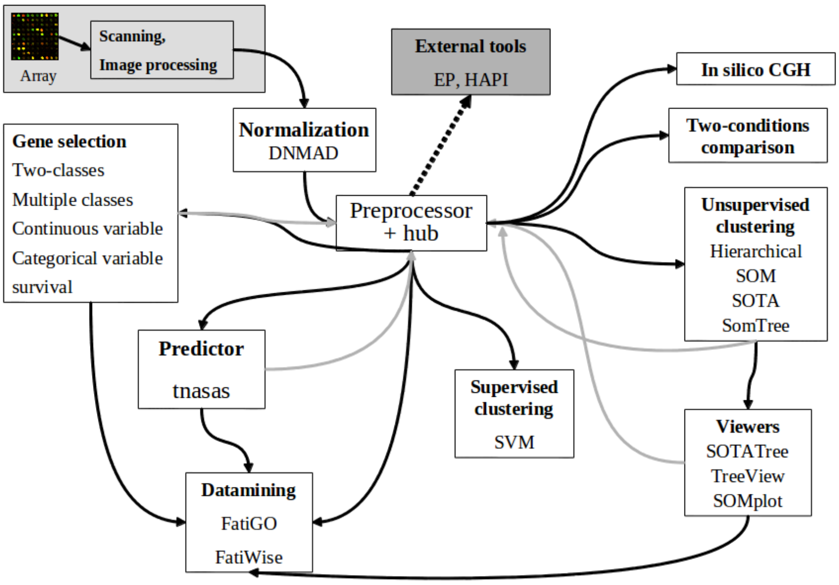


Figure 1.2: GEPAS pipeline in Herrero et al. 2004.

Vaquerizas et al. 2005 reinforced all the methodologies for the differential expression analysis and added the possibility of linking some GEPAS results with the tool FatiScan, designed for carrying out *gene set analysis* (GSA), see figure 1.3 on page 6. Nevertheless, up to that time, the interconnection between the different modules of GEPAS was limited. The suite was really a collection of several separated web pages administered as independent resources. GEPAS gathered tools like DNMAD (Vaquerizas et al., 2004) for the normalization of spotted microarrays, POMELO (Morrissey and Diaz-Uriarte, 2009) for the analysis of differential expression or InSilicoCGH (Vaquerizas et al., 2005), for the visualization of altered genomic regions.

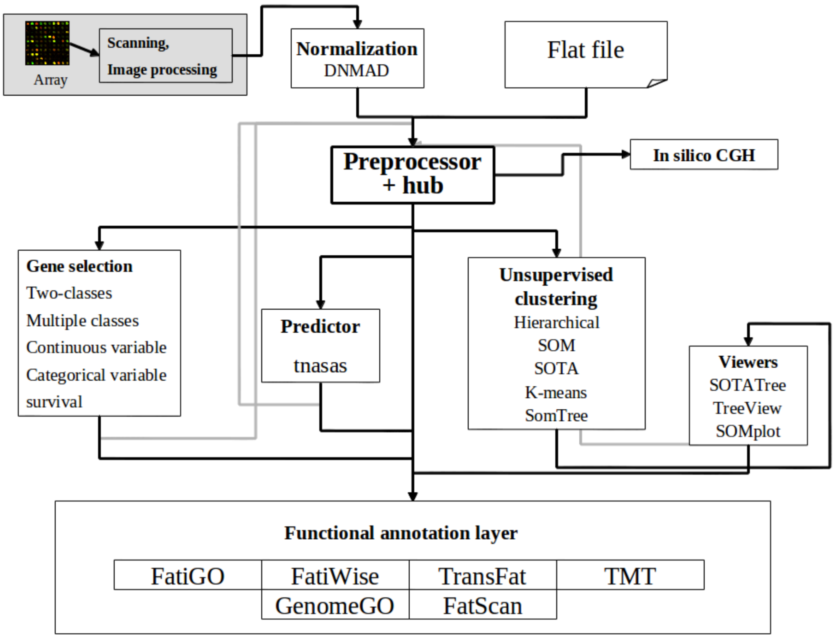


Figure 1.3: GEPAS pipeline in Vaquerizas et al. 2005.

It was in the release we did in **Montaner** et al. 2006, when GEPAS became a truly integrated platform. The software code was completely rewritten to make it more efficient and robust but, more importantly, the analysis pipeline was completely redesigned, see figure 1.4 on page 8. The aim was to offer a web platform able to deal with all kind of genomic studies. We wanted GEPAS to be powerful enough to cover any exhaustive analysis of microarray data; from raw data preprocessing to gene prioritization. We added normalization procedures for most common microarray platforms at that time, GPR, Agilent, Affymetrix . . . In this way, any gene expression study carried out using standard microarray devices, could be handled, preprocessed and shaped in the most convenient format to be used in all desired GEPAS analysis modules. We included most popular differential gene expression tests available at that time, and developed some others, like the one presented in Valls et al. 2008, for specific case studies. For the first time, GEPAS provided users with tools not only for clustering genes, but also samples or microarrays, thus meeting a growing demand for a change in the "direction" of such analyses. Several new supervised classification algorithms were included in the prediction module and methods implemented for the differential expression analysis were reused in this utility for the gene selection steps. In this way, every piece of the new GEPAS was a part of the same integrated tool, optimized and utilized in any analysis module which required it.

But we also wanted GEPAS to be flexible enough so that the user could combine any set of methods desired for the overall analysis. The only restriction we would put, of course, was that the analysis made "biological" sense.

Achieving such goal of usability was a task not less challenging than that of code integration and optimization. The web page, that is, our user interface, was redesigned to allow for the sought elasticity. The outcome of one analysis procedure could then be redirected and used as an input for all other tools in GEPAS, provided that the operation was meaningful. Analyses which, in previous releases of GEPAS, were made in an apparent single step, were now split in many, resulting in a much more operative toolkit. The clustering algorithms, for instance, were first optimized for any data matrix. Then they were applied to rows or columns of a file, depending on whether the user wanted to cluster genes or conditions. Those algorithms would not produce a tree plot straight forward, their output was a text file in *newick* format; a standard for coding tree shaped structures. It would be afterward, when

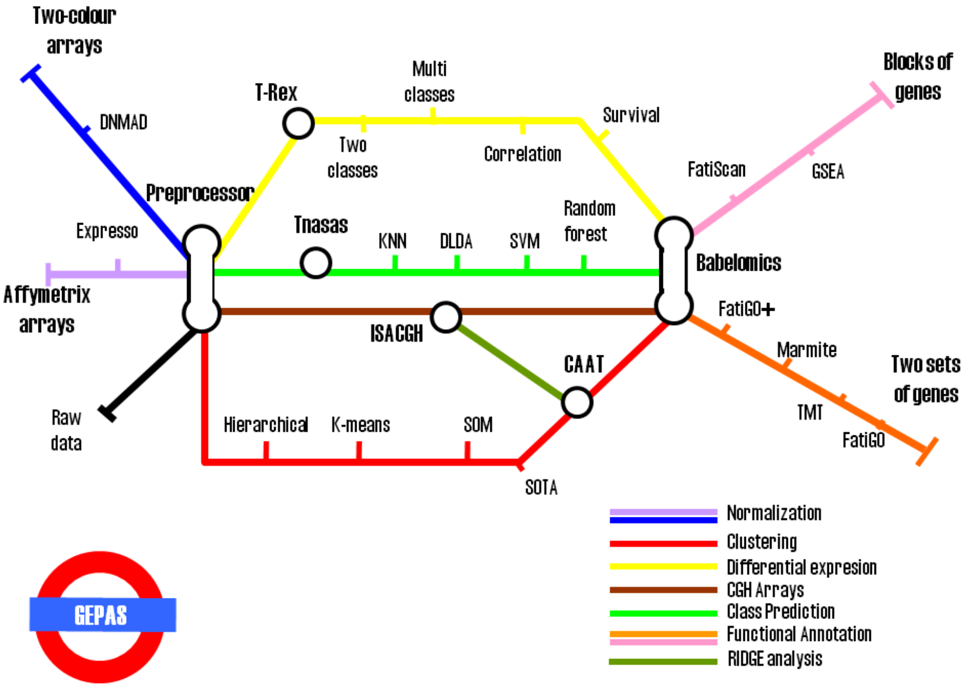


Figure 1.4: GEPAS pipeline in **Montaner** et al. 2006.

an interactive display tool[1] will draw the graph. In such way we allowed users of GEPAS to utilize independently our clustering methods and our drawing tools. Anyone could, for instance, calculate a tree for their microarrays and then, use a different software to display the structure or, simply, select a cluster and carry on an analysis we never thought about. Reversely, someone could have computed a clustering of their genes using a tool not included in GEPAS, and then use our displaying methods.

Thus, when GEPAS version presented in **Montaner** et al. 2006 was launched, we had implicitly developed a step by step methodology for the analysis of genomic data. A non linear pipeline which could be used partially or in its totality by any researcher, without the requirement of strong computational skills.

[1] In **Montaner** et al. 2006 this tool is referred as CAAT but it was later renamed as ETE and published as an independent tool in Huerta-Cepas et al. 2010.

The usefulness of GEPAS as a bioinformatic tool is supported by its wide usage in biological publications such as Montero-Conde et al. 2008, van Heerden et al. 2009 or Jantus Lewintre et al. 2009.

But beyond the recognition of the scientific community, the major revenue we got from developing GEPAS was, precisely, to realize the advantages of addressing complex genomic data analyzes in a modular way. We proved the advantage of splitting the problem in general pieces and tackle each of them step by step, creating the necessary data objects, the methods and the tool to handle them independently of the overall purpose of the research. Such approach allowed us to recycle many of the methods or tools we developed for a particular purpose, and, lately, take advantage of them in new contexts of analysis.

With such philosophy, our tools, firstly developed for the exploration of gene expression data, naturally grew and encompassed other areas of genomic research. For instance, in **Montaner** et al. 2006, GEPAS included methods for the analysis and exploration of Array-CGH data (Mantripragada et al., 2004). Utilities for estimating genomic copy number alterations, and tools for the visualization of such regions were easily inserted into our refurbished tool[1]. The new modularity of GEPAS allowed us not only to incorporate new methods into the suite, but also, to interact with other tools developed in our lab. An example of such interconnection was PupaSuite (Conde et al., 2004, 2006), which, combined with GEPAS, offered the possibility of analyzing polymorphisms within genomic regions selected according to gene expression criteria. So, for the first time, gene expression, copy number variation, and polymorphisms methodologies where integrated together, allowing for the setup of experiments based upon several sources of genomic data at once. But probably the most crucial advantage provided by GEPAS in **Montaner** et al. 2006, was the broad access to functional profiling methods.

In any genomic experiment, researchers state a hypothesis and collect

[1] A general overview of such methods and its interconnection with other GEPAS modules can be found in Conde et al. 2007.

data which information, properly interpreted, should lead us to support or reject such hypothesis. The collected data are new, in the sense that no one else has seen its information before; otherwise the experiment would had never been raised. But the analysis and interpretation of such original information, should be done in the light of what is already known and established as scientific knowledge. In the case of genomic studies, what is "known" is not what researchers remember by heart, but what is stored in digital databases and, evidently, the "interpretation" of the data cannot be done without bioinformatic routines and tools.

In the past years many computer programs and algorithms have been created in order to combine experimental data from genomic studies with *annotation* information form databases. FatiGO (Al-Shahrour et al., 2004, 2007b), makes a functional interpretation of experimental genomic results in terms of the Gene Ontology database. SNOW (Minguez et al., 2009), queries information about protein-protein interaction from several databases and performs statistical tests to explore the *interactome* of the samples under observation. GESBAP (Medina et al., 2009), extracts gene functional information from databases and extends it to single nucleotide polymorphisms via genomic position, allowing for the interpretation of variation data. Paintomics (Garcia-Alcalde et al., 2011), recovers information form the KEGG Pathway database and combines it with experimental data from transcriptomics and metabolomics experiments.

Most methods to jointly analyze experimental and database information use algorithms such that, the steps handling the experimental data are inseparable from those of the functional interpretation. For instance, the well known GSEA (Mootha et al., 2003; Subramanian et al., 2005), performs an enrichment test of Gene Ontology terms for two class differential expression data. Nevertheless, the assessment of significance in the "interpretative" part of the method relies up on permutations of the expression data, that is, the experimental data. Hence, the tool cannot be applied in experimental designs with, for instance, more than two classes or with continuous covariates, unless the algorithm itself is modified. This rigidity is a major drawback for the generalization and further use of such methodologies.

In our case, our modular way to tackle genomic studies was also reflected in our functional profiling approach. It is remarkable that, despite being GEPAS a tool for the explicit analysis of experimental data, a full section of **Montaner** et al. 2006 is devoted to genomic functional annotation. This is so because GEPAS has always been linked to Babelomics, an independent resource for functional annotation and analysis of groups of genes in high-throughput experiments.

At that time[1], Babelomics was able to query annotation from the Gene Ontology Data base (Ashburner et al., 2000), the Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000) and many other annotation databases. It also included functional information generated from text mining of biomedical literature and gene signatures compiled in paradigmatic experimental studies.

But Babelomics was not just a major database of genomic information, it was also outfitted with powerful tools that let the researcher take advantage of such information. In Al-Shahrour et al. 2006, Babelomics implemented different procedures for the functional interpretation of sets of pre-selected of genes. The rationale behind the tool derived also from the modular data analysis orientation: first, experimental information from high throughput devices would be appropriately deal with using any software tool such as GEPAS. This first step would yield, perhaps, a group of “selected genes” that would “mean something” according to the experiment; or it may, in other occasions, yield an “arrangement” of the genes in the study defined by a p-value, evaluating the departure form a biological null hypothesis. Then, a second layer of methods, those properly belonging to Babelomics, would interpret those “selected” markers or make sense of such “arrangement” of genes.

Hence, in order to achieve an optimal communication between the two suites, GEPAS in **Montaner** et al. 2006 had to format and reshape

[1] See Al-Shahrour et al. 2006 for more details.

the output from every single statistical or analytical method available. The objective was not only to provide the results of GEPAS modules in files “readable” by Babelomics. Moreover, we had to organize and classify the analytical methodologies in GEPAS according to the kind of functional profiling which could be meaningfully carried out using Babelomics. This took a great effort of organization of the tool and made us spend considerable time studying the characteristics of each statistical module in GEPAS; its input and output, its ultimate biological meaning, the way in which users will employ the module and how they will conceptualize the results form GEPAS and the redirection to Babelomics. GEPAS outputs had to be neatly organized to be interpreted on their own, but also to provide the intuition and clues about how to keep going with the next natural step in Babelomics.

Thus the work done for the implementation of GEPAS in **Montaner** et al. 2006 and Babelomics in Al-Shahrour et al. 2006, steeled the basis for many of our subsequent developments. There was still one more release of the tools as independent resources, (Al-Shahrour et al., 2008; Tarraga et al., 2008). Finally, as it was natural, both of them converged into a unique tool under the Babelomics “trademark”, (Medina et al., 2010).

In what concerns this thesis, the job in GEPAS from **Montaner** et al. 2006 and Babelomics from Al-Shahrour et al. 2006, highlighted the utility of analyzing genomic data from different sources at a time, and pointed out the lack of methodologies for the functional interpretation of such combined studies. This motivated my later research published in **Montaner** et al. 2009 and **Montaner** and Dopazo 2010, presented also in this thesis.

1.2 Gene set modules

There are thousands of databases storing genomic information. The most popular of them are open access and freely accessible to everyone via Internet. The Gene Ontology repository (Ashburner et al., 2000) is probably the mos widely used. It maintains and develops a controlled vocabulary of gene and gene product attributes. The KEGG Pathway Database (Kanehisa et al., 2004) describes networks of molecular interactions in the cell. Reactome (Joshi-Tope et al., 2005) has a navigable map of chemical reactions, pathways and biological processes. OMIM (McKusick-Nathans) keeps a catalog of human genes related to traits, phenotypes and diseases. But there are also many proprietary databases like, for instance, those provided by BIOBASE®, or even the same KEGG in its latest releases. And of course, every research group may have their unpublished genomic information related to their particular needs.

These databases represent the most up to date biological knowledge we have nowadays. It is anything but complete, as many pieces of the puzzle are still missing, but is the best asset upon which research can be developed. The quality of the stored information is generally quite variable. Part of the information kept has been curated by experts but, most of it, has been generated *in silico* using automatic routines which may perform with different levels of accuracy. The precision and the amount of information provided by databases is also very different from some sources respect to the others. It can range from the very detailed description of the *oxidative phosphorylation* pathway in KEGG (see figure 1.5 on page 14), which provides the structure and relationship between the gene products involved in the process, to the almost meaningless and "flat" annotation of the genes under the *female pregnancy* GO biological process (see figure 1.6 on page 14).

Fortunately, a great effort is made to improve and complete these sources of genomic information as in most areas of the molecular biology they are an invaluable tool for investigation. In any case, this database knowledge is the starting point to conform many research hypotheses and, more importantly, they provide the only means to draw a biological interpretation of most experimental results. In this work we are not so much interested in how this databases are generated or curated but in how to take practical advantage of them. Particularly, in how to use them to perform *gene set analyses*.

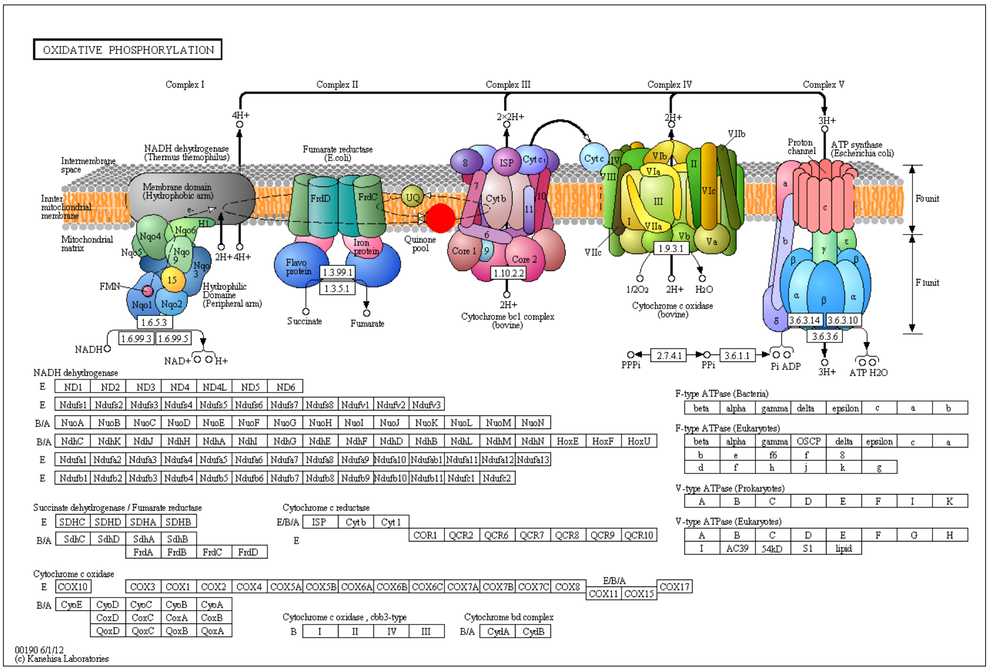

OXIDATIVE PHOSPHORYLATION
Complex I
Complex II
Complex III
Complex IV
Complex V
4H+
2×2H+
2H+
3H+
Intermembrane space
Inner mitochondrial membrane
Mitochondrial matrix
NADH dehydrogenase (Thermus themophilus)
Membrane domain (Hydrophobic arm)
Nqo4
Nqo6
Nqo5
Nqo9
Nqo3
15
FMN
Nqo1
Nqo2
Hydrophilic Domaine (Peripheral arm)
2H+4H+
1.6.5.3
1.6.99.3
1.6.99.5
NADH
NAD+
H+
Fumarate reductase (E.coli)
FrdD
FrdC
UQ
Iron protein
Flavo protein
1.3.99.1
1.3.5.1
Succinate
Fumarate
Quinone pool
ISP
Cyt c1
Cyt b
Core 1
Core 2
1.10.2.2
Cytochrome bc1 complex (bovine)
Cyt c
VIIb
VIb
VIa
VIII
III
II
VIc
VIIa
Vb
Va
VIIc
1.9.3.1
1/2O2
2H+
H2O
Cytochrome c oxidase (bovine)
Proton channel
ATP synthase (Escherichia coli)
F0unit
F1unit
α
β
δ
3.6.3.14
3.6.3.10
3.6.3.6
2.7.4.1
3.6.1.1
PPPi
PPi
Pi ADP
3H+
ATP H2O
NADH dehydrogenase
E ND1 ND2 ND3 ND4 ND4L ND5 ND6
E Ndufs1 Ndufs2 Ndufs3 Ndufs4 Ndufs5 Ndufs6 Ndufs7 Ndufs8 Ndufv1 Ndufv2 Ndufv3
B/A NuoA NuoB NuoC NuoD NuoE NuoF NuoG NuoH NuoI NuoJ NuoK NuoL NuoM NuoN
B/A NdhC NdhK NdhJ NdhH NdhA NdhI NdhG NdhE NdhF NdhD NdhB NdhL NdhM NdhN HoxE HoxF HoxU
E Ndufa1 Ndufa2 Ndufa3 Ndufa4 Ndufa5 Ndufa6 Ndufa7 Ndufa8 Ndufa9 Ndufa10 Ndufab1 Ndufa11 Ndufa12 Ndufa13
E Ndufb1 Ndufb2 Ndufb3 Ndufb4 Ndufb5 Ndufb6 Ndufb7 Ndufb8 Ndufb9 Ndufb10 Ndufb11 Ndufc1 Ndufc2
Succinate dehydrogenase / Fumarate reductase
E SDHC SDHD SDHA SDHB
B/A SdhC SdhD SdhA SdhB
FrdA FrdB FrdC FrdD
Cytochrome c reductase
E/B/A ISP Cyt b Cyt 1
E COR1 QCR2 QCR6 QCR7 QCR8 QCR9 QCR10
Cytochrome c oxidase
E COX10 COX3 COX1 COX2 COX4 COX5A COX5B COX6A COX6B COX6C COX7A COX7B COX7C COX8
E/B/A COX11 COX15 COX17
B/A CyoE CyoD CyoC CyoB CyoA
CoxD CoxC CoxA CoxB
QoxD QoxC QoxB QoxA
Cytochrome c oxidase , cbb3-type
B I II IV III
Cytochrome bd complex
B/A CydA CydB
F-type ATPase (Bacteria)
beta alpha gamma delta epsilon c a b
F-type ATPase (Eukaryotes)
beta alpha gamma OSCP delta epsilon c a
b e f6 f 8
d f h j k g
V-type ATPase (Prokaryotes)
A B C D E F I K
V-type ATPase (Eukaryotes)
A B C D E F G H
I AC39 54kD S1 lipid
00190 6/1/12
(c) Kanehisa Laboratories

Figure 1.5: KEGG pathway describing the *oxidative phosphorylation* process.

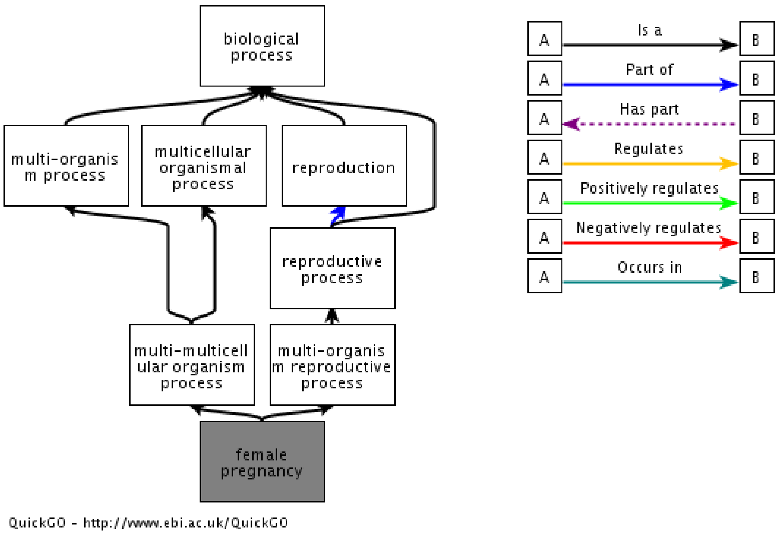


Figure 1.6: Graph structure showing the GO term *female pregnancy* and its *ancestors*. All terms in the plot are represented using square boxes; there is not any further internal organization among the genes annotated under each term.

Each gene or biological entity is ultimately unique and has qualities which differentiate it from the others. Nevertheless, in order to understand nature, generate theories and communicate knowledge, scholars and researchers need to distinguish and isolate general characteristics, properties and qualities that are shared among the biological pieces under study. Genomic databases capture such information in contingency tables where genes are linked to the labels identifying each property or attribute. In the functional genomics jargon those tables are referred to as *functional annotations*, and the groups of genes tagged under the different labels are called *gene sets*.

As discussed in the previous section, current genetic experiments have the capacity to collect information at a gene level. Hence, the straightforward approach to data analysis is to consider the gene as the intrinsic unit of interest in the study. In this first instance, hypotheses are stated over the genes and statistical tests are carried out at such *observational level*. But the complexity of interpreting statistical results on thousands of genes besides the scarce factual information provided by each of them, immediately pushes us to call the *annotation* databases. Generally the inclusion of the *annotation* in the study is done with two main purposes: to ease the interpretation of the results and to increase the amount of information included in the experiment. The advantages for the interpretation are clear; it is easier to discuss and draw conclusions about terms which are already defined to have a meaning, such as those included in the annotation, than to think of the abstract tokens which are the genes. Furthermore the shift in the observational unit, from the gene to the attributes described in the database, usually reduces the number of units to be considered in the study.

It is also apparent that adding the knowledge about the genes increases the amount of information in the study. Despite of that, such inclusion of the information cannot be done arbitrarily. Care should be taken not to induce any bias which will produce misleading results. Well designed *functional profiling* algorithms and *gene set* methodologies, guarantee that no distortion is introduced in the procedure and assure its statistical validity. We will discuss such issues in following sections but, for the purpose of these thesis, it is first important to point out two key aspects on how *gene sets* are modeled when included in *gene set analysis*:

Most databases internally keep extensive information on how and why genes are linked to an attributes. The Gene Ontology data base, for instance, includes an *evidence code* to indicate how the annotation of a gene to a particular term is supported. OMIM records are justified by different number of publications, each of which relates a gene to a disease with different nuances. Ideally, such extra information could be used to quantify how much we should trust each annotation and then, that reliability could be included into the *gene set analysis* model. But despite the quality of the information may be available, it is difficult to translate it in terms of biological relevance, and much more in a general manner which could be applied to most diverse experimental settings. In general, the best that can be achieved when the database information is injected into the data analysis, is to establish some filters for discarding low quality annotations. This returns a *true or false* discrete inclusion of the information into the analysis, and gives the same importance to all of the annotations that passed the filtering cutoff. Thus, once the annotation is filtered and included into the study, all of the genes tagged under the same attribute or label, that is all the genes in a *gene set*, are given the same reliability and importance. However, from the database quality flags themselves, we can realize that such assumption of homogeneity of the annotation is not the most accurate one.

On the other hand, some databases like for instance KEGG or Reactome include, not only the association of genes to certain attributes, but the description on how those genes interact among them. Thus, when using those databases or annotations, we soon realize that not all genes are equally important within a *gene set* or functional block. Some genes may behave as a hub[1], interacting with many others, some may be peripheral and perhaps not so biologically relevant in all experimental

[1] See figure 1.15 on page 33 for an example.

settings. That said, such asymmetry of the genes within the *gene set* is not easily modeled. Thus, for practical reasons, most *gene set* methodologies do no make use of it, and handle all the genes under an annotation tag similarly. Systems biology has always been dedicated to take advantage of such *topological* information but, due to the incompleteness of the databases and the lack of computational models describing the interactions, its practical use has been quite limited in the experimental world, being the *gene set analysis* the most popular way of merging database information and experimental data.

Hence, up to the time we wrote **Montaner** et al. 2009, *gene sets* where used as "flat" gene annotations, independently of the database used for the functional profiling or the experimental context of the data collection. In the first part of the article, we discussed this issues for the particular context of gene expression analyses. We empirically demonstrated that not all the Gene Ontology terms or KEGG pathways have such "flat" behavior, and that the non homogeneity of the genes within a functional block can be demonstrated form the different levels of coexpression among them.

In the same paper, we also collected a massive dataset form gene expression public repositories, and used it to create a general purpose co-regulation index for all the genes in the human genome. We used the gene to gene co-regulation score to provide a *continuous* estimate of the biological relevance of genes within each *gene set* module. Hence, for the first time, we included the concept *non-discrete membership* to a functional module. A gene was not just *in* or *out* of the *gene set* but *far from* or *close to* its core functionality.

But we did not just limit ourselves to spot the drawbacks of the, at that time, current methodologies of *gene set analysis*. In the second part of **Montaner** et al. 2009, we developed functional profiling methodologies which would be able to handle the novel concept of *non-discrete gene set*. We introduced the *weighted logistic regression* models as the natural extension of standard *gene set* methodologies to the context of continuous membership to a functional class.

1.3 Functional profiling

Functional profiling is a quite loose term coined to encompass all statistical and computational methods aiming to provide a “functional” interpretation of genome scale experiments. Here, the adjective *functional* should be understood as referring to the role that genes, or better, the gene products, play within the cellular machinery and along the biological processes.

As discussed in previous sections, once an experiment is carried out and genomic measurements collected, the first step in the data analysis procedure is, usually, to perform an individual gene level analysis. The next natural step is to try to extract some general sense out of those results, and see whether it fits with what is already known about the biological process under study. Such interpretation of results, in the light of what is already known, passes by the superimposition of database stored knowledge over our experimentally collected data. Functional profiling methodologies try to ensure that, the superimposition or combination of already established knowledge (the one extracted form the databases) and the new information (that of the experimental data) is properly done, without introducing any biases.

As an example, a classic bias introduced when naively incorporating Gene Ontology information into gene expression analyses is as follows: the researcher gets the differentially expressed genes, collects the GO terms annotated to them, finds out that 8% of the genes are annotated under the term *cell death*, and concludes that *cell death* is relevant to the process under study because such percentage is “high”. The bias is clear when we realize that 8% of the genes of the human genome is also annotated under the *cell death* flag. Evidently, we need to correct for the underlying distribution of the *annotations* over the reference genome.

This kind of correction is exactly what *functional enrichment analysis* does (Dopazo, 2010). *Functional enrichment analysis* is perhaps the simplest *functional profiling* approach. In this methodology, the interpretation is carried out in two steps: in the first one, some genes of interest are selected. This selection can be done according to very different criteria. It will depend, for sure, on the experimental context being analyzed; differentially expressed genes, genes with copy number alterations, mutated genes associated to disease status . . . But it will also rely on the statistical test or algorithm used in the gene level analysis, and, specially, on the cutoff chosen to finally select the genes relevant to the ongoing study.

In the second step, the *functions* annotated to the selected genes are contrasted against those annotated to a group of background or reference genes, usually the remaining genes in the genome of the species under survey. For each functional block or *gene set* annotated to the species, its proportion of occurrences in the selected genes is compared to the proportion of appearances in the reference list of genes, searching for an *enrichment* in one of the groups compared to the other. Such comparison is performed via statistical testing, usually a χ^2 test or a Fisher exact test, but many others, as the use of *logistic regression models* here introduced, are possible (see section 1.4 on page 22). *"This comparison with the background is essential because an apparently high proportion of a given functional module could easily be nothing but a reflection of a high proportion of this particular module in the whole genome but not a proper enrichment"* (Dopazo, 2010).

Apart from the need to control for the "standard" annotation background, there are few important remarks regarding *gene set profiling* approaches which can be already noticed in the simple *functional enrichment analysis*. First there is a need to ensure that the detected enrichment is not just a result of randomness. This is generally achieved by using statistical hypothesis testing. Thus, one p-value is provided for each *gene set* and then, all of them are corrected in order to control the amount of false positives arising in the multiple testing context. There are proposed many multiple testing methodologies, but may be the ones published in Benjamini and Hochberg 1995 and Benjamini and Yekutieli 2001 are the most popular ones in the genomic context. An implicit consequence is the *shift in the observational unit*, from the gene, to the *gene set*. At the gene level analysis, the unit of observation, the statistical variable, is the gene; the hypothesis tests are set for each gene, and for each gene are also derived the p-values. At the *gene set* level, one contrast is specified for each *gene set*, and a corresponding p-value is then derived. The unit of interest, the entity we think about now, is not any more the gene, but a block of genes acting together as a "biological machine".

Another relevant issue to be notice is that, despite the method is known as *enrichment* analysis, it may happen that the amount of genes annotated under certain *term* is lower in our group of genes than in the reference one. Such *term* is hence reduced in our genes. Put in a different way, the enrichment of a particular functional term or *gene set*, can occur in the group of selected genes or in the group of background genes. In the case where the background genes can be assumed to be the reference genome, the biological interpretation for this situation would be that the function is lost or deactivated in our selected genes.

The *functional enrichment analysis* is widely used in almost any genomic experiment as an straightforward functional profiling methodology. Nevertheless, the two-steps paradigm which forces the user to set a cutoff in the first stage, is far from being optimal (Dopazo, 2009). It may happen, for instance, that no gene surpasses the threshold. Hence, the second step of the functional interpretation cannot be done and the method itself becomes meaningless. Besides that, even when there are genes exceeding the threshold, generally those are a minimum part of all genes available in the study. As the functional profiling focuses just on those few genes, most of the functional information or *annotation* is discarded, making the method highly inefficient; see figure 3.1 on page 53 and section 3.1 for more details.

Despite of that, the major drawback of the two-steps functional profiling methodologies does not derive from those "statistical" limitations but from the biological conceptualization implied by the method itself. By focusing on just those genes surpassing a threshold, the *functional* *enrichment analysis* implicitly primes those genes showing the highest biological changes. In some research contexts, this may be the accurate model but, in general the highest changes may not be the mos relevant ones. In the case of gene expression studies, for instance, the selected genes would be those showing the highest statistical significance and the biggest changes in its expression across biological conditions. Sometimes that is what researchers are looking for and the two steps methodology may be the suited one. In a synaptic transmission, for instance, we expect all the genes involved to greatly increase their expression levels. But in many other biological contexts we would not expect a great change of the genes involved under a functional term. In a metabolic process, for example we would expect that the "biological machine" would be activated by an mall, but coordinated increment of all the genes in the corresponding metabolic pathway.

As a response to such considerations about *functional enrichment analysis*, Mootha et al. 2003 devised an analytical procedure which will lately concur in the first *gene set analysis* methodology. The ideas introduced in that first paper were fully developed in Subramanian et al. 2005 under the name of *Gene Set Enrichment Analysis* (GSEA).

The basic idea behind GSEA is to use the gene level statistic, not for selecting genes having high (or low) values in it according tho certain threshold, but to *rank* the whole list of genes in the study according to the biological condition accounted for in such statistic. *"The goal of GSEA is to determine whether members of a gene set S tend to occur toward the top (or bottom) of the list, in which case the gene set is correlated with the phenotypic class distinction"* (Subramanian et al., 2005).

The rationale behind GSEA, which also characterized later *gene set analysis* methods, is that, in any experiment, the statistic derived at a gene level bears always some amount of information. That information may not be relevant or significant at a gene level but, when considered at *gene set* level, it may reflect consistent patterns providing valuable insights of the underlying biology.

In this section we have just presented the concept of *functional profiling* of genomic experiments in an unpretentious manner. Chapter 3 further develops the concept of *gene set analysis* as a subclass of *functional profiling* methodologies. Then chapters 4 and 5 will present the methodological improvements proposed in this thesis for *gene set analysis* algorithms. In the remaining sections of this chapter we will introduce some statistical concepts necessary to fully understand those algorithms.

1.4 Logistic regression models

In statistics, a *binary variable* is that one which may take only two possible values. Tossing a coin, observing if there has been a failure in a machine or answering to a yes-no questionnaire are "experiments" which yield a binary variable as an outcome. Generally, the two values in a binary variable are coded numerically as 0 and 1. Despite of that, they are considered to lay on a nominal scale, representing qualitative differences without a prior order.

In genomic annotation studies for instance, the fact that a gene is annotated or not under certain biological function, may be represented using a binary variable. We can code the response for each gene as 1, if the gene is annotated under the biological term, or 0 if the gene does not have the biological function under consideration[1].

Logistic regression models, also called *logit* models, are a set of *generalized linear models* in which a binary variable is predicted as a response to one or several other independent variables. Those independent variables may themselves be continuous, ordered, categorical with two values, that is binary, or even categorical with more than two categories.

The *logistic model* relates the probability of the response variable being 1 (as opposite to 0) to the explanatory variable using a linear

[1] Alternatively we could use the 0 to represent the annotated genes and 1 the unannotated ones. This will only be taken into account when interpreting the binary variable itself.

equation. If we call π to the probability that the response variable takes value 1, and X to the independent variable, then, in the regression context the probability π is considered to be a function of X. Thus we will use the notation $\pi(x)$ to represent the probability of response 1 when the independent variable X takes the value x.

In its simplest representation, a *logistic model* states a relationship between π and x that can be formulated as:

$$\log \frac{\pi(x)}{1-\pi(x)} = \kappa + \alpha x \tag{1.1}$$

Where κ is called *intercept*, α is known as *slope* and the $\log \frac{\pi}{1-\pi}$ is called the *logit* transformation of the probability π.

A somehow more Bayesian formulation will state:

$$\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \kappa + \alpha x \tag{1.2}$$

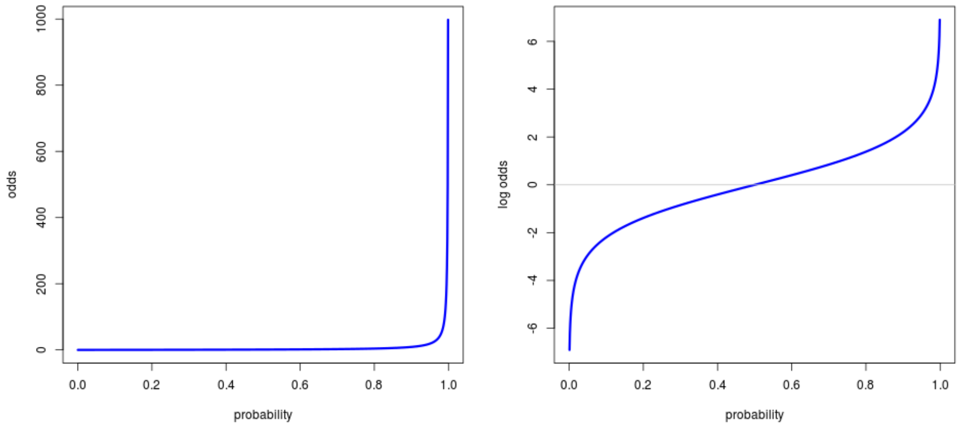
Where Y is the *binary* response variable taking values 0 or 1.

$\frac{\pi(x)}{1-\pi(x)}$ is called the *odds* of Y being 1 and the $\log(\frac{\pi(x)}{1-\pi(x)})$ is called the *log odds* of Y being 1. These two quantities, the *odds*, and the *log odds*, are alternatives to the *probability* that also "measure" how likely it is that the *binary* response variable takes value 1. Indeed there is a one to one increasing relationship between the probability and both, the *odds*, and the *log odds*. See figure 1.7 on page 24.

Thus, when the *logistic* model is fitted to our data, the slope parameter α, represents the increment in the *log odds* of Y being 1, when the independent variable X increases a unit.

If α is estimated to be positive, then the *log odds* of Y being 1 increases as X increases. Equivalently[1], the probability of Y being 1 also increases with X. Thus, the higher the value X takes, the more likely is the *binary* variable Y to take value 1. See figure 1.8a on page 25.

[1] Because of the one to one *increasing* relationship between the *log odds* and the probability indicated above.



(a) odds vs. prob

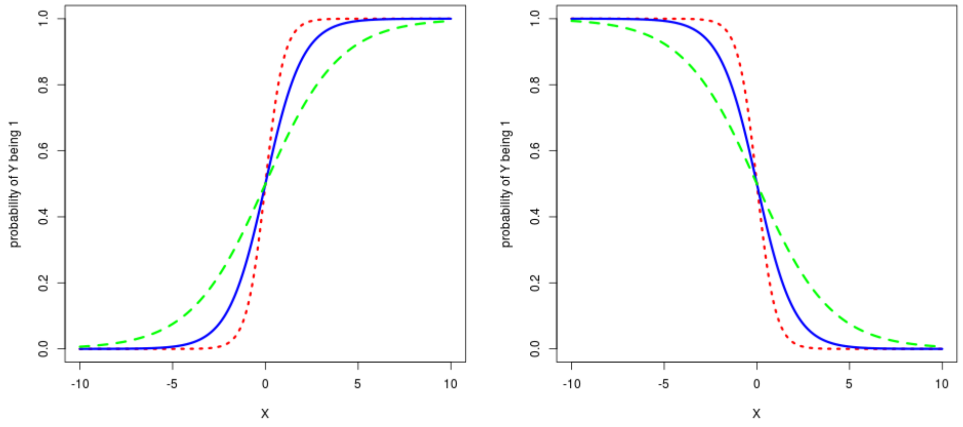
(b) log odds vs. prob

Figure 1.7: Representation of the relationship between the odds and the log odds and the probability.

On the other hand, If α is estimated to be negative, then the *log odds* of Y being 1 decreases when X increases. Hence, the higher it is X the more likely is Y to be 0 or, symmetrically, the lower it is X the more likely is Y to be 1. See figure 1.8b on page 25.

As in any other standard regression context, the interest when fitting the *logistic regression* model to our data, is to get an accurate estimate for the linear coefficients κ and α, being the last one (α) generally of more interest than the former one (κ). The standard approach is to fit parameters is via *maximum-likelihood* estimation but other approaches have been used: Bayesian, empirical least squares . . . Besides the parameters themselves, the different adjustment methods also yield some estimates for their variabilities. Using those variabilities it can be statistically tested whether the estimated coefficients are significantly different form zero or not, for instance using the so called *Wald statistic*, Agresti 2002. But many other approaches for testing hypothesis over the estimated coefficients are described: likelihood ratio test, Bayesian inference, permutation or bootstrapping . . .

Hence, when fitting a *logistic regression* model, we will get estimates for the *intercept* and the *slope*, and some index, such as p-values or



(a) α is positive

(b) α is positive

Figure 1.8: Representation of the relationship between probability of Y being 1 and the values taken by the independent variable X. Blue continuous lines represent a relationship with an α coefficient being 1 or -1. Red dotted lines represent a relationship with an α being 2 or -2. Green dashed lines represent a relationship with an α being 1/2 or -1/2.

posterior probabilities, indicating how likely are those estimates to be different form zero. If the estimates are *significantly* different form zero, then we report a *non random* relationship between the *binary* variable Y and the independent variable X. The sign and value of the coefficients are then suitable to describe the dependence of Y on X.

Logistic regression models are widely applied in numerous statistical contexts and used in most diverse disciplines. In what concerns this thesis, they are the workhorse for the functional profiling algorithms presented in **Montaner** et al. 2009 and **Montaner** and Dopazo 2010.

As argued in previous sections, the analysis of the data of most genomic experiments returns, ultimately, a numerical “index” value associated to each gene under study (see figure 1.9a on page 26). Afterwords, researchers may decide to *discretize* such index, for instance, setting a cutoff threshold and binarizing the information attached to each gene in a "pass" or "do not pass" the cutoff status[1] (see figure 1.9b on page 26). Such binarization is sometimes unavoidable because of the study requirements, but in many other cases it is unnecessary, or even inadvisable because of the lost of information which implies. Whatever the case, *continuous* or *discrete*, the gene level analysis ends up with a numerical variable X associated to each gene.

ID	X
gene1	1.23
gene2	2.74
gene3	-0.34
gene4	1.32
gene5	-2.02
gene6	0.45
gene7	0.93
...	...

(a) Continuous index attached to each gene.

ID	X
gene1	1
gene2	1
gene3	0
gene4	1
gene5	0
gene6	1
gene7	1
...	...

(b) Discretized index attached to each gene.

Figure 1.9: Numerical indexes attached to genes as a result of some statistical analysis.

After this analysis step, which involves just the "experimental" data, it comes the time to interpret the results in terms of what is already known about the genes under survey. For this purpose databases are queried and gene *annotations* extracted. Independently of the format in which we can have access to the *annotation*, the underlying structure of the annotated data is that of a two column file. In the first column we will get the gene identifiers, and in the second one, we will get identifiers of the different blocks of information known about the genes, that is, the *gene sets*. See figure 1.10 on page 27.

If we want for instance to interpret our experiment in terms of the Gene Ontology database, our second column may contain the identifiers for the different biological processes associated to our genes, that is,

[1] As mentioned above, this status is usually coded using a 0, 1 numerical variable.

gene1	label 1
gene2	label 1
gene4	label 1
gene1	label 2
gene2	label 2
gene3	label 3
gene4	label 3
...	...

Figure 1.10: A representation of a two columns *annotation* file.

the different biological processes within which our gene is known to be somehow involved in. See figure 1.11a on page 27. If we query the KEGG Pathway database, the second column of the annotation will bear the identifiers for each of the registered pathways, metabolic, signaling and so on. See figure 1.11b on page 27.

gene1	GO:0055114
gene2	GO:0055114
gene3	GO:0055114
gene1	GO:0044281
gene2	GO:0044281
gene4	GO:0006120
...	...

(a) A representation of a two columns *annotation* file for the Gene Ontology

gene1	01100
gene2	01100
gene1	01100
gene2	00710
gene4	00710
gene4	00540
...	...

(b) A representation of a two columns *annotation* file for the KEGG pathway database

Figure 1.11: Annotation matrices describing GO and KEGG terms.

As one gene may be involved in several of the biological processes registered in the data base, in the most general case, gene identifiers will appear duplicated in the firs column of such annotation structure. Also the *annotation* identifiers, those of the second column, should appear repeatedly, as there will be several genes involved in any of the registered processes[1].

All the genes *tagged* under one same annotation identifier correspond to what we have been calling *gene set*, a group of genes that are known to conform a higher level biological entity, interesting to be studied as a whole. If we consider the "universe" of genes measured in the study[2], and we focus on a particular *gene set* or annotation tag, then, each gene of the research universe can just be "inside" or "outside" the predefined *gene set*. That is, once a *gene set* is fixed, belonging to it is a *boolean* condition for the genes in the study. Such condition can be recorded into a *binary* variable, Y, using a 0, 1 coding as explained at the beginning of this section.

When considering the complete *annotation* or database, as the "universe" of genes is fixed, all 0, 1 variables for the different *gene sets* can be arranged into a matrix structure, being the genes in the rows and the annotation tags in the columns[3]. See figure 1.12 on page 29 as an example of such data structure representation.

Hence, in a usual genomic experiment we end up having, for each gene

[1] Annotation databases aim to abstract biological characteristics which are shared by several genes. Hence, ideally, there should not be described in the database any *block* or *gene set* involving a unique gene. Eventually there may be *annotations* tagging just a single gene, but this generally occurs because it is expected that, at some point, some other genes will fall under such *biological classes*. Generally, biological classes having "few" genes are excluded in the *functional profiling* steps of the analysis.

[2] Usually the theoretical set of genes measured in an experiment is the complete genome of the species under study. Nevertheless, many times in practice, the collection of genes effectively measured is constrained by the *high throughput* technology employed. In the case of microarrays, for instance, we are restricted to the genes spotted in the glass. When using Next Generation Sequencing methods, just genes *known* for the species under observation are generally available in the preprocessed dataset.

[3] The transposed matrix, genes by columns and annotation tags in the rows, is obviously possible too.

	label 1	label 2	label 3	...
gene1	1	1	0	...
gene2	1	1	0	...
gene3	0	0	1	...
gene4	1	0	1	...
...	...	...	...	...

Figure 1.12: A *boolean* matrix representing the same annotation information than the figure 1.10 on page 27. We can see, for instance, how "gene1" is annotated under "label 1" and "label 2" but not under "label 3". At the same time, "gene3" is the only one belonging to the biological class named "label 3".

under survey, a *ranking*[1] index X accounting for some biological property evaluated in the experiment, and a binary variable Y, representing the prior knowledge of whether the gene is annotated or not to certain *gene set*[2]. See figure 1.13 on page 29.

	label 1 (Y)	index (X)
gene1	1	1.23
gene2	1	2.74
gene3	0	-0.34
gene4	1	1.32
...	...	...

Figure 1.13: Combined information from experimental results in figure 1.9a (page 26) and annotation information in figure 1.12 (page 29)

The *functional profiling* step of the interpretation of the experimental results in terms of the annotation database is then formalized in the statistical question of whether there is a relationship between the binary variables Y and the ranking index X. As already mentioned, there are many statistical frameworks proposed by the different authors in order to

[1] If the data analysis step is properly done and X meaningfully derived, X should effectively be a variable which orders all the genes in the experiment according to the researchers interests.

[2] More generally, we can say that, after the data analysis and the query of the information in the annotation database, we will have the numerical index X, and several binary variables Y_i, one for each of the labels $1, 2, \ldots N$ representing the *gene sets* in the database, as in figure 1.12.

estimate the strength of the dependence between Y and X. The methodologies proposed this thesis, particularly in **Montaner** et al. 2009 and **Montaner** and Dopazo 2010 rely up on the usage of *logistic regression* models.

Following the notation introduced in equation (1.2) on page 23, we can state the *logistic* model for the dependence between Y and X as:

$$\log \frac{P(Y=1)}{P(Y=0)} = \kappa + \alpha X \tag{1.3}$$

Or, if we wanted to consider the more general context including the multiple *gene sets* of the annotation, we will write:

$$\log \frac{P(Y_i=1)}{P(Y_i=0)} = \kappa_i + \alpha_i \; X \tag{1.4}$$

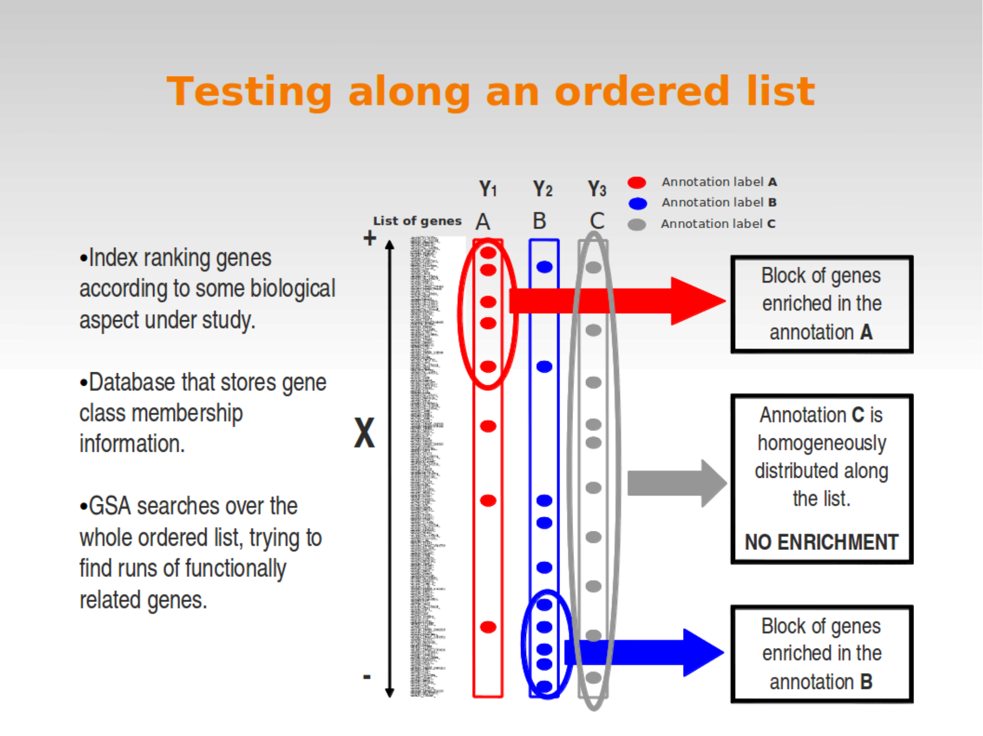
Where $i \in \{1, 2, \ldots N\}$ indicates that there is a *logistic* model *fit* for each of the N *gene sets* or *labels* in the annotation[1].

From the formulation in equation (1.4) is easy to interpret the α_i coefficients in the model; the interesting ones for our purpose of using the *logistic* model as a *gene set analysis* tool. When α_i is *significantly* positive, equation (1.4) implies that, if X increases, so it does the probability of Y taking value of 1 (see figure 1.8a on page 25). Applied to our *gene set analysis* context, it means that the greater it is the "experimental" index X for a gene, the greater it is the probability of the gene being annotated into the *gene set* i. If we where for instance in a differential expression analysis context, X could be a *fold change* accounting for gene over-expression of cases compared to controls. The interpretation of a *significantly* positive α_i then would be that, the genes more "up-regulated" (those showing higher X values) would be more likely to belong to *gene set* i (those having $Y_i = 1$). Thus, the *gene set*, as a whole, shows increased gene expression levels in cases compared to controls.

[1] Note that the index i iterates over the *gene set* indicators but not over the X variable. The continuous index is the same for all models.

Conversely when α_i is estimated to be *significantly* negative, the probability of Y taking value of 1 increases as X decreases (see figure 1.8b on page 25). In our example of the differential gene expression analysis, a negative α_i value would then indicate that the *gene set* is consistently more expressed in controls than in cases. That is, the *gene set* is "down-regulated" in cases compared to controls.

Figure 1.14 on page 31 represents the three different possible trends in the "distribution" of the genes of a *gene set*, over a *ranking* of genes according to a continuous variable X measuring differential expression.



Testing along an ordered list

- Index ranking genes according to some biological aspect under study.
- Database that stores gene class membership information.
- GSA searches over the whole ordered list, trying to find runs of functionally related genes.

Y1
Y2
Y3
List of genes
A
B
C
Annotation label A
Annotation label B
Annotation label C
+
X
-
Block of genes enriched in the annotation A
Annotation C is homogeneously distributed along the list.
NO ENRICHMENT
Block of genes enriched in the annotation B

Figure 1.14: Possible *gene set* trends over an ordered list of genes.

In this section we have introduced the straight forward usage of *logistic regression* models for "standard" *gene set analysis*. As mentioned above there are many alternative statistical approaches which can be used for conventional *functional profiling*. The advantages of using *regression* models relies on their flexibility and easiness of extension to more complex settings or experimental designs. In the following sections we introduce some of those extensions up on which part of the work developed in this thesis relies on.

Section 1.5 explains how *weights* can be introduced into the *logistic model*. This is the basis for the developments we introduced in **Montaner** et al. 2009, presented in chapter 4. Section 1.6 develops the *multidimensional regression* framework. This extension of the model is used in **Montaner** and Dopazo 2010 to develop *gene set analysis* methods which combine several sources of experimental information. This part of the work is presented in in chapter 5.

1.5 The weighting schema

As pointed out in the previous section, in principle, *functional profiling* algorithms model gene membership to a *gene set* as a binary variable. In such conceptualization, for any gene, belonging to a functional block is a discrete condition: the gene is within the biological machinery described by the *gene set*, or is out of it. Nevertheless, such understanding of the *gene sets* as homogeneous entities, supposes a clear simplification of the underlying biology described in the databases. Each of the pathways described in the KEGG database, for instance, has its own *topological* structure. The genes integrating the pathway relate each other conforming a network and, depending on the biological process under study, some *nodes* or genes of the net are more relevant than others as can be seen in figure 1.15 on page 33.

The internal structure of the *gene sets* is not so explicitly defined in the GO database, but the underlying structure of the *ontology* of terms implies substructures within each of the gene ontology *gene sets*. The genes in the *regulation of apoptosis* term, for instance, are further separated into the exclusive terms of *positive regulation of apoptosis* and *negative regulation of apoptosis* (see figure 1.16 on page 34). Thus the first functional block of “regulation of apoptosis” has clearly two main subunits which internally structure it and disrupt its homogeneity.

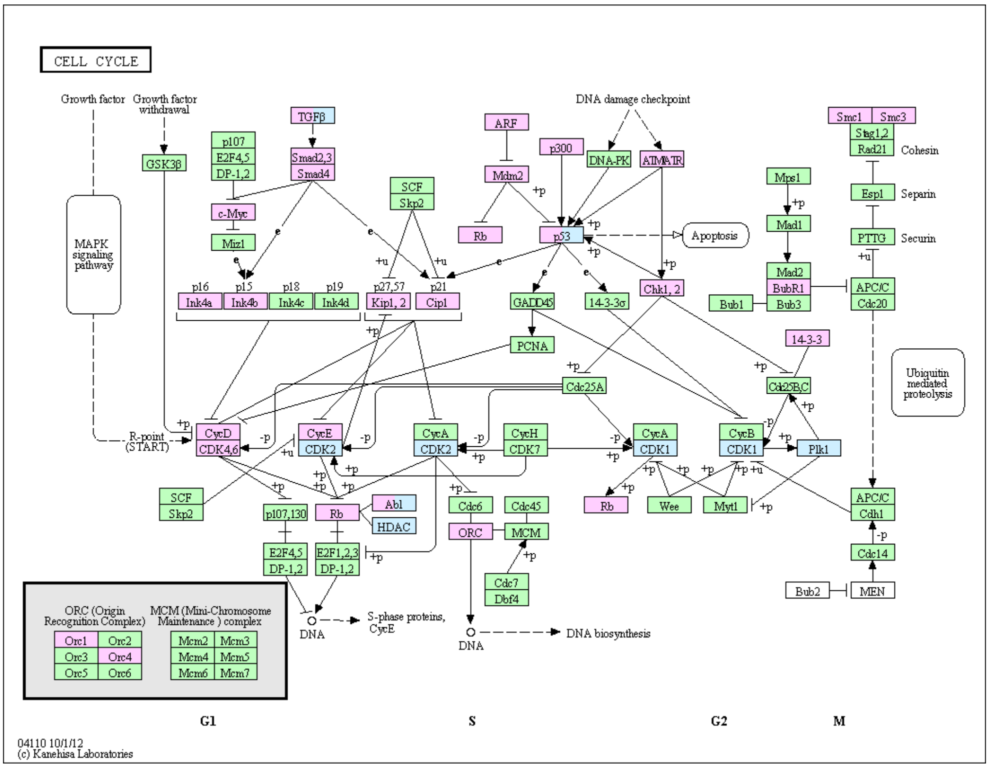
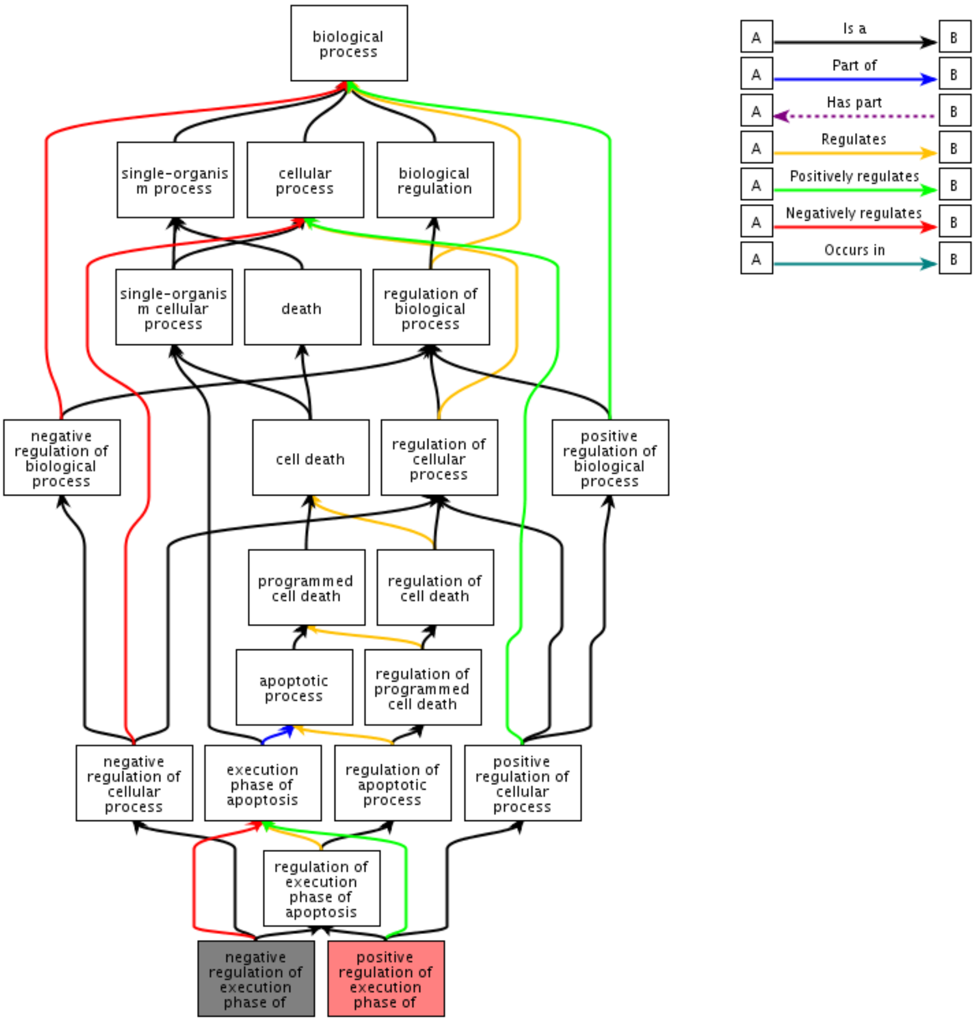


Figure 1.15: *Cell cycle* KEGG pathway. We can appreciate how gene **p53** is a hub in the pathway topology.

Despite the evident inhomogeneity of *gene sets*, very little work has been carried out to develop analytical methods accounting for it. This is partially due to the expectation created by some new *systems biology* trends which aim to use database information more exhaustively, in a *path graph* approach which will take advantage of the network structure of the *gene sets* when available. But is also due to the fact that, in order to correct for the lack of homogeneity of a *gene set*, its internal “irregularity” has to be somehow estimated before it can be controlled for. Such estimation has to be done empirically, what, considering the state of the genomic data repositories, is a cumbersome task. Moreover, it is ultimately dependent of the type of genomic data analyzed and the biological “meaning” of the ongoing experiment.

But above technical difficulties, providing estimations of the internal coherence of *gene sets* is an interesting goal on its own; it sheds light


biological process
single-organism process
cellular process
biological regulation
single-organism cellular process
death
regulation of biological process
negative regulation of biological process
cell death
regulation of cellular process
positive regulation of biological process
programmed cell death
regulation of cell death
apoptotic process
regulation of programmed cell death
negative regulation of cellular process
execution phase of apoptosis
regulation of apoptotic process
positive regulation of cellular process
regulation of execution phase of apoptosis
negative regulation of execution phase of
positive regulation of execution phase of
A
Is a
B
Part of
Has part
Regulates
Positively regulates
Negatively regulates
Occurs in

QuickGO - http://www.ebi.ac.uk/QuickGO

Figure 1.16: *Regulation of apoptosis* GO term.

on the biological processes themselves, as well as in our understanding and modeling of them through databases. Furthermore, once certain empirical description of gene relationships within *gene sets* is available, using it to improve *gene sets analysis* methods is an independent but closely related duty.

This double exercise of describing internal coherence of *gene sets* and providing *gene sets analysis* methods to take advantage of such description is what we did in **Montaner** et al. 2009 for the case of gene expression analyses. In this work we indicate how general gene expression studies implicitly assume a hypothesis of co-expression of the genes constituting *gene sets*. Then we used public data from thousands of microarrays to demonstrate that the internal co-expression of GO terms and KEGG pathways is not as coherent as firstly thought. We provided empirical estimation of the correlation between pairs of genes and developed a correlation distance from each gene to each *gene set*[1]. Finally we showed how such *distance* between genes and *gene sets* can be fruitfully incorporated into *gene set analysis* via *logistic regression* methodologies using a *weighed* schema.

As indicated in the preceding section, *logistic regression* models allow for many different extensions and generalizations. One of them is the possibility of weighting *cases* when fitting the model[2]. In this schema, besides the *dependent* and *independent* variables, a *weight* is provided for each case or observed individual in the dataset (see figure 1.17 on page 36 for an example). Such *weights* indicate the relevance we want to give to each of our observations or cases when estimating the model parameters from data. Observations with *high weights* will be considered to be more important and the estimated model will be prone to "explain" the dependence between the *response* and *independent* variables for such cases. Conversely, *low weights* cases will be less important to the algorithm estimating the model parameters; hence, the final model will not fit such observations as well as the highly weighted.

To give an intuition of how weights affect the model estimation in practice, we could say that, if case A doubles the weight of case B, then, the "importance" of A in the fitting is such as if the observations B would appear duplicated in our dataset. The figure 1.18 on page 36 shows a dataset that will provide a similar *logistic regression* fitting than the data represented in figure 1.17 on page 36.

[1] Using the Gene Ontology database and the KEGG Pathways repository as the paradigmatic *gene set* databases.

[2] Indeed the possibility including weighted observation is proper of the *generalized linear models*, not just of the *logistic* regression ones.

Response	Independent	Weights
1	1.23	0.1
1	2.74	0.3
0	-0.34	0.2
1	1.32	0.1
0	-2.02	0.1
0	0.45	0.1
0	0.93	0.1

Figure 1.17: A representation of a weighed dataset. Each case or row has values for its dependent and independent variables and also, an assigned weight. In this example, for instance, the second case is three times more important than the first one and the third case is two times more important than the first one. All remaining observations are given the same weight as the first one and hence, all of them are considered to be equally important.

Response	Independent	Weights	
1	1.23	0.1	
1	2.74	0.1	∗
1	2.74	0.1	∗
1	2.74	0.1	∗
0	-0.34	0.1	+
0	-0.34	0.1	+
1	1.32	0.1	
0	-2.02	0.1	
0	0.45	0.1	
0	0.93	0.1	

Figure 1.18: A representation of a dataset *equivalent* to the one presented in figure 1.17 (page 36). Cases marked with ∗ are the *unfold* of the case weighed with 0.3 in figure 1.17, cases marked with + are the *unfold* of the case weighed with 0.2 in figure 1.17.

In the work presented in **Montaner** et al. 2009, we empirically derived a *correlation distance* between genes and *gene sets*. In the context of gene expression, such *distance* accounts for the amount of coordination between the level of expression of the gene and the levels of expression of the remaining genes within the *gene set*. Genes being "close" to the *gene set* are those highly correlated with most of the genes in the block. In practical terms it means that, across biological conditions, the gene follows the same expression level pattern than the main bulk of genes of the *gene set*, acting as one more piece of the "biological machine" described by the annotation block. Genes having a "long" distance to the *gene set* are less correlated to the main core of genes within the functional block, indicating that the gene may be acting as part of the block just in few biological conditions. Long distances may even derive from a negative correlation between the gene and the *gene set*, which occurs for instance when a gene is an *inhibitor* within a pathway.

Also in **Montaner** et al. 2009 we used the *inverse*[1] of such correlation distance as a weight to indicate the importance a gene has when analyzing each *gene set*. For the analysis of each *gene set*, a *logistic regression* model was applied as described in section 1.4, but this time, the weights empirically derived from the correlation distances would be provided to the algorithm. Thus, the estimated model would reflect the real biological influence that the genes have over *gene sets*, and the global interpretation of the *gene set analysis* results would be biologically more consistent, as we demonstrated in the paper.

It should be noticed here that the same gene will have different assigned weights when considering different *gene sets*. This is so because the different *gene sets* are constituted by different groups of genes and hence, the estimated correlations between one gene and several functional blocks will differ among them.

Some other authors have considered *"reducing a gene set to its core members that chiefly contribute to the statistical significance of the differential expression of the initial gene set"* (Dinu et al., 2009). The relevance of our approach comes first from its versatility due to the flexibility of the *logistic regression* models. Any weighting schema other than the one we proposed for gene expression studies, may be straightforward applied using our *gene set methodology*. Thus, we keep on with our philosophy of tackling complex analyses following a "modular approach" (see previous

[1] The inverse of a distance D is the quantity $1/D$. Using such transformation to define the *weight* we obtain high weights for the genes that are "close" to the *gene set* and low weights for the genes that have "long" distance the *gene set*.

section 1.1 on page 2). Also our *gene set analysis* approach is not just restricted to gene expression studies. It may be easily applied to any kind of genomic data as the *gene level* analysis and the *gene set* level analysis are separated; in section 3.4 (page 58) we further consider this issue.

But the most outstanding improvement derived from our methodology is that, for the first time, we consider a *non discrete* membership of a gene to a *gene set*. We showed how, for any gene, the simplistic "flat" boolean state of being *in* or *out* of a *gene set* registered in databases, may be modulated in a meaningful *continuous* way when extra empirical information is available, yielding in more robust *gene set analysis* results.

1.6 The multidimensional context

Hitherto we have been approaching *gene sets analysis* from a *unidimensional* perspective. This means that, so far, we have been considering the simplest context in which a unique genomic measurement needs to be interpreted. Such simple analytical scenario rises in biological experiments where a single genomic characteristic is measured. Most biological experiments are such. They record gene expression measurements, genomic variants, methylation status and so on; but in an isolated way. Generally this is enough for testing concise hypotheses, focused on precise cellular mechanisms. Nevertheless, a growing amount of experiments are currently collecting several of those genomic measurements at a time. Improvement of the technologies besides their lowering of prices are making complex experiments more appealing to researchers.

As explained in section 1.4 (page 22), most genomic experiments end up yielding a *ranking index* which organizes the genes according so some biological condition: differential expression, methylation status or copy number are some of the possibilities given by modern technologies. When such *index* is put together with the binary variable indicating gene membership to a *gene set* (see figure 1.13 on page 29), *gene set analysis* methods can be applied in order to make a functional interpretation of the experimental data. Particularly, in this thesis, we have described how to use *logistic regression models* in order to carry out such *gene set analysis*.

When not one but several genomic measurements are taken in the same experiment, each of them is finally summarized by a different *ranking statistic*. We can imagine, for instance, an experiment in which researchers collect gene expression data and genomic variation data in a diseased versus control design. In such setup, the primary data analysis will return, for each gene[1], a first measurement of differential expression and a second indicator of genomic variant association to disease. When performing the functional interpretation of such results, for each *gene set* considered, a third binary variable indicating gene membership to the *gene set* will be added. Thus, the data table of figure 1.13 (page 29) would be extended as in figure 1.19.

Response var.	Independent var. 1	Independent var. 2
1	1.23	-2.3
1	2.74	1.4
0	-0.34	-0.1
1	1.32	1.5
0	-2.02	2.0
0	0.45	-1.9
0	0.93	0.8
...	...	...

Figure 1.19: A representation of the combined information from two ranking indexes. The experimental information reflected in the two continuous variables goes along with the annotation indicator.

Conventional *gene set analysis* methods will not be able to to deal straightforward with such *multidimensional* setup. Of course, the functional interpretation of the two genomic measurements can always be done separately, but if the experiment was designed to collect both ge-

[1] We are here simplifying the example by assuming that the gene would be the feature of interest to researchers, but transcripts, for instance, could also be used.

nomic characteristics, probably the interest remains in jointly analyzing them, considering possible synergies between them.

In **Montaner** and Dopazo 2010 we demonstrated how *logistic regression models* can easily be extended to to handle such *multidimensional* scenario. The basic idea in the paper can be formulated extending equation 1.1 on page 23 as follows

$$\log \frac{\pi}{1-\pi} = \kappa + \alpha \; x_1 + \beta \; x_2 \tag{1.5}$$

As in equation 1.1, in this formula, the parameter π represents the probability that a gene is annotated under the *gene set* being tested. Its *logit* transformation, $\log \frac{\pi}{1-\pi}$, is modeled as a function of, not one variable x, as in equation 1.1, but two: x_1 and x_1, each of them representing one of the two genomic *indices* under consideration. The parameters α and β are the *slopes* and their fitted value describes how the analyzed *gene set* is related to both genomic characteristics. The parameter κ, the *intercept* is generally not useful for the interpretation of the model.

Hence again, our modular approach to data analysis immediately let us take advantage of the flexibility of the *regression models*. Thus we can straightforward incorporate two[1] genomic characteristics into the *gene set analysis*. But, as indicated in **Montaner** and Dopazo 2010, the *gene set analysis* results derived from equation 1.5 will not be substantially different from those carried out applying conventional methods independently to each of the genomic measurements. Because of that, the actual model proposed in **Montaner** and Dopazo 2010 for the combined *functional profiling* of two genomic measurements, incorporates an *interaction term* as follows:

$$\log \frac{\pi}{1-\pi} = \kappa + \alpha \; x_1 + \beta \; x_2 + \gamma \; x_1 x_2 \tag{1.6}$$

[1] Indeed, theoretically, the model can be extended to as many genomic characteristics as desired.

In this equation, the *interaction term* γ represents the departure form the *additive* model stated in equation 1.5. For any *gene set*, the fitted value of γ reflects how much of the probability of genes being annotated under the *functional term*, is explained by the "combined" effect of the two genomic conditions x_1 and x_2.

The inclusion of the *interaction term* in the model is what supposes a real advantage over the utilization of conventional *gene set* methods applied independently over the two conditions. In biological sense, the *interaction* accounts for the combined effect of the two genomic characteristics in modulating the biological function described by the *gene set*. Our new approach lets the researchers discover relationships which will go unnoticed in the independent analysis of each of the genomic characteristics at a time. We can for instance explore gene expression and genomic variation measurements over the same individuals, and discover *gene sets* which are only activated when their genes are differentially expressed and mutated. As highlighted in our work, the relevance of such *gene sets* will be dismissed in the conventional analysis of the expression and the variation separately.

Our paper **Montaner** and Dopazo 2010 is fully reproduced in chapter 5 on page 77.

Chapter 2

Genomic Data and Analysis Tools

2.1 Montaner 2006 overview

This chapter presents the first publication from those appointed to outline this thesis: **Montaner** et al. 2006. The paper described the fourth version of GEPAS, a web tool devised for the analysis of microarray data.

Experience form previous releases of the tool had proved the advantages of integrating diverse analytical methods into the same suit. GEPAS users could, at that point, easily perform complex statistical analyses covering most general transcriptomic experiments. Keeping this spirit, GEPAS 2006 hold important technological improvements in its programming together with an enlarged collection of statistical methodologies addressing each analysis step.

But there where two aspects of the new version that became remarkable in the subsequent development of our concept of genomics: The first main change was to include tools for the analysis of other than transcriptomic data. Genomic copy number, for instance, could be since then analyzed using GEPAS. The second most relevant characteristic was the extension of the options available for *functional profiling*. Since 2006, results form any analysis ran by GEPAS could be further *functionally* explored using Babelomics modules (see chapter 1).

Thus we became aware that, following the analysis of any genome scale experimental data, a functional profiling step is almost compulsory. We also showed how the functional interpretation of a genomic analysis can be done independently of the methods or algorithms firstly used to explore the experimental data. Using GEPAS and Babelomics we had, for instance, the possibility of using many statistical algorithms to test for differential expression, and still, apply a unique functional interpretation approach . . . provided that it was flexible enough.

This acknowledgments settled down the requirements to be made to the GSA algorithms we developed later: Being able to deal with several genomic measurements at a time. And being general enough not to depend of the kind of genomic data, nor the statistics employed in the gene level analysis.

2.2 Paper

Montaner et al. 2006 paper is printed in this section.

GEPAS website may still be accessed at:
`http://www.gepas.org`

A link to Babelomics, the web tool within which GEPAS is now embedded, can be found at:
`http://www.babelomics.org`

The online version of the paper can be found at
`http://nar.oxfordjournals.org/content/34/suppl_2/W486.full`

W486–W491 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl197

Next station in microarray data analysis: GEPAS

David Montaner[1,2], Joaquín Tárraga[1,2], Jaime Huerta-Cepas[1,2], Jordi Burguet[1], Juan M. Vaquerizas[1], Lucía Conde[1], Pablo Minguez[1], Javier Vera[3], Sach Mukherjee[4], Joan Valls[5], Miguel A. G. Pujana[5], Eva Alloza[1], Javier Herrero[6], Fátima Al-Shahrour[1] and Joaquín Dopazo[1,2,*]

[1]Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Autopista del Saler 16, E46013, Valencia, Spain, [2]Functional Genomics Node, INB, CIPF, Autopista del Saler 16, E46013, Valencia, Spain, [3]INB—BSC, Jordi Girona 29, Edifici Nexus II, E-08034 Barcelona, Spain, [4]Pattern Analysis and Machine Learning Group, Department of Engineering Science University of Oxford, Oxford OX1 2JD, UK, [5]Translational Research Laboratory, Catalan Institute of Oncology, Institut d'Investigació Biomèdica de Bellvitge, L'Hospitalet, 08907 Barcelona, Spain and [6]Ensembl Team, EMBL-EBI, Hinxton, Cambridge, UK

Received February 14, 2006; Revised and Accepted March 21, 2006

ABSTRACT

The Gene Expression Profile Analysis Suite (GEPAS) has been running for more than four years. During this time it has evolved to keep pace with the new interests and trends in the still changing world of microarray data analysis. GEPAS has been designed to provide an intuitive although powerful web-based interface that offers diverse analysis options from the early step of preprocessing (normalization of Affymetrix and two-colour microarray experiments and other preprocessing options), to the final step of the functional annotation of the experiment (using Gene Ontology, pathways, PubMed abstracts etc.), and include different possibilities for clustering, gene selection, class prediction and array-comparative genomic hybridization management. GEPAS is extensively used by researchers of many countries and its records indicate an average usage rate of 400 experiments per day. The web-based pipeline for microarray gene expression data, GEPAS, is available at http://www.gepas.org.

INTRODUCTION

It is quite common that the introduction of a new technology is accompanied by claims and promises which on many occasions cannot be fulfilled. This hype is then followed by a wave of disappointment against the technology. Fortunately, as it is reaching a certain degree of maturity, DNA microarray technologies do not seem to have followed this fate. During an initial period, DNA microarray publications were dealing with issues such as reproducibility and sensitivity. Many classical microarray papers dating from the late nineties were mere proof-of-principle experiments (1,2), in which only cluster analysis was applied. Later, sensitivity became a main concern as a natural reaction against quite liberal interpretations of microarray experiments made by some researchers, such as the fold criteria to select differentially expressed genes. It was soon obvious that genome-scale experiments should be carefully analysed because many apparent associations happened merely by chance (3). In this context, different methods for the adjustment of *P*-values, which are considered standard today, started to be extensively used (4,5). More recently the use of microarrays as predictors of clinical outcomes (6), despite not being free of criticisms (7), fuelled the use of the methodology because of its practical implications. There are still some concerns with the cross-platform coherence of results but it seems clear that intra-platform reproducibility is high (8) and, despite the fact that gene-by-gene results are not always the same, the biological themes emerging from the different platforms are increasingly consistent (9). That points to the importance of the interpretation of experiments in terms of their biological implications instead of a mere comparison of lists of genes (10,11).

Keeping a pace with the trends mentioned above, Gene Expression Profile Analysis Suite (GEPAS) has been growing during the last 4 years. In the first release it was more oriented towards clustering and data preprocessing (12). Successive releases showed a package more oriented towards gene selection, class prediction and the functional annotation of experiments (13,14). The version presented here include several new

[*]To whom correspondence should be addressed. Tel: +34 963289680; Fax: +34 963289701; Email: jdopazo@cipf.es

modules, some of which are new while other ones constitute already available tools completely rewritten including new functionalities. GEPAS is not a simple web server, but it constitutes one of the largest resources for integrated microarray data available over the web. It has been working for more than four years having by the end of year 2005 an average of 400 experiments analysed per day summing up over all of their modules. GEPAS is used by researches worldwide as can be seen in the usage map, where all the sessions are mapped to its geographic location (http://bioinfo.cipf.es/access_map/map.html). It also offers on-line tutorials that can be used in courses. In the new version (3.0) we present new modules for the normalization of Affymetrix experiments, for differential gene expression, for the evaluation of cluster quality and another module for array-comparative genomic hybridization (Array-CGH) data management. Also, another conceptual novelty is the connection of GEPAS to the PupaSuite tools (15–17), which offers the possibility of analysing polymorphisms at the light of the results of the gene expression analysis.

GENERAL OVERVIEW

GEPAS aims to tackle the most common problems in microarray data analysis in a simple but rigorous way. Thus, after an essential step of normalization, there are different 'workflows', or sequences of steps, that can be followed, depending on the aim of the experiment: class discovery, differential gene expression, class prediction or genomic copy number estimation, just to cite the most common objectives of microarray experiments. Class discovery, either in genes or in experiments, is achieved by using clustering methods. GEPAS includes some commonly used clustering methods such as hierarchical clustering (18), *SOTA* (19,20), *SOM* (21), *K-means* (22) and SOM-Tree (23). The evaluation of cluster quality, a scarcely addressed issue, has been implemented here in the Cluster Accuracy Analysis Tool (CAAT) module (see below). Differential gene expression implies finding genes with significant differences in expression between two or more classes, related to a continuous experimental factor (e.g. the concentration of a metabolite) or to survival data. A new, more complete module for differential gene expression is presented in this new version of GEPAS (see below). The module *Tnasas* for class prediction implements different classifiers, such as diagonal linear discriminant analysis (DLDA) (24), nearest neighbour (NN) (25), support vector machines (SVM) (26), random forest (27) and shrunken centroids (PAM) (28) of known efficiency as class predictors using microarray data (24). Cross-validation error is calculated in a way to avoid the well-known selection bias problem (29,30). See *Tnasas* help (http://tnasas.bioinfo.cipf.es/cgi-bin/docs/tnasashelp) for a more detailed description of the methods and error estimation strategy. *Array-CGH* (31) can be analysed through the module *ISACGH* that allows predicting copy number, relating these values to gene expression and performing functional annotation through the babelomics (11) suite. Finally, functional annotation is carried out with the babelomics suite which can be used either as an independent suite or as an integrated part of the GEPAS. Figure 1 illustrates, following the metaphor of a subway line, the interconnections of the different tools in the GEPAS environment.

NORMALIZATION AND PREPROCESSING

GEPAS now implements normalization facilities for both two-colours and Affymetrix arrays. *DNMAD* (32) module performs normalization in two-colour arrays using print-tip loess (33) with a number of different options. *DNMAD* can input Genepix (Axon instruments) GPR files. The module *expresso* normalizes Affymetrix CEL files using standard Bioconductor (34) tools; in particular the package affy (35). Besides its friendly web interface we provide the user with the speed and above all the physical memory available in our server.

More information can be found in the corresponding tutorial web pages (http://bioinfo.cipf.es/docus/courses/on-line.html).

In addition, the *preprocessor* (36) module performs some preprocessing of the data (log-transformations, standardizations, imputation of missing values and so on).

CLUSTERING AND CLUSTER QUALITY ESTIMATION

Despite the fact that clustering is one of the most popular—albeit often improperly used (30)—methodologies in the analysis of microarray data there are very few alternatives for the estimation of the quality of the results found. We have included a module, *CAAT*, which provides many options for the visualization and intuitive manipulation of hierarchical and non-hierarchical clustering results. Many visualization modes, browsing options and cluster extraction possibilities are currently available. Moreover, *CAAT* provides some descriptive measures about each partition (average profiles, standard deviation profiles, inter and intra-cluster distances) as well as a global estimation of cluster quality by the silhouette method (37), which performs well, in noisy situations, such as microarray analysis (38). *CAAT* submits data to other tools such as the Babelomics (11) functional annotation suite or to *ISACGH* (Figure 1).

There is more detailed information in the *CAAT* documentation (http://bioinfo.cipf.es/docus/courses/on-line.html).

DIFFERENTIAL GENE EXPRESSION

This version of GEPAS includes new methods for differential gene expression analysis under different conditions. The old module *pomelo* has been replaced by the new module *T-rex* (Tools for RElevant gene seleXion) which is much faster and offers new tests for different situations. *T-rex* distinguishes among four conceptually different testing cases:

- *Finding genes differentially expressed between two discrete classes* (e.g. case/control and so on). A number of authors (39,40) have found that the classical t-statistic, which was widely used in early work on the analysis of differential expression, can be highly unreliable for microarray data. Problems arise mainly as a consequence of statistical issues relating to the SD term in the denominator of the *t*-statistic. For example, many non-differentially expressed genes may by chance have small observed SDs, which may cause these genes to be erroneously selected. GEPAS now also implements different new tests:
  - The *t*-test, which is still available.

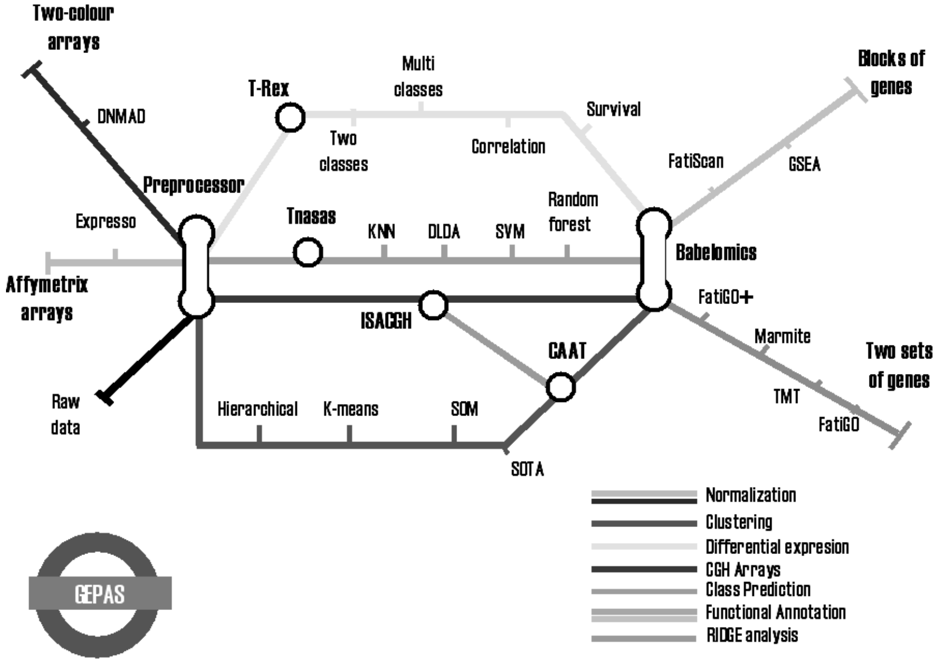


Figure 1. Map of GEPAS functionalities as a subway line. Data (Affimetrix, two-colour or raw) are introduced from the left side and pass through the preprocessor. Then different types of analyses can be performed: gene selection (T-rex) in different situations (two or more classes, correlation or survival; see text for details) or class discovery (Tnasas) are two types of supervised analyses. Array-CGH data can be analysed through the red line ISACGH. Unsupervised analysis can also be performed using different methods. CAAT allows to map co-expressed genes on their chromosomal coordinates allowing the study of RIDGES (54). All the tools end up in Babelomics (11), that allows for two different types of analysis: comparison of two sets of genes of analysis or blocks of functionally related genes.

- An empirical Bayes methodology that allows fitting hierarchical mixture models to identify differentially expressed genes (41). One of the advantages of this methodology is that it fits a global model taking into account all genes in the dataset.
- A novel test for the analysis of microarray data by combining inference for differential expression and variability (CLEAR-test) (J. Valls, M. Grau, X. Sole, P. Hernandez, D. Montaner, J. Dopazo, M. A. Peinado, G. Capella, M. A. G. Pujana and V. Moreno, manuscript submitted). Most tests evaluate differential expression by using estimated variability, but no inference is made in terms of the variability itself. CLEAR-test evaluates both whether genes show large fold changes and whether their variability is high.
- A data-adaptive approach to the analysis of differential expression, in which an effective test statistic is learned directly from microarray data. This approach has been shown to ameliorate many of the problems associated with both the t-statistic and simple moderated statistics like SAM (42), and to produce good results under a range of conditions (43).

- *Finding genes differentially expressed between more than two classes* (e.g. different types of cancers and so on) Together with the classical ANOVA methodology we make available the same CLEAR test mentioned above (41). While the mathematical treatment of this kind of data is similar to that of two classes, in our tools, we separate the case when more than two classes are available because of its different conceptual implications.
- *Finding genes whose expression is correlated to a continuous variable* (e.g. the level of a metabolite). Regression analysis of gene expression on any numerical independent variable has been implemented. C routines have been compiled for the particular architecture of our computers in order to achieve the maximal speed. Estimates of Pearson's and Spearman's correlation coefficients as well as *P*-values for testing the null hypothesis of no correlation can also be obtained with *T-rex*.
- *Finding genes whose expression is related to survival times*. GEPAS uses C routines to estimate a Cox proportional hazards regression model (44). Right censored data are allowed as well as replicates in the survival times. Censoring variables should be provided by the researcher together with survival times that may be replicated.

When appropriate, *P*-values adjusted for multiple testing are provided. Three methodologies are implemented. One of them controls the FWER (family-wise error rate) (45) while the others control the FDR (false discovery rate) (46). Our implementations make use of the *p.adjust* function in the *stats* R package and the *qvalues* package (47) from Bioconductor.

FUNCTIONAL ANNOTATION

Functional annotation of the experiments gives clues to the researcher for the interpretation of the experiment. There are a number of tools that make use of gene functional annotations to try to understand the global changes in gene expression in microarray experiments (48), but probably one of the most complete packages in this respect is the Babelomics suite (11,49). This suite of programs for functional annotation of genome-scale experiments has undergone a deep modification described in detail elsewhere (49). In brief, Babelomics can now compare two groups of genes and test simultaneously for the significant over-abundance of diverse biological themes such as GO terms, KEGG pathways, Interpro motifs, Swissprot keywords, Transfac® motifs, CisRed motifs, relative abundance in tissues and bioentities extracted from PubMed, with the proper multiple testing adjustment. This is carried out by the *FatiGO+* module, the evolution of the *FatiGO* program (50). Additionally there are two modules designed to search for functionally related blocks of genes that are co-ordinately over- or under-expressed using both the *FatiScan* (51) or the *GSEA* (52) algorithms.

Despite its general scope (Babelomics is not restricted to microarrays but applicable to any type of large-scale experiment), and the possibility of being used alone as an independent resource, the Babelomics suite has been fully integrated into GEPAS. Modules of gene selection (*T-rex*) or class prediction (*tnasas*) can submit the genes selected as relevant to the *FatiGO+* module for testing against the rest of genes. Likewise, the modules for clustering (*hierarchical, k-means, SOM, SOTA*) through their cluster' viewers or through *CAAT*, can submit the genes within the selected cluster to be tested against the rest of genes. Similar operation can be performed from within *ISACGH*, with the genes contained in the selected chromosomal region. Moreover, arrangements of genes can be sent from *T-rex* to the *FatiScan* to test blocks of functionally related genes tha are co-ordinately over- or under-expressed. Sets of arrays can also be submitted to *GSEA* with the same purpose.

ARRAY-CGH

Genetic aberrations, which are the molecular basis of many diseases, have classically been studied through CGH. The introduction of microarray-based CGH methods (array-CGH) has revolutionized this methodology in terms of resolution and throughput (31,53) but, at the same time, has generated a need for new algorithms and software for dealing with this type of data. We have included in GEPAS a new module, *ISACGH*, which completely replaces the old viewer InSilicoCGH (14). *ISACGH* includes two new and efficient methods for accurate estimation of genomic copy number from array-CGH hybridization data, integrated into a web-based system that allows, for the first time, the combined study of gene expression and genomic copy number. Several visualization options offer a convenient representation of the results. Moreover, the link to the Babelomics (11,49) tools allows, for the first time in a tool of this type, the production of functional annotations (using different relevant biological information such as gene ontology, pathways, etc.) for the detected chromosomal regions of interest (amplified or deleted). We use the DAS technology (Distributed Annotation System; see http://www.biodas.org/), that allows a remote mapping of information (our predictions) from a server (our server) to a client (Ensembl), to represent the ISACGH predictions and data onto the Ensembl chromosomal coordinates.

ISACGH generically maps data onto their chromosomal coordinates. So, beyond to map genomic hybridisations any other data can be mapped. Thus *CAAT* can send to *ISACGH* groups of co-expressing genes, which might be useful for defining regions of increased gene expression, also known as RIDGES (54).

Polymorphisms affecting gene expression

Although the study of regulatory polymorphisms is not new, there has been a recent revival of interest in them mainly because of the availability of high-throughput data and methodologies that allows their characterisation (55). The corresponding GEPAS modules (*CAAT, tnasas* and *T-rex*) have a unique feature in this regard: the possibility of connecting the genes found to be regulated in a microarray experiment to possible regulatory SNPs in such genes. In particular, clustering and gene selection methods can be connected to the *PupaSuite* (15–17).

DISCUSSION

GEPAS is a long-term project that aims to provide the scientific community with an advanced set of tools for microarray data analysis, without renouncing to an easy and intuitive use. It has been running uninterruptedly for more than four years and has grown to include more tools as new algorithms were introduced in the microarray data analysis arena (12–14). The GEPAS team has intended to deliver a coherent set of state-of-the-art and widely established algorithms, running away from building a simple collection of as-much-as-possible tools. Actually, any new tool included is the response to a new or emerging requirement requested by our users. As the Functional Genomics node of the Spanish Institute of Bioinformatics (INB; http://www.inab.org) and being part of the Spanish Network of Cancer Centers (RTICCC; http://www.rticcc.org) we have a direct contact with researchers from which we get much of the feedback necessary to build up a useful tool. GEPAS, integrated with the Babelomics suite (11,49), provides the tools for performing the most common analyses of microarray data. Moreover, it has been conceived as a workflow that helps the user to carry out a series of consecutive steps of analysis with simple mouse clicks. GEPAS has been designed to take full advantage of the properties of the web: connectivity, cross-platform functionality and remote usage. Its modular architecture allows easy implementation of new tools and facilitates the connectivity of GEPAS from and to other web-based tools.

The user of GEPAS ranges from the experimentalist with not much experience in bioinformatics and no deep statistical skills, interested only in data analysis, to the bioinformatician that invokes some of the tools remotely for different purposes.

GEPAS is running in a high-end cluster (with 20 dedicated AMD Opteron CPUs at 2.4 GHz) with a large amount of RAM (6 GB). This allows to use tools (e.g. normalization tools are highly RAM-consuming) that usually are beyond the capabilities of the hardware available to many end users.

In addition, there is a teaching programme related to GEPAS (see http://bioinfo.cipf.es/docus/courses/courses.

html) with on-line tutorials that can be freely used (http://bioinfo.cipf.es/docus/courses/on-line.html).

Although other alternatives are available for microarray data analysis, there is no other similar resource over the web with the number of possibilities offered by GEPAS.

ACKNOWLEDGEMENTS

We are deeply acknowledged to our external advisors Sandrine Dudoit, Yves Moreau and John Quackenbush for their comments and support. This work is supported by grants from Fundació La Caixa, Fundación BBVA, MEC BIO2005-01078 and NRC Canada-SEPOCT Spain. The Functional Genomics node (INB) is supported by Genoma España. Funding to pay the Open Access publication charges for this article was provided by Genoma España.

Conflict of interest statement. None declared.

REFERENCES

1. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
2. Perou,C.M., Jeffrey,S.S., van de Rijn,M., Rees,C.A., Eisen,M.B., Ross,D.T., Pergamenschikov,A., Williams,C.F., Zhu,S.X., Lee,J.C. *et al*. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA*, **96**, 9212–9217.
3. Ge,H., Walhout,A.J. and Vidal,M. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*., **19**, 551–560.
4. Benjamini,Y. and Yekutieli,D. (2001) The control of false discovery rate in multiple testing under dependency. *Ann. Stat*., **29**, 1165–1188.
5. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
6. van't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton M.J., Witteveen,A.T. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
7. Simon,R. (2005) Roadmap for developing and validating therapeutically relevant genomic classifiers. *J. Clin. Oncol*., **23**, 7332–7341.
8. Moreau,Y., Aerts,S., De Moor,B., De Strooper,B. and Dabrowski,M. (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet*., **19**, 570–577.
9. Bammler,T., Beyer,R.P., Bhattacharya,S., Boorman,G.A., Boyles,A., Bradford,B.U., Bumgarner,R.E., Bushel,P.R., Chaturvedi,K., Choi,D. *et al*. (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods*, **2**, 351–356.
10. Al-Shahrour,F. and Dopazo,J. (2005) In Azuaje,F. and Dopazo,J. (eds), *Data analysis and visualization in genomics and proteomics*. Wiley, West Sussex, UK, pp. 99–112.
11. Al-Shahrour,F., Minguez,P., Vaquerizas,J.M., Conde,L. and Dopazo,J. (2005) BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res*., **33**, W460–W464.
12. Herrero,J., Al-Shahrour,F., Diaz-Uriarte,R., Mateos,A., Vaquerizas,J.M., Santoyo,J. and Dopazo,J. (2003) GEPAS: A web-based resource for microarray gene expression data analysis. *Nucleic Acids Res*., **31**, 3461–3467.
13. Herrero,J., Vaquerizas,J.M., Al-Shahrour,F., Conde,L., Mateos,A., Diaz-Uriarte,J.S. and Dopazo,J. (2004) New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res*., **32**, W485–W491.
14. Vaquerizas,J.M., Conde,L., Yankilevich,P., Cabezon,A., Minguez,P., Diaz-Uriarte,R., Al-Shahrour,F., Herrero,J. and Dopazo,J. (2005) GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. *Nucleic Acids Res*., **33**, W616–W620.
15. Conde,L., Vaquerizas,J., Dopazo,H., Arbiza,L., Reumers,J., Rousseau,F., Schymkowitz,J. and Dopazo,J. (2006) PupaSuite: finding functional SNPs for large-scale genotyping purposes. *Nucleic Acids Res.*, in press.
16. Conde,L., Vaquerizas,J.M., Ferrer-Costa,C., de la Cruz,X., Orozco,M. and Dopazo,J. (2005) PupasView: a visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purposes. *Nucleic Acids Res*., **33**, W501–W505.
17. Conde,L., Vaquerizas,J.M., Santoyo,J., Al-Shahrour,F., Ruiz-Llorente,S., Robledo,M. and Dopazo,J. (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res*., **32**, W242–W248.
18. Sneath,P. and Sokal,R. (1973) *Numerical Taxonomy*. W.H. Freeman, San Francisco.
19. Dopazo,J. and Carazo,J.M. (1997) Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol*., **44**, 226–233.
20. Herrero,J., Valencia,A. and Dopazo,J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.
21. Kohonen,T. (1997) *Self-organizing maps*. Springer-Verlag, Berlin.
22. Hartigan,J. and Wong,M. (1979) A k-means clustering algorithm. *Appl. Stat*., **28**, 100–108.
23. Herrero,J. and Dopazo,J. (2002) Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns. *J. Proteome Res*., **1**, 467–470.
24. Dudoit,S., Fridlyand,J. and Speed,T. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc*., **97**, 77–87.
25. Ripley,B. (1996) *Pattern recognition and neural networks*. Cambridge University Press, Cambridge.
26. Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, NY.
27. Breiman,L. (2001) Random forests. *Machine Learning*, **45**, 5–32.
28. Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.
29. Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.
30. Simon,R., Radmacher,M.D., Dobbin,K. and McShane,L.M. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst*., **95**, 14–18.
31. Mantripragada,K.K., Buckley,P.G., de Stahl,T.D. and Dumanski,J.P. (2004) Genomic microarrays in the spotlight. *Trends Genet*., **20**, 87–94.
32. Vaquerizas,J.M., Dopazo,J. and Diaz-Uriarte,R. (2004) DNMAD: web-based diagnosis and normalization for microarray data. *Bioinformatics*, **20**, 3656–3658.
33. Smyth,G., Yang,Y. and Speed,T. (2003) In Brownstein,M. and Khodursky,A. (eds), *Functional Genomics: Methods and Protocols*. Humana Press, Totowa, NJ, Vol. 224, pp. 111–136.
34. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.
35. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
36. Herrero,J., Diaz-Uriarte,R. and Dopazo,J. (2003) Gene expression data preprocessing. *Bioinformatics*, **19**, 655–656.
37. Rousseeuw,P. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math*., **20**, 53–65.
38. Azuaje,F. (2002) A cluster validity framework for genome expression data. *Bioinformatics*, **18**, 319–320.
39. Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.
40. Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*., **4**, 210.
41. Kendziorski,C.M., Newton,M.A., Lan,H. and Gould,M.N. (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med*., **22**, 3899–3914.

42. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
43. Mukherjee,S., Roberts,S.J. and van der Laan,M.J. (2005) Data-adaptive test statistics for microarray data. *Bioinformatics*, **21**, ii108–ii114.
44. Klein,J.P. and Moeschberger,M.L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York.
45. Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.
46. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R Stat. Soc. [Ser B]*, **57**, 289–300.
47. Storey,J., Taylor,J. and Siegmund,D. (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R Stat. Soc. [Ser B]*, **66**, 187–205.
48. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
49. Al-Shahrour,F., Minguez,P., Tarraga,J., Montaner,D., Alloza,E., Vaquerizas,J.M., Conde,L., Blaschke,C., Vera,J. and Dopazo,J. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res.,* in press.
50. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
51. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2005) Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*, **21**, 2988–2993.
52. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
53. Albertson,D.G. and Pinkel,D. (2003) Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.*, **12**, R145–R152.
54. Caron,H., van Schaik,B., van der Mee,M., Baas,F., Riggins,G., van Sluis,P., Hermus,M.C., van Asperen,R., Boon,K., Voute,P.A. *et al.* (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, **291**, 1289–1292.
55. Wang,D.G., Fan,J.B., Siao,C.J., Berno,A., Young,P., Sapolsky,R., Ghandour,G., Perkins,N., Winchester,E., Spencer,J. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–1082.

Chapter 3

Gene Set Methods Rationale

This chapter advances some ideas and concepts that may help reading the next two articles presented for the thesis: **Montaner** et al. 2009 and **Montaner** and Dopazo 2010. Hopefully, it will also clarify the link between GEPAS, reported in the previous chapter, and the other two works. Some parts of the text are taken form a previous report we fist published as a book chapter, **Montaner** et al. 2008. The original writing (available in Appendix A) was entitled "New trends in the analysis of functional genomic data." Sometime later I am glad to see that those "trends" derived into robust analytical methods.

3.1 Replications of the same statistical test

Most analyses carried out using high throughput data (analyzed using GEPAS, for instance) consist of the repetition of the same statistical test for all genes in the dataset. As a result of such replicated analysis we get, for each gene, several estimates of statistical parameters: statistics, fold changes, p-values or confidence intervals. Thus, a huge amounts of *numbers* is yield by the analysis itself.

As in all biological experiments, after getting the results, researchers have to interpret them to answer those questions addressed by their rehearsal. To interpret results means facing what is *already known* with the *new information* provided by the experimental outcome. If there are few variables in the game, recalling what is *already known* is straight forward and so it is to face it with the *new information* carried by the data. On the other hand, in genome scale experiments, what is *already known* is what is stored in databases, and to face it with the high throughput results is a cumbersome task, unfeasible without the application of automatic procedures for *functional profiling*.

Trying to simplify things and being aware that most statistical methods were developed to test for a single hypothesis, researchers will usually correct p-values for multiple testing[1] before choosing a cut-off that will indicate the rejection of the null hypotheses, whichever it is.

Once chosen the genes with alternative pattern (meaning different form the one stated in the null hypothesis) the next step is to biologically interpret such departure from hypothesis. Different repositories of functionally relevant biological information such as Gene Ontology (Ashburner et al., 2000), KEGG (Kanehisa and Goto, 2000), InterPro, (Mulder et al., 2007) or Reactome (Vastrik et al., 2007) are available and can be used for the functional annotation of genome-scale experiments. Thus the functional properties of the selected genes can be analyzed.

Using the tools we developed at the CIPF, you could, for instance, carry out a differential expression using GEPAS, select the differentially expressed genes and send them to Babelomics to be functionally interpreted with the FatiGO module (Al-Shahrour et al., 2004, 2007b).

Such cut-off based approach has been recently referred to as “Over-Representation Analysis” by Khatri et al. 2012; in section 1.3 (page 18) we have referred to it as *functional enrichment analysis*. In this methods we can appreciate an interesting change in the philosophy of the analysis. The focus of attention is not anymore a single gene but a block of genes

[1] Classical p-value correction methods widely used in genomics are those proposed by Benjamini and Hochberg 1995 and Benjamini and Yekutieli 2001 Still there is a lot of ongoing work on this topic; see for instance Carvajal-Rodriguez and de Una-Alvarez 2011.

with a common biological meaning. This new way of looking at data provides, among others, obvious advantages for the biological interpretation of results as well as for the p-value adjustment. We just need to correct by the number of blocks, usually smaller than the number of genes.

But the "Over-Representation Analysis" carries its own drawbacks. Dopazo 2009 provides a detailed discussion of them, being the most detrimental the fact that, by discarding genes with p-values above the cut-off, we loose most of our information. Not only we loose the measurements taken over the discarded genes but also, the functional annotation that could be linked to them from repositories, making more difficult the biological interpretation of results (see figure 3.1 on page 53).

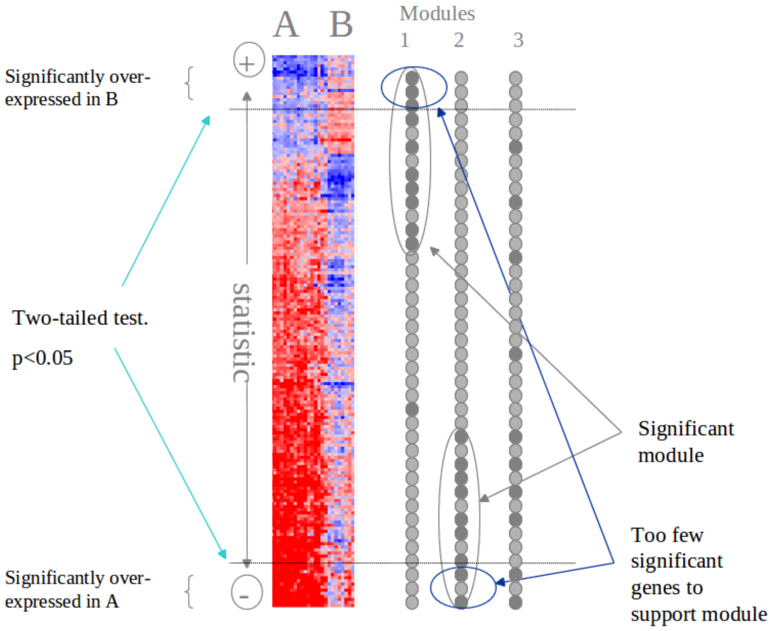

A
B
Modules
1
2
3
+
-
statistic
Significantly over-expressed in B
Two-tailed test.
p<0.05
Significantly over-expressed in A
Significant module
Too few significant genes to support module

Figure 3.1: Over-Representation Methods. The horizontal lines represent the cut-off points. Just the information in the extremes of the ranking is used for the *functional interpretation*; all the other information is dismissed.

Also, in biological sense, calling most relevant to those genes overtaking certain threshold is, somehow, a biased approximation. It is generally accepted that genes showing biggest changes between experimental conditions may play a key roll in biology. Nevertheless, as Khatri et al. 2012 indicates, it is likely as well that, small but coordinated changes in functionally related genes, that is, in *gene sets*, have decisive biological implications.

3.2 Whole-genome methods: Gene Set Analysis

Aiming to prevent such waste of information, some authors proposed to directly analyze the behavior of blocks of functionally related genes in a *whole-genome* context. In this new approach, the cut-off steep is skipped and all genes available in the experiment are considered. Hence, experimental and annotated information enter complete in the analysis which becomes, therefore, more powerful.

The Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003), the FatiScan (Al-Shahrour et al., 2007a) or the Global Test (Goeman et al., 2004), constitute examples of this type of approach inspired from systems biology. Khatri et al. 2012 refer to this second generation of algorithms as "Functional Class Scoring (FCS) Approaches", but the term "Gene Set Analysis" (GSA) is probably the most broadly used in publications.

All this methodologies address the general issue of whether the general expression pattern of a group of genes, for example a GO term or a KEGG pathway, changes across biological conditions. They study the relationship between the expression of the genes of the block of interest and a characteristic associated to each biological sample. Such characteristic may be a categorical condition, like the class of the microarray in the context of differential gene expression, or a continuous variable such as a level of a metabolite.

In a wonderful paper, Goeman and Buhlmann 2007 classify GSA methodologies as "competitive" or "self-contained":

> *A **competitive** test compares differential expression of the gene set to a standard defined by the complement of that gene*

> *set. A* ***self-contained*** *test, in contrast, compares the gene set to a fixed standard that does not depend on the measurements of genes outside the gene set.*

After this publication there has been an extensive debate on the advantages and disadvantages of each of the two approaches. Goeman et al. 2004, Mansmann and Meister 2005 or Dinu et al. 2007 advocate *self-contained* while Subramanian et al. 2005, Al-Shahrour et al. 2007a or Sartor et al. 2009 prefer the *competitive*.

The methods presented in this thesis, **Montaner** et al. 2009 and **Montaner** and Dopazo 2010, refine the ideas behind Al-Shahrour et al. 2007a and share underlying statistical models with Sartor et al. 2009; they also fall into the *competitive* methodologies classification.

Ultimately, to choose one or the other approach depends on the experimental context and on how researchers understand the *enrichment* or over/under expression of a *gene set*. As Dinu et al. 2009 states:

> *This fundamental disagreement on the concept of, not the methods for identifying, differentially expressed gene sets [. . .] is a key point in the debate between the self-contained versus competitive methods.*

In general, the *self-contained* view implies a stronger statement (Goeman and Buhlmann, 2007) meaning a more restrictive contour of the statistical hypothesis tested or a more ambitious description of the experimental reality. Hence, *self-contained* algorithms need to be more customized for each experimental design. They also need to be more exhaustive in the usage of the experimental information.

Usually this requirements are fulfilled via “subject sampling” p-value computation (Goeman and Buhlmann, 2007) and require the gene level analysis and the set level analysis be embedded within the same algorithm. But, the gene level analysis, which is always conditioned by the design of the experiment, cannot easily be modified, making these methodologies quite inflexible.

The approach in Goeman et al. 2004 for instance, was designed to handle a class comparison design; it required complete amendment of the methods and re-programing of the software before it could be used to analyze survival data (Goeman et al., 2005). On the contrary, this two experimental frames, class comparison and survival analysis, were easily tackled using our competitive approach: A first step will conduct differential expression or survival analysis using standard gene level methodologies.[1] In a second step FatiScan or a *logistic* model like those in Sartor et al. 2009, **Montaner** et al. 2009 or **Montaner** and Dopazo 2010 would be applied to carry out the GSA part of the job.

Finally, in the *self-contained* algorithms, the blending of the gene and *gene set* levels of analysis besides the imperative subject sampling approach, transport the potential shortcomings of the original expression data, to the final results of the analysis. If the dataset has small sample size, for instance, some methods are not applicable; if the expression data are not available p-values cannot be directly obtained (Fridley et al., 2010).

3.3 Competing over a ranking index

The block of genes is also the unit of interest of the *competitive* methods. The GSEA, FatiScan or the *logistic* model based methods[2] are similar to the Global Test and other *self-contained* approaches in that they are also used to discover groups of genes which overall expression pattern changes across biological conditions. Nevertheless, GSEA, FatiScan and the *logistic* models consider all genes in the data when analyzing each of the blocks. They compare the pattern of the genes of one block with the general pattern of the genes in the whole dataset, and call it significant

[1] GEPAS could be used for this purpose but also any other software that the researchers considered more suitable for their study; this exemplifies the complete separation of gene and the *gene set* levels in our functional profiling approach.

[2] Sartor et al. 2009, **Montaner** et al. 2009 and **Montaner** and Dopazo 2010.

if it is a *winner in that competition*. GSEA was particularly designed for the two class comparison context while FatiScan and the *logistic* models may be applied in a wider range of studies.

The rationale underlying this methodologies is that, if a property of genes can be described using a continuous index, a *ranking index*, then, the statistical distribution of such index within a functional block of genes can be compared to the general distribution of the index across all genes in the dataset. Hence, the almost[1] complete genome, provides the background distribution towards which we can compare the *gene set* of interest and compute statistical significance for it. Thus, we can asses whether the property described by the index is related to the characteristic which a priori defined the block of genes, that is, the biological function described by the annotation.

As said before GSEA is developed for the two class comparison. In this methodology, a signal-to-noise ratio comparing mean expression across classes is computed for each gene in the dataset. This statistic can be seen as a continuous index that ranks the genes according to their differential expression, from those more expressed in one of the biological conditions to those more expressed the second condition, and passing through those genes not differentially expressed. Then, given a block of genes, for instance a functional class that we may be interested in, we can compare the distribution of the signal-to-noise ratio of the genes in the block to the distribution of the same statistic in the remaining genes. If the values of the signal-to-noise ratio are, for instance, systematically higher in the genes of the block compared to the genes in the whole dataset, we will conclude that, as a block, the genes of the functional class of interest are over-expressed in one of the biological conditions.

GSEA uses a modification of the Kolmogorov-Smirnov test to asses differences between the signal-to-noise ratio in the class of interest and in the rest of the genes. Significance of the modified Kolmogorov-Smirnov

[1] The genes of the *gene set of* interest are drawn from such background.

statistic is computed in GSEA using permutations of the expression data.[1] The original expression data is permuted several times, the signal-to-noise ratios are calculated over each permuted expression dataset and the modified Kolmogorov-Smirnov statistic is computed over each new distribution of the signal-to-noise ratio. Thus GSEA can estimate the random variability of the Kolmogorov-Smirnov statistic and test its significance in the original data.

3.4 Detaching concepts and algorithms

FatiScan and the *logistic regression* methods follow the same analytical philosophy than GSEA but with a more general and flexible approach. FatiScan implements a segmentation test which checks for asymmetrical distributions of biological labels associated to genes ranked by any index, just as GSEA does. But there is a major difference between FatiScan and GSEA. FatiScan does not implement a permutation test to asses such asymmetry. Therefore, the algorithm that computes the index and the algorithm that analyses the distribution of the index are completely separated so the calculations can be done in two different steps. This means that FatiScan can be used to study the relationship between biological labels associated to genes and any type of experiment whose outcome is a sorted list of genes or a variable that can be used to rank genes according to some characteristic of interest. Block of genes sorted by differential expression between two experimental conditions can be studied as it would be done using GSEA. But with FatiScan we can also consider many other gene properties or characteristics.

We can easily explore the correlation between gene expression and a clinical continuous variable such as the level of a metabolite. First, for each gene we will compute the correlation between its expression mea-

[1] It is a "subject sampling" method according to the definition of Goeman and Buhlmann 2007. Thence derives its lower flexibility when compared to FatiScan or the *logistic* methods.

surements and the levels of the metabolite. Thus we can range the genes from those which expression is more positively correlated to the levels of the metabolite to those inversely correlated, passing by genes which expression does not correlate with the clinical variable. In a second step, FatiScan explores the distribution of such correlation measurements, testing whether the distribution of correlations within a block of genes is different from the overall distribution of correlation in the dataset.

We can fit a Cox proportional hazard model to each gene in our data in order to study the relationship between gene expression and survival times. The estimates of the slope coefficients may be used as an index that ranks genes from those which increased expression is associated with long time survival to those which increased expression is associated to an early death. After computing this rank-index, FatiScan will find those blocks of genes for which the distribution of the slopes differs from the global distribution of the slopes.

Many other application examples can be found in Al-Shahrour et al. 2007a.

The complete separation of the two steps in FatiScan analysis is the key point which provides its flexibility to the method. Such flexibility makes possible to handle many different sources of information, not only microarray gene expression data: Any lists of genes ranked by any other experimental or theoretical criteria can be studied. Genes can be, for example, arranged by physico-chemical properties, mutability, structural parameters and so on, in order to understand whether there is some biological feature, characterized by the blocks of genes, which is related to the experimental parameter studied.

3.5 Avoiding the segmentation step

The GSA methodology FatiScan (Al-Shahrour et al., 2007a), evolved from the FatiGO tool (Al-Shahrour et al., 2004, 2007b).

FatiGO is an Over-Representation method that uses a Fisher's exact test for 2 by 2 contingency tables as underlying statistic. Once a cutoff is set, and relevant genes[1] are selected, the proportion of annotated genes in the selected list is compared to the proportion of annotated genes in the rest of the genome.

FatiScan extrapolated this paradigm by iterating the choice of the cutoff over the ranking statistic[2] provided by any gene level analysis; *scanning* it over a series of equidistant points, segments or partitions. This approach, called the segmentation test, was somehow arbitrary in the choice of the number of partitions as well as in their distribution over the ranking statistic. Such arbitrariness was criticized by some users of the method which was embedded in the Babelomics suit.

Form a statistical perspective, gene membership to a functional class, such as a GO or a KEGG, is modeled as binary variable. Being differentially expressed or not is also a binary variable, and that is why 2 by 2 contingency tables are suitable in the Over-Representation context.

But the statistical generalization of the Fisher's exact test when one of the two variables is not *binary* but *continuous*, comes from the *generalized linear models* theory. In particular, *logistic regression models*, are used to study dependencies between a *binary* variable and a *continuous* variable. See Agresti 2002 for a full description.

Hence, the natural way of extending the FatiGO methodology is not the FatiScan but the *logistic* model approach.

[1] Differentially expressed genes for instance if we where in a two class comparison context.

[2] See section 3.3 for clarification on the ranking index.

Chapter 4

Gene Set Internal Coherence

4.1 Montaner 2009 overview

In this chapter, the reader can find the second paper that makes up this thesis: **Montaner** et al. 2009.

At the moment of the publication it was widely accepted that the functional modules or *gene sets* described in public databases such as GO and KEGG, did not conform homogeneous classes of genes (see section 1.5 in page 32). This fact has always been expected because, unlike engineering, life[1] does not take uniform pieces to yield standardized solutions. Living beings' *parts* are rather custom made.

Some previous publications like Mateos et al. 2002 or Brown et al. 2000, had already explored real data in order to asses the homogeneity of *gene sets*. Nevertheless, such research efforts had always been conducted under particular experimental conditions and using a reduced amount of data.

The first point addressed in our publication was to thoroughly evaluate the general extent of such *gene set* inconsistency or "incoherence". We compiled the greatest possible dataset for the human transcriptome

[1] Or we would rather say evolution.

which could be collected at that time. More than 3000 microarrays[1] where downloaded from GEO, covering a wide range of diseased but also healthy biological conditions. Correlations among genes of this dataset where used to develop a *coherence index* reflecting the biologically measured homogeneity of each GO term and each KEGG pathway.

We surprisingly found a lower than expected homogeneity within *gene sets*. This fact questioned some of the biological hypotheses under which *gene set analysis* (GSA) studies are set up.

The second objective of our paper was to exploit the empirical information in our transcriptome dataset to enhance GSA methods performance. We achieved that by means of *logistic regression* models. Such statistical tools extended the methodological ideas used at the time, and allowed for the possibility of easily handle in the model the relevance of each gene within each *gene set*.

From our biological collection of data, we developed a *weighting* schema that measured the agreement of each gene expression profile with the general expression pattern of the *gene sets*. Then such *weights* where included in the *logistic* model to reflect the importance of each gene in the evaluation of each *gene set*.

4.2 Paper

A copy of the article, **Montaner** et al. 2009, follows next.

Supplementary materials provided with the paper may still be found at `http://bioinfo.cipf.es/data/coherenceindex`

The online version of the paper can be found at `http://www.biomedcentral.com/1471-2164/10/197`

[1] In our latest update of the collection we gathered 30000

BMC Genomics

Research article

Gene set internal coherence in the context of functional profiling

David Montaner[1,2], Pablo Minguez[1], Fátima Al-Shahrour[1] and Joaquín Dopazo*[1,2,3]

Address: [1]Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, E-46013, Spain, [2]Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe (CIPF), Valencia, E-46013, Spain and [3]CIBER de Enfermedades Raras (CIBERER), Valencia, E-46013 Spain

Email: David Montaner - dmontaner@cipf.es; Pablo Minguez - pminguez@cipf.es; Fátima Al-Shahrour - falshahrour@cipf.es; Joaquín Dopazo* - jdopazo@cipf.es

* Corresponding author

Published: 27 April 2009

BMC Genomics 2009, **10**:197 doi:10.1186/1471-2164-10-197

Received: 10 September 2008
Accepted: 27 April 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/197

Abstract

Background: Functional profiling methods have been extensively used in the context of high-throughput experiments and, in particular, in microarray data analysis. Such methods use available biological information to define different types of functional gene modules (e.g. gene ontology -GO-, KEGG pathways, etc.) whose representation in a pre-defined list of genes is further studied. In the most popular type of microarray experimental designs (e.g. up- or down-regulated genes, clusters of co-expressing genes, etc.) or in other genomic experiments (e.g. Chip-on-chip, epigenomics, etc.) these lists are composed by genes with a high degree of co-expression. Therefore, an implicit assumption in the application of functional profiling methods within this context is that the genes corresponding to the modules tested are effectively defining sets of co-expressing genes. Nevertheless not all the functional modules are biologically coherent entities in terms of co-expression, which will eventually hinder its detection with conventional methods of functional enrichment.

Results: Using a large collection of microarray data we have carried out a detailed survey of internal correlation in GO terms and KEGG pathways, providing a coherence index to be used for measuring functional module co-regulation. An unexpected low level of internal correlation was found among the modules studied. Only around 30% of the modules defined by GO terms and 57% of the modules defined by KEGG pathways display an internal correlation higher than the expected by chance.

This information on the internal correlation of the genes within the functional modules can be used in the context of a logistic regression model in a simple way to improve their detection in gene expression experiments.

Conclusion: For the first time, an exhaustive study on the internal co-expression of the most popular functional categories has been carried out. Interestingly, the real level of coexpression within many of them is lower than expected (or even inexistent), which will preclude its detection by means of most conventional functional profiling methods. If the gene-to-function correlation information is used in functional profiling methods, the results obtained improve the ones obtained by conventional enrichment methods.

Background

The popularisation of high-throughput technologies such as DNA microarrays has lead to a parallel demand of methods for data analysis. In particular, the necessity of providing a functional interpretation at molecular level that accounts for the macroscopic observations in high-throughput experiments has promoted the development of different methods for the functional profiling of this type of experiments during the last years [1,2].

It is widely accepted that genes do not operate alone within the cell, but they carry out their functions through a complex interplay whose most obvious experimental evidence is the intricate network of protein interactions that we only just have started to decipher [3,4]. Most of the biological functionality of the cell arises from complex interactions between their molecular components that define operational interacting entities or modules [5]. Functions collectively performed by such modules can conceptually be represented in different ways, being possibly Gene ontology (GO) [6] and KEGG pathways [7] the most popular and widely used ones. For practical purposes, functional modules are defined as sets of genes sharing GO or KEGG annotations. There are, obviously, many other categorizations of gene modules in different domains; for example Reactome pathways [8], Biocarta pathways http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways, etc.

In an attempt to understand the functional basis of high-throughput experimental results different functional profiling methods have been proposed [1]. Depending on the way the experimental data are selected and used two main families of methods, generically known as functional enrichment methods and gene set methods, can be distinguished. Functional enrichment methods have been implemented in several programmes such as GOMiner [9], FatiGO [10] and others. These are used to study whether a previously selected list of genes of interest is significantly enriched in one or more functional modules. Typical criteria for the selection of such gene lists in microarray experiments are differential expression between two classes, co-expression across experiments, etc. Thus, by means of this simple two-step approach, a reasonable biological interpretation of a microarray experiment can be achieved. Gene set methods were proposed more recently and directly aim to detect sets of functionally related genes (modules) with a coordinate and significant over- or under-expression across a list of ranked genes. Gene lists are ranked by differential expression between two classes, compared in microarray experiments [11-16]. In that way, the first step, in which genes are selected according to thresholds that ignore its cooperative behaviour, was avoided.

However, all these methods use functional modules as categorical variables. This fact, in the most typical microarray experimental designs (e.g. up- or down-regulated genes, clusters of co-expressing genes, etc.) or in other types of genomic experiments (e.g. Chip-on-chip, epigenomics, etc.), leads to the implicit assumption that such functional modules must be composed by sets of genes with a strong level of co-expression (otherwise they would never appear together in clustering or differential expression experiments or co-regulated by transcription factors, etc.). Nevertheless this assumption might not be necessarily true for all these modules. In fact, early attempts to deduce gene functionality (that is, functional module membership) from gene co-expression revealed that many functional modules did not even show a detectable degree of internal co-expression [17,18]. Therefore, if a non-negligible number of functional modules cannot be considered to be discrete categories there are two potential problems that affect all the methods for functional profiling: There is, on one hand, a problem of power in the statistical tests used, given that a number of functional modules tested will never be found simultaneously activated or deactivated, but are taken into account in the p-value adjustment procedures. On the other hand there is a potential problem of sensitivity, because many functional modules include genes with different degrees of intra-module co-expression, while the methods are many times applied to datasets in which the complete module is assumed to be over- or under-expressed as a whole. Since functional profiling methods do really produce results, one may conjecture that the results that are being obtained under the present unrealistic assumptions are only an underestimation of the results that could be really obtained if functional classes were properly tested.

Surprisingly, there are no systematic studies on the extent of this lack of internal co-expression within the most commonly used functional module definitions. The aim of this study was to produce a detailed survey on the GO and KEGG functional module definitions so as to determine which ones among them can be considered coherent modules of co-expression across a wide range of representative human samples. In principle, such coherent functional modules will define the subset of GO and KEGG functional modules susceptible of being detected using common strategies for functional profiling. In order to do so, we have derived, for each functional module as defined in GO and in KEGG, a co-expression (or coherence) index which could be used to assess the strength of its internal correlation. This index has been further used to filter for functional modules with a weak internal degree of co-expression. The index was derived from a gene pair-wise correlation matrix representing the overall correlation structure of the human transcriptome as estimated from microarray expression measurements of 3034 samples collected under the most diverse biological conditions. In addition, a second main aim of this work was to use this information to re-define the functional modules as non-discrete entities. Even in the case of the functional modules with a high degree of internal coherence, these cannot be considered as co-expression modules but rather as entities with a core of co-expressed genes along with a variable number of genes with lower correlation (that probably modulate, complement or provide alternative functionality). In other words, not all the genes need to be expressed at the same time for the function to be activated. Then, for each gene annotated within a functional module, we estimate its degree of correlation with the main bulk of genes annotated under such module. In this way we provide an index which is useful for quantifying how essential each gene is in the activation of the functional module. At the same time, we introduce a framework within which functional modules can be treated as non-discrete entities. Under the prism of this new vision of gene function, we propose a simple modification of the functional profiling methodologies in order to enhance the use of biologically relevant information as described in the functional modules. Finally, we present some examples about how these modifications can enhance the detection of biologically meaningful functions which would have remained unnoticed using currently available techniques for functional enrichment.

Results

Coherence index applied to functional modules defined by GO and KEGG annotations

A coherence index that gives an idea of the internal correlation of the genes belonging to a functional module has been proposed and estimated for all the GO terms and KEGG pathways. This coherence index may have several interpretations but certainly the most direct one is its understanding as the complement of a p-value. We firstly calculate the all-against-all correlation matrix for all the 10866 transcripts across the 3034 arrays used (see material an methods section), which is available as online supplementary material http://bioinfo.cipf.es/data/coherenceindex/. When the median correlation between the transcripts of a functional module is compared to the empirical distribution of correlations, estimated over randomly sampled sets of genes, we are assessing how strong the departure of our estimate from the null hypothesis of module correlation is. In other words, we can test if the internal correlation of such module is significantly higher than the correlation observed in a similar number of functionally unrelated genes. The coherence index proposed is the percentile represented by the module correlation within the random distribution. This index accounts for the complement of the probability of observing, under the null hypothesis, a value as extreme as the observed median. The cut-off of 0.05 usually chosen to reject a null hypothesis when the observed p-value is lower would be represented, in this case, by the level 95 of our coherence index. We would reject the null hypothesis for a functional module when its estimated index is higher than such value.

The application of the coherence index to the functional modules as defined by KEGG pathways showed that only 57% of them presented a correlation index greater than 95 (see Figure 1A). That is, if we were performing statistical analysis searching for KEGG pathways having internal correlation stronger than the overall correlation of the transcriptome, we would find no evidence of significant strong internal correlation in 43% of the cases. Thus 43% of the KEGG pathways do not co-express more than they would do if they were composed of functionally unrelated genes. Supplementary Dataset S1 contains the list of the KEGG pathways with their corresponding coherence index and median correlation values. Even more drastic are the results obtained for the GO terms. Only 32% (30% in Biological Process; 30% in Molecular Function; 46% in Cellular Component) of the functional modules defined by GO showed a correlation index greater than 95 (see Figure 1B). Supplementary Datasets S2, S3 and S4 contain the list of the GO terms corresponding to the "biological process", "molecular function" and "cellular component" ontologies respectively, along with their corresponding coherence indexes and median correlation values.

It is also worth pointing out that for many functional modules correlation indexes below 50 were observed. This means that for those modules the internal correlation is even lower than the overall genome correlation, which suggests the existence of a pattern of negative correlations among a significant amount of genes in the modules.

As expected, large functional modules (more than 100 transcripts) tend to have a strong internal correlation whereas small modules show more variability (see Figure 2). This was also observed for the three ontologies (Biological Process, Molecular Function and Cellular Component) of GO (see Additional file 1)

Not surprisingly it was found that, in general, when the internal median correlation of a functional module was low (correlation index below 50) its estimated standard deviation was high (see Additional file 2). More interesting is the finding that many functional modules with high internal correlation had also high standard deviations. This last observation makes it clear that, even within the functional modules which have a strong internal co-expression, there exist a non-negligible number of genes which do not co-express with the main bulk of genes of the module. We may conclude that, while a number of GO terms or KEGG pathways are defining true functional

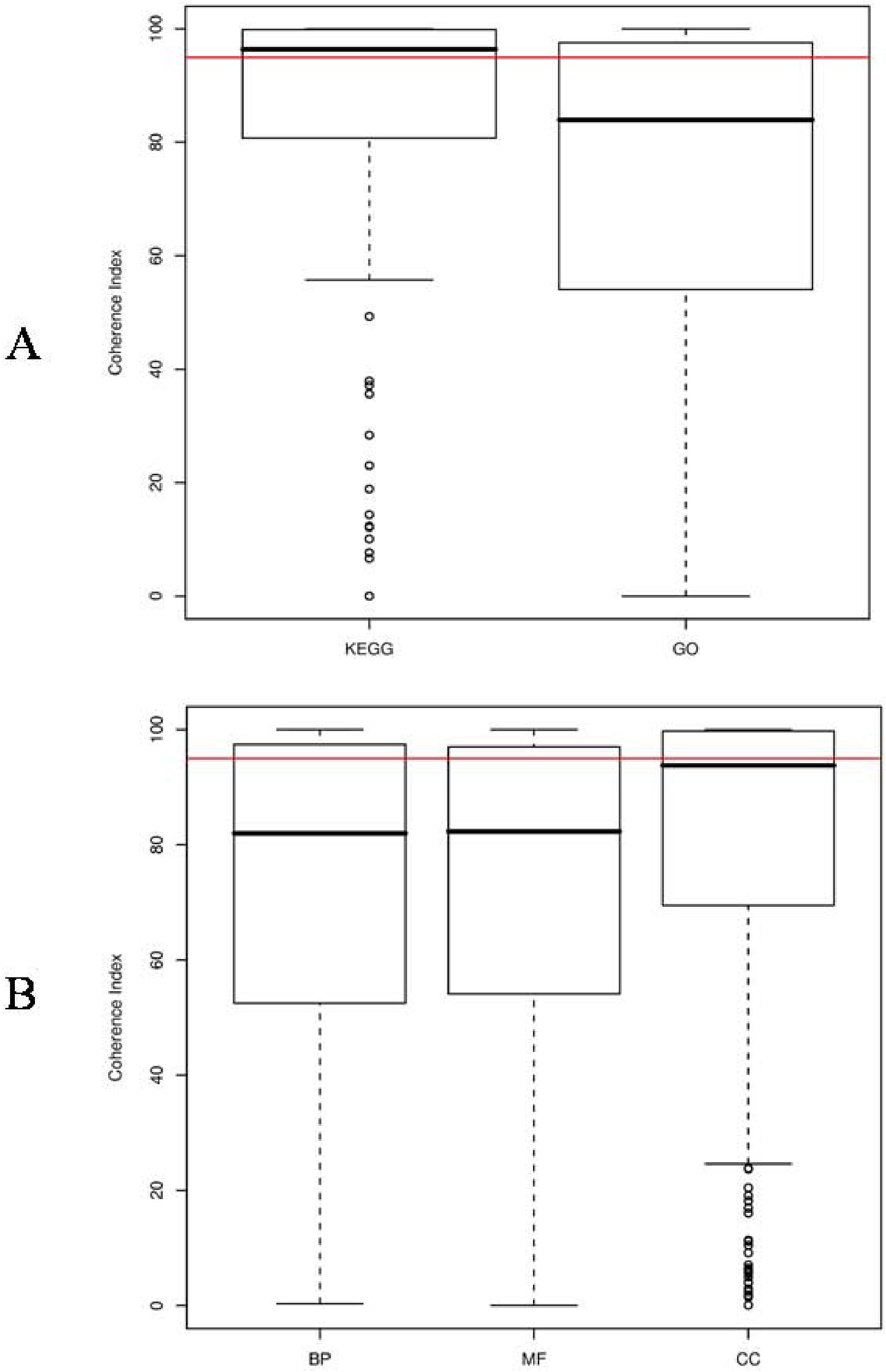


Figure 1
Distribution of coherence indexes. Coherence indexes for **A)** KEGG pathways and GO and **B)** the three GO Ontologies.

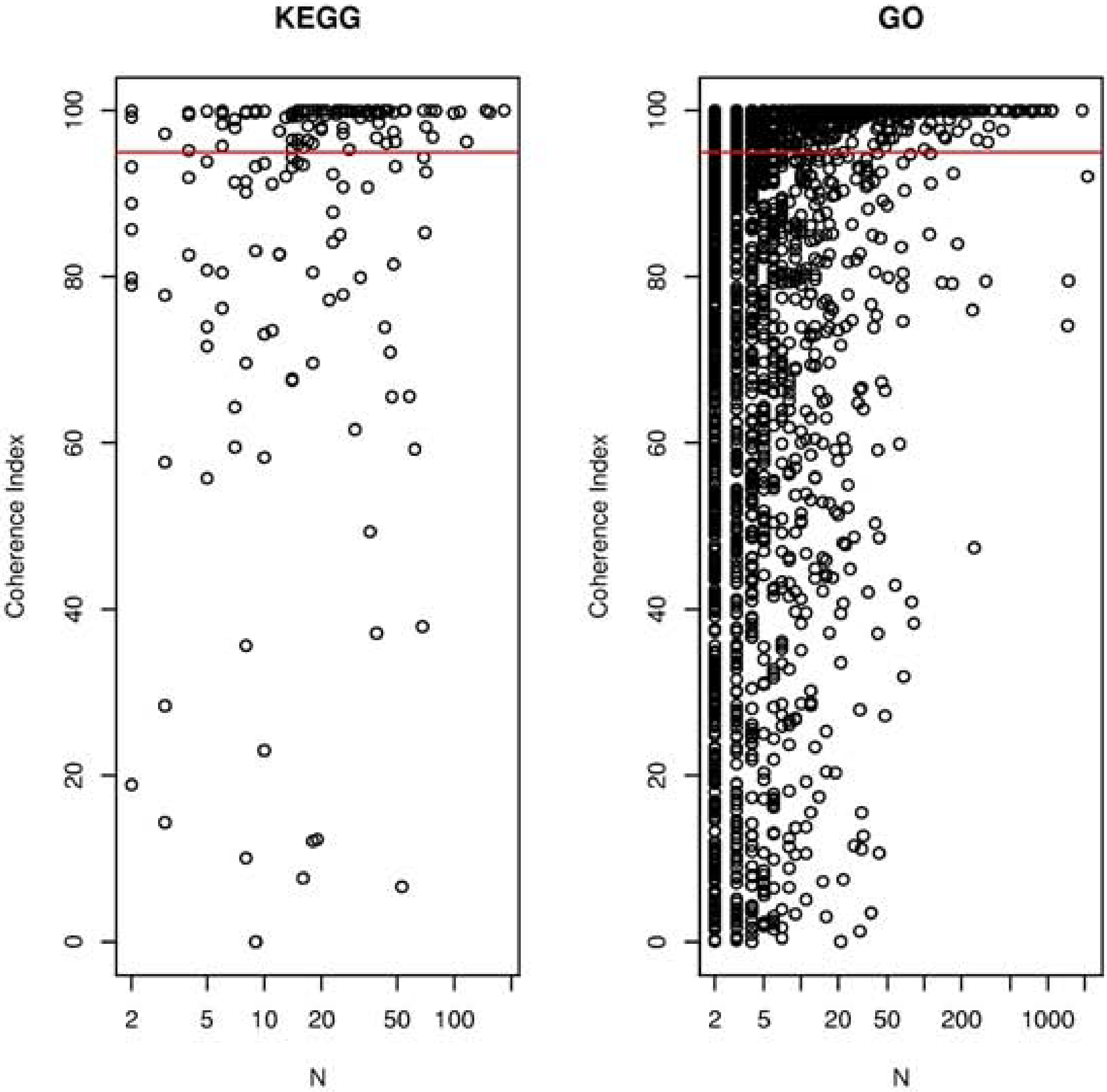


Figure 2
Coherence index values as a function of functional module size obtained for KEGG (left) and GO (right) categories.

modules of genes which need to be co-ordinately expressed in order to activate their corresponding functional roles, most of the currently used functional modules are not formed by sets of co-expressing genes.

Coherence index and the level of annotation in GO

In the particular case of GO, where functional terms are related to each other following a special type of hierarchical structure called directed acyclic graph (DAG) [6], we have studied the relationship between the proposed coherence index and the level of annotation of each term. Here, the level of annotation of a GO term is defined as the maximum number of nodes that can be found in the DAG between the term and the root of the corresponding ontology. Under this definition, high levels in the ontology represent more specific GO terms. Our findings show (see Figure 3) that there is not a direct relationship between the degree of internal correlation of a GO term

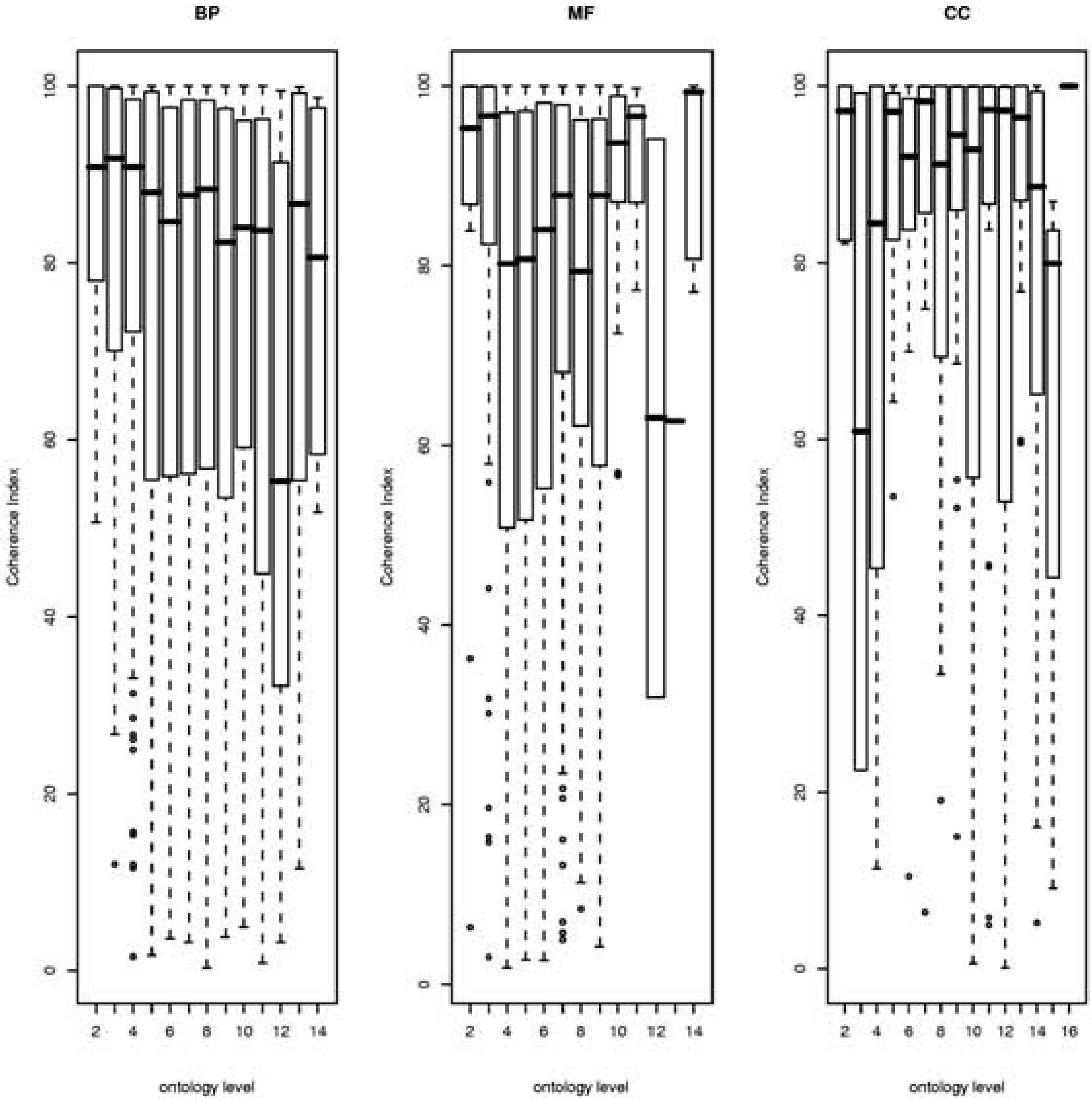


Figure 3
Coherence index as a function of the level (the deeper the more specific the functional definition) in the GO hierarchy obtained for the three ontologies: Biological process (left), molecular function (center) and cellular component (right).

and its level in the ontology hierarchy. It is interesting to remark that, contrarily to what it was expected, more specificity in a GO term does not imply a tighter co-expression. This is probably a reflection of the fact that many definitions in the ontology are not accounting for simple cooperative processes such as the ones carried out for example, by a complex of proteins.

Using gene-to-function information to best detect functional modules

In the following examples we show how to use this gene-to-function inter-dependence in order to incorporate the non-discrete nature of the membership of a gene to a functional module in the context of functional enrichment analysis.

Case example 1: functional profiling of genes differentially expressed in patients infected with Human Papillomavirus

A study of 36 Head and Neck Squamous Cell Carcinoma (HNSCC) tumour samples, 8 of them corresponding to patients infected with Human Papillomavirus (HPV+) and the remaining 28 to non infected patients (HPV-) [19] was used to illustrate the application of the proposed methodology. The authors assessed differential gene expression between HPV+ and HPV- tumours using Affymetrix 133 Plus 2.0 chips, and reported 89 genes over-expressed in the HPV+ group of tumours. A significant number of such genes were cell cycle regulators and transcription factors. The Affymetrix IDs for these 89 genes can be found in the supplementary material provided by the authors. The gene expression data are available in the GEO database [20] under the accession number GSE3292. In this case, raw files (.CEL) were not available and consequently these were not used in our estimation of the correlation between genes. Therefore, weights used in the analysis were obtained independently from the analyzed data set. The internal correlation value of any transcript to the rest of the transcripts in a functional module is used to assign a weight to it (see material and methods). Transcripts positively correlated to the rest of the module are given a weight of 2 (that is, are given double importance in the calculations), while negative correlations are penalised with a weight of 0.5 (half of the importance in this case). For the rest of genes a weight of 1 is used. A logistic regression model, which allows the use of weights, is utilised here instead the classical Fisher's test of equivalents (see material and methods).

We have systematically explored the GO and KEGG functional annotations of these 89 genes over-expressed in HPV+, testing for differences against the whole genome, that is, the remaining genes represented in the Affymetrix chip. A total 733 GO Biological Process terms and 161 KEGG pathways (with sizes comprised between 10 and 500 genes) were tested as described in the methods section.

A total of four GO terms were found as significantly over-represented in the group of over-expressed genes by the application of a standard, un-weighted test for functional enrichment with the permutation correction (see Table 1). In agreement with the discussion of the authors on the functionality of the genes differentially expressed [19], the terms related to DNA metabolism/replication (*DNA replication initiation*, p < 0.001, and *DNA strand elongation*, p < 0.001) were found. Also *SRP-dependent cotranslational protein targeting to membrane* (p < 0.001), probably accounting for the production of viral proteins is found. Finally, a term with no clear interpretation, *regulation of smooth muscle contraction*, was also found.

The application of the alternative weighted analysis proposed here detects a new term: *negative regulation of protein kinase activity* (p < 0.001), while *regulation of smooth muscle*

Table 1: Gene ontology functional terms and their respective significances under the standard (unweighted) and the weighted tests obtained for the HPV experiment [19] with the permutation test.

			Weighted			Unweighted		
GO name	BP	size	Log Odds	p-value	Adjusted p-value	Log Odds	p-value	Adjusted p-value
negative regulation of protein kinase activity	**GO:0006469**	100	3.030	<0.001	<0.001	2.548	0.006	0.083
DNA replication initiation	**GO:0006270**	44	4.281	<0.001	<0.001	4.162	<0.001	<0.001
SRP-dependent cotranslational protein targeting to membrane	**GO:0006614**	12	4.103	<0.001	<0.001	4.032	<0.001	<0.001
DNA strand elongation	**GO:0006271**	13	4.079	<0.001	<0.001	3.945	<0.001	<0.001
regulation of smooth muscle contraction	**GO:0006940**	18	2.875	0.010	0.088	3.597	<0.001	<0.001

contraction disappears. It is long known the relationship between MAP kinase and growth factor activity, two terms descendant of *negative regulation of protein kinase activity* and HPV infection [21] (see Table 1).

In the equivalent analysis of functional modules defined using KEGG, the pathway Heparan sulfate biosynthesis (that remained unnoticed in the unweighted test) was found to be significantly over-represented in the genes over-expressed in HPV+ by the weighted test significant. It has recently been reported that Human Papillomavirus infection requires cell surface heparan sulfate [22]. *Urea cycle and metabolism of amino groups* is significant in both the weighted and the unweighted analysis.

Case Example 2: functional differences between two types of cancers

A second example on a matched-pair analysis of 24 breast tumours to study the transition between *in situ* ductal carcinoma (DCIS) and invasive ductal carcinoma (IDC) [23] was analysed. In the study Affymetrix HG U133A and HG U133 Plus 2.0 chips were used to assess gene expression differences between these two conditions. The authors reported 445 Affymetix probe-sets up-regulated in IDC and 101 down-regulated in IDC. In their analysis authors also indicate cell-to-cell signalling and interaction as being the more significant functions of the differentially expressed genes. As in the previous example, Affymetrix IDs of the differentially expressed probe-sets where provided and gene expression data are available in the GEO database [20] under the accession number GSE3893.

We have tested for enrichment in GO and KEGG terms in the up-regulated genes and in the down-regulated genes. A total of 733 GO terms of Biological Process and 161 KEGG pathways of sizes comprised between 10 and 500 genes where included in this study.

Using a standard, un-weighted test for functional enrichment with the permutation correction two KEGG pathways:, *Focal adhesion* ($p < 0.001$) and *ECM-receptor interaction* ($p < 0.0001$), as well as two GO terms: *transmembrane receptor protein tyrosine kinase signaling pathway* ($p < 0.001$) and *regulation of cell shape* ($p < 0.001$), all of them related with the maintenance of cellular structures and cell motility, were found as differentially expressed. Again, the application of the alternative weighted analysis proposed here detects a new term: *proteoglycan metabolism* ($p < 0.001$). Proteoglycans are known to determine mitogenic responses of breast carcinoma cells to fibroblast growth factors, mediated by tyrosine kinase-signaling receptors [24].

Discussion

Functional annotations, such as GO or KEGG pathways, have been used for the definition of modules of genes in functional enrichment methods [1,9,10]. The detection of such functional modules within lists of genes by means of different tests relies upon the implicit assumption that common functionality implies a high degree of co-expression among all the members of each module [25]. While this assumption can be considered true as a general observation, it does not necessarily imply that the conventional definitions of functional classes used for this purpose (GO, KEGG, etc.) do all correspond to co-expressing sets of genes. It was previously reported that a large number of functional modules showed a low degree of internal co-expression, contradicting thus the expected cooperation among the genes to carry out their functions together [17,18]. Despite this observation, a systematic study on the degree of internal co-expression of the most commonly used functional modules and the impact of this bias on real biological data has not been carried out to date. Here we aimed a redefinition of functional modules, understood as groups of genes carrying out, cooperatively, a function in the cell. It is widely recognized that the biological circumstance of coexpression of two genes is properly defined by the coefficient of correlation among them [26]. So, we use it here to measure gene coordinate activity within a functional module. In this paper we present a general methodology to quantify the strength of the internal correlation of a functional module and we propose a simple way of using this information for functional profiling purposes that allows finding functional modules activated or deactivated that would remain otherwise unnoticed.

We have derived the correlation structure of the largest possible fraction of the human transcriptome, estimating its parameters from measurements from 3034 DNA microarrays stored in public data repositories. One of the strengths of the present study is, precisely, the big sample size (especially large if the difficulties in finding comparable microarrays in the databases are considered [27]) on which all estimations relay on. Of not less importance is the wide range of biological conditions considered in the study which includes several types of normal tissues, different kinds of cancer cells, male and female individuals as well as different cell lines. In order to ensure as much as possible the compatibility of the data gathered for the analysis, we have used one of the more extensively used expression arrays currently available (Affymetrix HG U133 Plus 2.0). For the same reasons, we have only collected datasets for which raw data were available so we could normalize and pre-process all of them together with the same method. This collection of samples constitutes a large dataset that allows us to perform a robust profiling of a large fraction of the human transcriptome, covering an ample spectrum of clinically and biologically relevant conditions.

The correlation structure of the transcriptome has been used to derive a coherence score which measures the internal co-expression of 173 KEGG pathways and 2221 GO terms. Our estimations indicate that only 57% of the KEGG pathways and just 32% of the GO terms can be considered to have internal correlation stronger than random modules of functionally unrelated genes of the same size. We also provided separate estimates for each of the Gene Ontologies (30% in BP; 30% in MF; 46% in CC), showing that, in general, GO Biological Processes or Molecular Function have a weaker internal correlation than KEGG pathways or GO Cellular Component. Another interesting finding was the fact that many modules have high internal correlation but also high variability.

Different reasons may account for these observations. In some cases there are functional modules defined in GO that are composed by independent or even antagonistic sub-modules and, consequently, their genes will never be found co-expressing in any experiment. Examples are transporters, which are composed by different independent types of sub-modules or any GO term starting by "regulation of", which usually has two antagonistic descendants called "positive regulation of" and "negative regulation of". In other cases, there are functional modules that require of a core of genes for properly carrying out the function and other genes of the module are only activated under particular physiological conditions, stresses, etc., displaying a lower degree of correlation. Modules composed by sub-modules can also exist, and many other situations can be imagined. In any case, the vision of a functional module as a discrete class, to which genes belong or do not belong to, is definitively not supported by the observations made. Thus, it is urgent to take a new approach that accounts for the non-discrete nature of the functional modules as defined by the most commonly used functional annotations (GO and KEGG).

In addition we highlighted how the level of annotation of a GO term in the ontology structure may not be the most suitable indicator, at least in terms of co-regulation, of the described function, despite being often used as a measure of its specificity.

Under the above mentioned considerations, most currently used functional profiling methods which model functional modules as groups of co-expressing genes, seem clearly inappropriate. The need of new methodologies for functional profiling and, above all, the essential requirement of a new notion of membership of a gene to a functional module is still an open issue. The proposed coherence score can be used in a first instance as a filtering criterion when the aim is to relate functionality to gene expression by discarding functional modules that will never be found as co-expressing units. Beyond this obvious use, this index can also be used to derive a weighting scheme that introduces the idea of non-discrete functional modules within the context of functional profiling methodologies in a straightforward manner. The proposed weighting scheme has the desirable property of using information on gene coexpression in the algorithm when such information is available but not introducing any bias when the information is missing. Relying on this new concept and using gene expression correlation information, we have shown with two examples how the proposed weighted approach discovers GO terms and pathways unnoticed under the equivalent standard un-weighted functional profiling method.

The approach shown here is quite general and could easily be extended to any other species or different platforms just by calculating the corresponding correlation matrix in a straightforward manner. The methodology could also be easily extended to any other types of modules defined by functionality, regulatory motifs, etc. Obviously, the use of newer strategies for functional profiling such as the different versions of gene set enrichment analysis [11-14,16], would benefit of considering this weighted definition of functional modules instead of using the classical categorical, un-weighted definitions.

Although the weighting schema proposed is quite simple, it proves efficient in finding functional modules in a standard functional enrichment analysis framework [1,10], as shown by the examples. Obviously, these examples have only an illustrative purpose of the application of the method that uses information on gene coexpression to improve functional module detection. However, in the worst scenario in absence of such information, this approach would be strictly equivalent to a conventional functional enrichment test and, therefore, its application would be equally valid. The use of most sophisticated weighting schemes, in which the continuous distribution of values of co-expression of all the genes in the module (and possibly outside the module) were taken into account, would probably improve even more the results. Also, a similar philosophy could also be applied to improve the detection of modules in gene-set enrichment methods although it falls beyond the scope of this manuscript.

Conclusion

The aim of the manuscript was, on one hand to show the discrepancy between functional modules as defined in some popular repositories (GO and KEGG pathways) and real co-expressing modules and, on the other hand, to propose a new vision of such modules that combines the original definition of the function with the actual dynamics of co-expression. In this more realistic scenario, functional modules with a coherence index that makes them undistinguishable from functionally unrelated gene modules would be excluded from a functional analysis, thus increasing the power of any test in the process of adjustment for multiple testing. In the remaining functional modules to be tested, more importance will be given to the core of co-expressing genes while uncorrelated genes and negatively correlated genes (probably representing genes that express under particular physiological conditions or stress situations, or perhaps other sub-modules with an independent dynamics of expression) will be penalised in the analysis.

Despite functional profiling of genome-scale experiments is an active field in which new proposals arise continuously [1,2], the concept of functional modules as binary discrete classes has remained unchanged along the last years. With the coherence index and the weighted schema proposed here we have introduced a conceptually new operative definition of functional module, biologically more meaningful, that clearly increases the sensitivity of functional profiling methods.

Methods

Expression values

All data used in this study was downloaded from the Gene Expression Omnibus (GEO), public repository of the NCBI [20]. At the time of doing this study, there were 169 GEO "series" containing microarray data generated using the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array (GPL570 platform in the GEO data base). Only for 74 of those series raw data (Affymetrix .CEL files) were available, comprising a total of 3034 array hybridized to all kind of human samples. We downloaded the raw data (.CEL files) for the 3034 arrays, normalized them in batches of size 100 (because of memory size limitations) using the function RMA in the *affy* library of Bioconductor [28] and finally rescaled all batches together using the "quantile" method implemented in the *limma* library of Bioconductor [29].

The data covered an ample spectrum of biological conditions including different tissues, and diseases, male and female individuals as well as cell lines.

ID mapping

Affymetrix probe-set identifiers were linked to their corresponding transcripts according to the Ensembl database, release 44 [30]. Among the 54675 probe-set IDs in the Affymetrix chip just 31542 had a corresponding Ensembl Transcript ID. Such IDs where unique just for 15477 Affymetrix IDs; that is, there are 16065 of the Affymetrix IDs that correspond to at least two different Ensembl Transcripts. A requirement of this study was to generate transcript expression measurements independent one of each other. Therefore we used just the 15477 Affymetrix IDs mapping to unique Ensembl IDs and, when several of them mapped to the same transcript, summarize them by its mean. In this way we manage to compute expression levels for 10866 transcripts, corresponding to 10486 genes of 3034 human samples.

Definition of functional modules using GO and KEGG annotations

GO and KEGG pathways annotation for the Affymetrix HG U133 Plus 2.0 array (the most abundant microarray in the databases) was taken from the Bioconductor metadata package "hgu133plus2" (version 1.14.0, see http://www.bioconductor.org/packages/devel/data/annotation/) which is assembled using data from public data repositories. 2221 GO terms (1014 Biological Process; 925 Molecular Function; 282 Cellular Component, Built: 8-Aug-2006) and 173 KEGG pathways (Release 38.1, June 1, 2006) that had annotated at least two of the 10866 selected transcripts where used in this study. While KEGG pathways are conceptually considered as independent entities, GO terms are related among them by a hierarchical relationship (known as directed acyclic graph, or DAG, in which a term can have more than one parent). Terms closer to the root define more general concepts and terms towards the leaves define more specific terms. In the particular case of GO terms, the usual procedure is to consider that each gene annotated to a given level is automatically annotated to all its parents [1]. All the GO terms have been used without making any distinction among distinct evidence codes. Since an overwhelming majority are electronic annotations (IEA), neither here, not in the most common programs for functional profiling [1] are taken into account. Functional modules are therefore defined as sets of genes sharing GO or KEGG annotations.

Computing correlations and assessing their strength

The main motivation in this work is the redefinition of the essence of a functional module, understood as a group of genes carrying out, cooperatively, a function in the cell. Typically, the coefficient of correlation [26], which accounts for the coexpression of genes across the experimental conditions measured, is used to measure such gene cooperation within a functional module. Figure 4 illustrates the way in which we proceed for computing the internal correlations for all the functional modules and estimating its significance. Thus, for all pairs of transcripts, the correlation of their expression levels along the 3034 arrays was computed and stored in a 10866 by 10866 correlation matrix. Distribution of this correlation coefficients within the functional modules considered in this study (GO terms and KEGG pathways) was studied and summarized by a median correlation value for each of the terms. For each functional module (GO term or KEGG pathway) consisting of N transcripts we randomly sam-

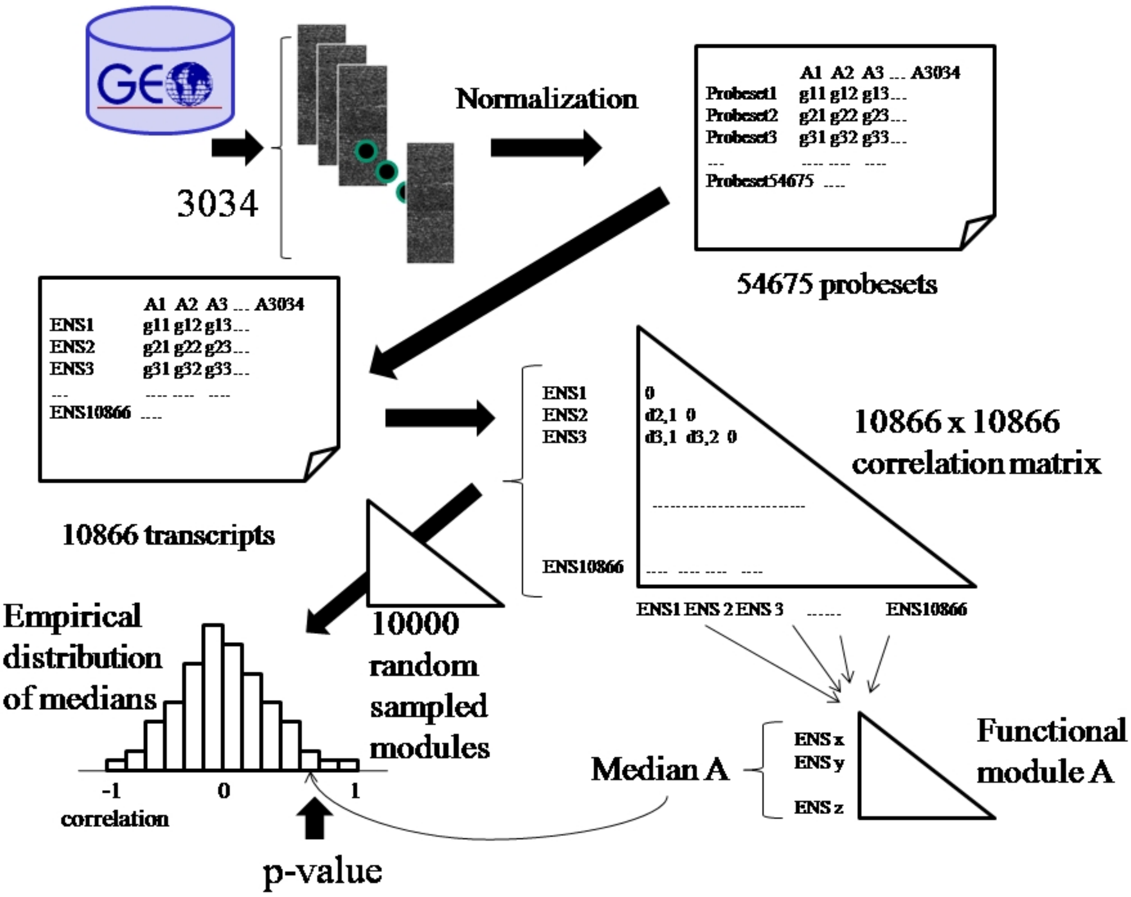


Figure 4

Schematic representation of the procedure followed for obtaining the internal correlation for each functional module and its significance. See material and methods.

pled, from the whole collection of transcripts in the study, 10000 modules of the same size N. Then, for each of the 10000 resampled modules, we computed the median value of the correlation between its transcripts. In this way we obtained a sampling distribution of the median correlation within equivalent modules of transcripts of size N not functionally related. In order to assess how strong the real internal median correlation of each functional module is, any of these values was compared to the sampling distribution of median correlations of random (functionally unrelated) modules of the same size. The percentile of the sampling distribution represented by the true median correlation in the functional module is, finally, taken as a measurement of the strength of its internal correlation and provided as coherence index.

The weighted approach: using co-expression information to improve functional profiling analysis

The most widely used tools for functional profiling classify genes into 2 by 2 contingency tables according to their functional annotation (functional module membership) and to the list to which they belong to. Then some statistical test, like a chi-square, Fisher or other equivalent test, is used to find statistically significant over-representations of any functional annotation in the lists of genes compared. Here we use logistic regression models [31] to estimate the log odds ratio of association between being or not annotated within a functional module and belonging to one or the other list of genes. When applied to binary data, this approach is equivalent to other 2 by 2 contingency table methods but has the advantage of allowing for the use of weighted observations. It has been shown that, when correlated genes are introduced in 2 by 2 contingency tables, standard tests inflate type I error rates [2]. In this paper we computed p-values based on the subject sampling model (1000 permutations) described by Goeman [2] in order to avoid such bias.

Here, we propose a very simple modification of the use of functional modules that can be applied within the context of functional enrichment analysis. The rationale for this modification is to give more importance to those genes that, being annotated in a functional module, are positively correlated to the main bulk of genes in the module. Likewise we seek to penalise the negative contribution to the detection of a functional module of those genes negatively correlated to this module. In order to achieve this, we have first to determine a measure of the internal correlation of genes within functional modules. Then, instead of using a discrete definition of functional modules, the correlations will be used to weight the membership of each gene to the module. When using the logistic model to test for each functional module, each gene was weighted depending on whether it was annotated or not within the module and whether it was positively or negatively correlated with it. Genes belonging to the functional module were given weight 2 if they were positively correlated to it and weighted by 0.5 if the correlation with the module was negative. The genes that were not in the functional module were given a neutral weight of 1. As in the classical functional enrichment test scenario, all computed p-values where corrected for multiple testing using the False Discovery Rate (FDR) method [32].

Abbreviations

DAG: Directed acyclic graph; **FDR**: False Discovery Rate; **GEO**: Gene Expression Omnibus; **GO**: Gene Ontology; **KEGG**: Kioto Encyclopaedia of Genes and Genomes.

Authors' contributions

DM is the author of the algorithm and has participated in all the analyses, FAS and PM has participated in the analysis of the data and JD has conceived and coordinated the study and written the manuscript. All the authors have read and approved the final manuscript.

Additional material

Additional File 1

Additional Figure 1. Coherence index values as a function of functional module size obtained for the three GO ontologies: Biological process (left), molecular function (center) and cellular component (right).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-197-S1.jpeg]

Additional File 2

Additional Figure 2. Relationship between the coherence index and its standard deviation for KEGG (up left), GO biological process (up right), molecular function (down left) and cellular component (down right).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-197-S2.jpeg]

Acknowledgements

This work is supported by a grant from the Spanish Ministry of Science and Innovation (BIO2008-04212 to JD). The CIBER de Enfermedades Raras is an initiative of the ISCIII. The RTICC is an initiative of the ISCIII. The National Institute of Bioinformatics http://www.inab.org is a platform of Genoma España.

References

1. Dopazo J: **Functional interpretation of microarray experiments.** *Omics* 2006, **10(3)**:398-410.
2. Goeman JJ, Buhlmann P: **Analyzing gene expression data in terms of gene sets: methodological issues.** *Bioinformatics* 2007, **23(8)**:980-987.
3. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, *et al.*: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature* 2005, **437(7062)**:1173-1178.
4. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, *et al.*: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122(6)**:957-968.
5. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402(6761 Suppl)**:C47-52.
6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25(1)**:25-29.
7. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Res* 2004:D277-280.
8. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, *et al.*: **Reactome: a knowledge base of biologic pathways and processes.** *Genome Biol* 2007, **8(3)**:R39.
9. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, *et al.*: **GoMiner: a resource for biological interpretation of genomic and proteomic data.** *Genome Biol* 2003, **4(4)**:R28.
10. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.** *Bioinformatics* 2004, **20(4)**:578-580.
11. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information.** *Bioinformatics* 2005, **21(13)**:2988-2993.
12. Goeman JJ, Geer SA van de, de Kort F, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome.** *Bioinformatics* 2004, **20(1)**:93-99.
13. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, *et al.*: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** *Nat Genet* 2003, **34(3)**:267-273.
14. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, *et al.*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102(43)**:15545-15550.
15. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies.** *Proc Natl Acad Sci USA* 2005, **102(38)**:13544-13549.

16. Kim SY, Volsky DJ: **PAGE: parametric analysis of gene set enrichment.** *BMC Bioinformatics* 2005, **6**:144.
17. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci USA* 2000, **97(1)**:262-267.
18. Mateos A, Dopazo J, Jansen R, Tu Y, Gerstein M, Stolovitzky G: **Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons.** *Genome Res* 2002, **12(11)**:1703-1715.
19. Slebos RJ, Yi Y, Ely K, Carter J, Evjen A, Zhang X, Shyr Y, Murphy BM, Cmelak AJ, Burkey BB, *et al.*: **Gene expression differences associated with human papillomavirus status in head and neck squamous cell carcinoma.** *Clin Cancer Res* 2006, **12(3 Pt 1)**:701-709.
20. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles–database and tools update.** *Nucleic Acids Res* 2007:D760-765.
21. Gu Z, Matlashewski G: **Effect of human papillomavirus type 16 oncogenes on MAP kinase activity.** *J Virol* 1995, **69(12)**:8051-8056.
22. Giroglou T, Florin L, Schafer F, Streeck RE, Sapp M: **Human papillomavirus infection requires cell surface heparan sulfate.** *J Virol* 2001, **75(3)**:1565-1570.
23. Schuetz CS, Bonin M, Clare SE, Nieselt K, Sotlar K, Walter M, Fehm T, Solomayer E, Riess O, Wallwiener D, *et al.*: **Progression-specific genes identified by expression profiling of matched ductal carcinomas in situ and invasive breast tumors, combining laser capture microdissection and oligonucleotide microarray analysis.** *Cancer Res* 2006, **66(10)**:5278-5286.
24. Mundhenke C, Meyer K, Drew S, Friedl A: **Heparan sulfate proteoglycans as regulators of fibroblast growth factor-2 receptor binding in breast carcinomas.** *Am J Pathol* 2002, **160(1)**:185-194.
25. van Noort V, Snel B, Huynen MA: **Predicting gene function by conserved co-expression.** *Trends Genet* 2003, **19(5)**:238-242.
26. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302(5643)**:249-255.
27. Larsson O, Sandberg R: **Lack of correct data format and comparability limits future integrative microarray research.** *Nat Biotechnol* 2006, **24(11)**:1322-1323.
28. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy–analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, **20(3)**:307-315.
29. Smyth G: **Limma: linear models for microarray data.** In *Bioinformatics and Computational Biology Solutions using R and Bioconductor* Edited by: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. New York: Springer; 2005:397-420.
30. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, *et al.*: **Ensembl 2007.** *Nucleic Acids Res* 2007:D610-617.
31. Agresti A: **An Introduction to Categorical Data Analysis.** New York: Wiley-Interscience; 1996.
32. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society Series B* 1995, **57**:289-300.

Chapter 5

Multidimensional Gen Set Analysis

5.1 Montaner 2010 overview

In this chapter we present the third article from those constituting this thesis: **Montaner** and Dopazo 2010.

In the paper we introduced the usage of *logistic regression* methodologies to perform a *multidimensional gene set analysis*.

As explained before (see section 1.6 in page 38). the rationale underlying conventional GSA methods is that of exploring the relationship between certain experimental characteristic measured for the genes of a genome and their membership to a *gene set* or *functional block*.

The experimental characteristic is reflected by a *ranking index*, that is, a numerical value quantifying certain biological property measured in the experiment. Such *ranking index* may be a p-value accounting for the differential expression of each gene in the experiment, a statistic computed to estimate SNP allele association to disease, or simply the copy number of each gene under our experimental conditions.

Up to the time we presented our idea, GSA methods could explore just one of such *ranking indexes* at a time. Therefore, just one experimental characteristic could be explored at a time in the light of, for instance, Gene Ontology terms.

The novelty of our approach was to introduce the possibility of including two (or more) *ranking indexes* in a combined analysis. Hence, our method enabled the researcher carrying out combined analysis of, for instance, differential expression and methylation, searching for *gene sets* enriched in this two dimensions of the experiment.

Nevertheless, the real advantage of the methodology we proposed was not to ease the computation of the analysis but, to allow for the exploration of the cooperative effect in modulating functional block behavior that had the two genomic characteristics under study when acting together.

5.2 Paper

A copy of the article **Montaner** and Dopazo 2010 is presented below.

Supplementary materials submitted besides the article can be found in Appendix B.

Among those supplementary materials, an R package implementing the algorithms was developed. The help manual of this package can be found in Appendix C and the source code may still be available at:
`http://bioinfo.cipf.es/supplementary/multidimensional_gsa`

Some of the methods from the R package where also included in the web tool Babelomics (Medina et al., 2010); they may still accessible at:
`http://www.babelomic.org`
The online version of the paper can be found at
`http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0010348`

PLoS one

Multidimensional Gene Set Analysis of Genomic Data

David Montaner[1,2], **Joaquín Dopazo**[1,2,3]*

1 Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain, 2 Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain, 3 CIBER de Enfermedades Raras (CIBERER), Valencia, Spain

Abstract

Understanding the functional implications of changes in gene expression, mutations, etc., is the aim of most genomic experiments. To achieve this, several functional profiling methods have been proposed. Such methods study the behaviour of different gene modules (e.g. gene ontology terms) in response to one particular variable (e.g. differential gene expression). In spite to the wealth of information provided by functional profiling methods, a common limitation to all of them is their inherent unidimensional nature. In order to overcome this restriction we present a multidimensional logistic model that allows studying the relationship of gene modules with different genome-scale measurements (e.g. differential expression, genotyping association, methylation, copy number alterations, heterozygosity, etc.) simultaneously. Moreover, the relationship of such functional modules with the interactions among the variables can also be studied, which produces novel results impossible to be derived from the conventional unidimensional functional profiling methods. We report sound results of gene sets associations that remained undetected by the conventional one-dimensional gene set analysis in several examples. Our findings demonstrate the potential of the proposed approach for the discovery of new cell functionalities with complex dependences on more than one variable.

Citation: Montaner D, Dopazo J (2010) Multidimensional Gene Set Analysis of Genomic Data. PLoS ONE 5(4): e10348. doi:10.1371/journal.pone.0010348

Editor: Jörg Hoheisel, Deutsches Krebsforschungszentrum, Germany

Received December 3, 2009; **Accepted** March 30, 2010; **Published** April 27, 2010

Funding: This work is supported by grants from project BIO BIO2008-04212 from the Spanish Ministry of Science and Innovation. The National Institute of Bioinformatics (www.inab.org), is a platform of Genoma Espana. The CIBER de Enfermedades Raras is an initiative of the ISCIII. The RTICC is an initiative of the ISCIII. This work is also partly supported by a grant (RD06/0020/1019) from Red Tematica de Investigacion Cooperativa en Cancer (RTICC), Instituto de Salud Carlos III (ISCIII), Spanish Ministry of Science and Innovation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jdopazo@cipf.es

Introduction

The development of new genomic technologies, such as microarrays of gene expression, genotyping or array-CGH, along with the new next-generation sequencing techniques is increasing the volume of data throughput amazingly. As a direct consequence of this, the bottleneck in functional genomics has shifted from the data production phase to the data analysis steps. In particular, the necessity for providing a functional interpretation at molecular level that accounts for the genome-scale experimental designs has promoted the development of different methods for the functional analysis of this type of data in the last years [1,2].

It is widely accepted that most of the biological functionality of the cell arises from complex interactions among their molecular components that define operational interacting entities or modules [3]. Functions collectively performed by such modules have conceptually been represented in different ways. Gene ontology (GO) [4] and KEGG pathways [5] are the most popular and widely used module definitions although many other are available in different repositories (e.g., Reactome [6], Biocarta, etc.) For practical purposes, functional modules are henceforth defined as sets of genes sharing functional annotations extracted from any of these repositories. Functional profiling methods exploit different definitions of modules in an attempt of understanding the functional basis of high-throughput experimental results [7]. Initially, functional enrichment methods, in different implementations [7,8], have been used for this purpose. More sensitive approaches, generically known as gene-set analysis (GSA) methods, pioneered by the Gene Set Enrichment Analysis (GSEA) [9], were later proposed [1,10]. In the original formulation, GSA methods aimed to directly detect sets of functionally related genes (modules) with a coordinate and significant over- or under-expression across the complete list of genes ranked according to their differential expression [9,11,12,13,14,15]. GSA methods can detect such modules even if their gene components are not significantly differentially expressed when tested individually. GSA has been successfully applied to the analysis of microarray experiments and has contributed to the adoption of a systems-biology perspective in distinct fields such as cancer [16]. Recent findings, brought about by the application of GSA methods on microarray experiments [17] are consistent with the idea that pathways, rather than individual genes, appear to govern the course of tumorigenesis [18]. The use of GSA has been extended to other areas beyond transcriptomics, such as evolution [19], QTL analysis [20] or genotyping [21].

Nevertheless, the different versions of GSA published to date [1,2,10] are inherently one-dimensional. Its application to the analysis of genomic datasets is at present limited to the study of a unique variable measured for the genes. The experimental conditions studied, even if corrected by other variables (e.g. age, gender, treatments, etc.), are typically summarized into a unique value for each gene (e.g., differential expression in a case-control, risk in the case of survival analysis, etc.) which is used to rank them accordingly.

Nowadays, the extensive use of different high-throughput methodologies allows the obtention of different measurements for the genes such as methylation status, splicing variants, linkage to diseases, etc., in a straightforward manner. As an illustration of this, a pilot study by The Cancer Genome Atlas (http://cancergenome.nih.gov/) consortium on glioblastomas has recently been published [22]. In it, different types of transcriptomic and genomic profiling were obtained and analyzed in an example of application of different genomic methodologies that would become routine soon. In addition, different measurements of the same type in different experimental contexts can easily be done. For instance, gene expression measurements in case-controls of different, but mechanistically related experimental conditions, phenotypes, diseases, treatments, etc. can be easily obtained. In such scenario, more than one measurement could be obtained to rank the genes involved in the study. Under the conventional GSA paradigm the different ranked lists of genes could be analyzed one at a time and still a good deal of information might be obtained. Nevertheless, by taking this approach any list of ranked genes is considered independent from each other and, consequently, behaviour of functional modules which are dependent on the combination of the studied ranking variables will, most likely, remain undetected.

Here we focus on a conceptually different strategy for GSA by extending the gene set based functional analysis to a multidimensional scenario in which more than one variable or genomic measurement is available for all genes in the study. Logistic regression allows for fitting models that include more than one variable. We show here, by means of several examples, how the application of the multidimensional GSA (MD-GSA) uncovers biological processes activated by different combinations of parameters (measured for all the genes and derived from microarray of other experiments) that would have remained undetected if the parameters would have been analysed one at a time, independently.

Results

Gene-set activation dependent on the transcription rates and mRNA activities in yeast

Gene expression is a process that involves two steps of synthesis which end when the appropriate level of protein required for performing a given function is reached. Some processes in the cell require of a quick activation and/or deactivation, while others remain in activity for longer periods and their activation processes do not involve any urgency. Thus, it is expectable different cell functionalities will use different strategies of gene and protein expression and degradation. Measurements of these parameters can be found in a recent genome-wide analysis on common gene expression strategies in yeast [23]. Using these data, we have studied two relevant and opposite biological processes that account for the steady-state mRNA level in the cell: transcription and stability [24]. The authors used a functional enrichment strategy [25] to test the GO terms associated to the parameters measured and to their correlations. Essentially, they used quintiles as cut-off values and tested for enrichments in the genes showing a high or low correlation in rates (transcription and translation) or abundances (mRNA and protein copy number), finding a total of 22 GO terms significantly over-represented at different combinations of rates and abundances. Nevertheless, other interesting situations in which the measurements are not correlated (e.g. transcription rate and mRNA stability) could not be analysed with this approach that, in addition, has the disadvantage of requiring an arbitrary threshold.

Here we analysed the dependences of GO terms on two measurements, transcription rate (TR) and mRNA stability (RS), as well as on the interaction between them. When the logistic

Table 1. Significant GO terms when transcription rate and mRNA stability are taken into account in the model.

	Log odds ratio (model coefficients)			Adjusted p-value					
GO id	**TR**	**RS**	**inter**	**TR**	**RS**	**inter**	**pattern**	**new**	**GO name**
GO:0019953	−11.87	−0.82	3.29	0.04	0.01	0.02	q3i	yes	sexual reproduction
GO:0051704	−11.98	−0.69	3.23	0.04	0.02	0.02	q3i	yes	multi-organism process
GO:0000819	−30.49	−0.87	7.1	0.02	0.03	0.02	q3i	yes	sister chromatid segregation
GO:0006260	−20.35	−0.97	4.99	0	0	0.01	q3i	no	DNA replication
GO:0006261	−25.15	−1.31	6.28	0	0	0.01	q3i	no	DNA-dependent DNA replication
GO:0022613	−4.69	−1.78	1.61	0.08	0	0.03	q3i	no	ribonucleoprotein complex biogenesis and assembly
GO:0042254	−5.05	−1.91	1.75	0.09	0	0.03	q3i	no	ribosome biogenesis
GO:0000746	−11.48	−0.73	3.17	0.06	0.02	0.03	q3i	yes	conjugation
GO:0000747	−11.39	−0.74	3.16	0.06	0.02	0.03	q3i	yes	conjugation with cellular fusion
GO:0042221	−6.65	−0.12	2.05	0.02	0.6	0.01	q3i	yes	response to chemical stimulus
GO:0000070	−30.23	−0.78	7.01	0.03	0.07	0.03	q3i	yes	mitotic sister chromatid segregation
GO:0019725	−9.13	−0.38	2.71	0.02	0.15	0.01	q3i	yes	cellular homeostasis
GO:0042592	−8.75	−0.3	2.59	0.02	0.27	0.01	q3i	yes	homeostatic process
GO:0006325	8.01	−0.47	−3.09	0	0.03	0.01	q4i	no	establishment and/or maintenance of chromatin architecture
GO:0065004	12.12	−0.49	−4.6	0	0.21	0.02	q4i	no	protein-DNA complex assembly
GO:0006323	12.63	−0.48	−4.96	0	0.15	0.01	q4i	no	DNA packaging
GO:0006333	12.44	−0.4	−4.84	0	0.23	0.01	q4i	no	chromatin assembly or disassembly
GO:0031497	12.51	−0.44	−4.84	0	0.2	0.01	q4i	no	chromatin assembly

A total of 18 GO terms were found as significant at FDR-adjusted $p<0.05$, nine of them were also found by the multivariate analysis. Column new indicates if the term as been found only because of the interaction factor (yes) or if it was found also in the univariate analysis in one or both dimensions independently.
doi:10.1371/journal.pone.0010348.t001

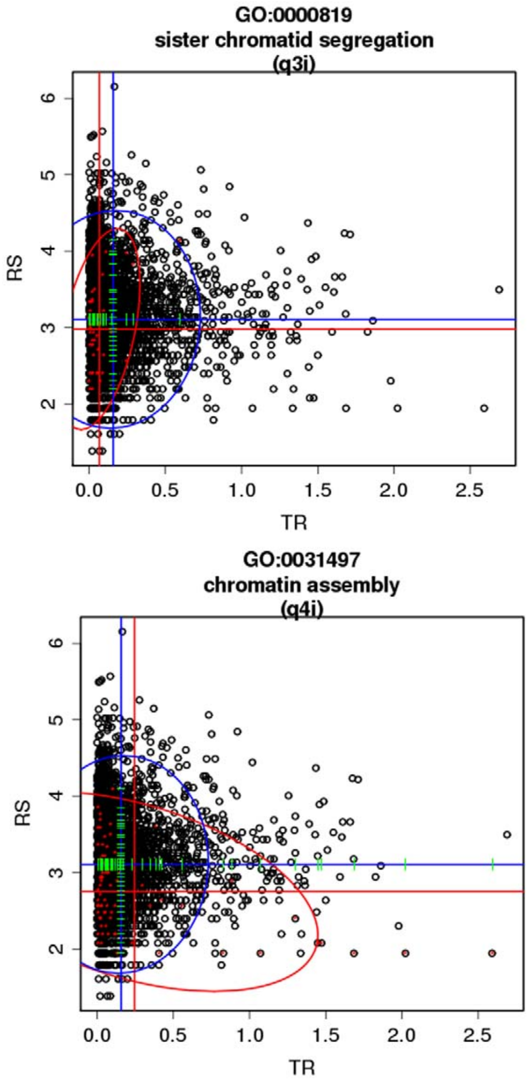


Figure 1. Combined analysis of transcription rates and mRNA stability in yeast with the logistic model. RS (mRNA stability) is represented in vertical axis and TR (transcription rate) is represented in the horizontal axis for GO terms sister chromatid segregation (top) and chromatin assembly (bottom). Blue lines intersect in the mean of the distribution of all the values and red lines intersect in the mean of the distribution of values of the genes corresponding to the GO term analysed. Blue ellipse delimits the confidence interval for all the values and red ellipse delimits the confidence interval for the GO term analysed. The red ellipse marks the trend of the relationship between both variables. MD-GSA assigns patterns q3i and q4i respectively to these functional modules.
doi:10.1371/journal.pone.0010348.g001

model was applied to the mRNA stability and to the transcription rate independently, we obtained 170 and 80 GO terms significantly associated to extreme values of these variables (see Table S1). This increase in the number of GO terms found was due to the well known fact that GSA strategies are much more sensitive than threshold-based functional enrichment strategies [1,10]. Actually, similar results were obtained when other equivalent GSA strategies were used (data not shown) [11,19].

Nevertheless, the most interesting aspect of this study is the analysis of the interaction between both variables. Table 1 shows 18 GO terms which were significantly associated to the interaction between transcription rate and mRNA stability. Figure S1 depicts the GO terms within the GO hierarchy. Nine of these GO terms could only be detected when the model takes into account simultaneously both parameters. In most of the cases, the GO was associated to both low transcription rate and mRNA stability (pattern *q3i*, see methods for an explanation of the patterns) such as *sister chromatid segregation* (Figure 1 top) in a subtle way that can only be detected when both parameters are included in the model. On the other hand, other processes, such as *DNA packaging*, *Chromatin assembly* (Figure 1 bottom), *Chromatin assembly or disassembly* and *Establishment and/or maintenance of chromatin architecture* (which are related terms, see File S1), or *protein-DNA complex assembly* are associated to high transcription rates but low mRNA stability (pattern *q4i,* seemethods). This last strategy, opposite to the first one, suggest a transient necessity of these processes, whose genes are produced at a fast rate but quickly discarded after their functions have been carried out.

Different strategies of production and degradation, corresponding to different biological requirements of the cell, can be thus detected by the combined analysis of these parameters.

Gene-set dependences on differential expression and splicing index. Recent studies have shown that more that 70% of the multi-exon genes, corresponding to about 50% of all human genes, are predicted to be alternatively spliced [26]. It is well known that alternative splicing participates in many pathways and processes. Also alterations in splicing function has been implicated in many diseases, including neuropathological conditions such as Alzheimer disease, cystic fibrosis, defects in growth and development, and many human cancers [27].

The magnitude of the alterations in the splicing process can be studied through the splicing index. This index accounts for changes at the exon level that are relative to the expression of the gene. In particular, the intensity value of an exon's probeset is divided by an estimate of the expression level of the transcript cluster to which the exon belongs to. In this way, a gene-level-normalized intensity that can be compared across samples or conditions is created. Changes in this value between case and control samples provide a quantitative measure of alternative splicing between the two conditions [28]. Thus each gene in the data set can be studied both in terms of its differential expression and its alternative splicing. Our multidimensional logistic model can be used to explore this two dimensional gene space.

Here we reanalyze data obtained using Affymetrix exon arrays [29] in which human breast cancer cell lines are compared to non tumorigenic human breast epithelial cell lines. The parameters studied by means of the multidimensional logistic model are: differential gene expression estimates obtained upon the application of a t-test for the above mentioned comparison and a splicing index, that accounts for changes at the exon level that are relative to the expression of the gene [30].

A total of 141 GO terms were found to be significantly associated to high values of the differential gene expression dimension (pattern yh, yl; see methods section). These terms are equivalent to those that would be found by conventional one-dimensional GSA methods and, as expected, GO definitions related to cell proliferation, cell signalling, apoptosis, cellular adhesion, etc., were found among them. One significant GO term, *regulation of viral reproduction*, was significant in the splicing index dimension alone. The trend of the enrichment was towards the positive values of the splicing index (pattern *xh*; see methods section) meaning that genes in the GO term are "subordinately" more spliced in the tumour than in the normal tissue (see File S2A).

Another 12 terms were found by the MD-GSA (see Table 2), whose relationships within the GO hierarchy is depicted in Figure

S2. The processes discovered here were related (but yet undetected) to other processes already detected by the conventional analysis of differential expression (see File S2A). For example, *positive regulation of cell adhesion* and its parent *regulation of cell adhesion* are descendants of *cell adhesion*, and two sister processes (*cell-matrix adhesion* and *cell-substrate adhesion*) were found by the model when the two variables were taken into account, and would have remained undetected if a conventional, unidimensional GSA approach would have been used. The patterns for these terms are bimodal in the two dimensional space (pattern *b24*, see methods section) indicating that the genes annotated to them behave as if they were in two sub-modules. For example, *positive regulation of cell adhesion* and its parent processes *regulation of cell adhesion*, which are known to be related to cancer, show a bimodal pattern towards the quadrants 2 and 4 (pattern *b24*). This means that part of the annotated genes are more spliced but underexpressed in the tumour samples while the other part is more spliced but underexpressed in the control samples (see Figure 2).

An equivalent analysis for KEGG can be found in File S2B.

Gene-sets differentially activated in related diseases: a case study with psoriasis and dermatitis. The study of gene expression at genomic level in both psoriasis [31] and dermatitis [32] and further functional analysis reveals a considerable number of deregulated pathways when both diseases are compared to their corresponding healthy samples. Thus, when the multivariate logistic model was applied to gene lists arranged by differential expression 172 GO terms were found to be significant only for dermatitis (patterns *xh*, *xl*; see methods section) and 202 only for psoriasis (patterns *yh, yl*). Another 77 GO terms were found to be significant in both, dermatitis and psoriasis but did not show an interaction effect (patterns *q1f, q2f, q3f, q4f*) Most of this terms will also be found by the independent unidimensional analysis of the dermatitis dataset and the psoriasis dataset. In the case of dermatitis, terms related to signalling, cell proliferation, immune system and development of epidermis were found, among others (see Files S3A and S3B). Similar terms can be found in psoriasis with some variations (see Files S3A and S3B). A detailed comparative functional analysis of these diseases is beyond the scope of this manuscript and we will only focus on the results obtained when both diseases are simultaneously analysed.

Table 3 shows the GO terms that are significant when both diseases are taken into account in the logistic model (column labelled with "inter"). Figure S3 shows the GO terms within the GO hierarchy. The GO terms *M phase of mitotic cell cycle* (and their parent terms *M phase* and *cell cycle phase*) and *cell division* where associated to both diseases in their main effects and also in their interaction effect (pattern *q1i*, seemethods) reinforcing their relevance in the biological mechanisms underlying both skin syndromes. Some other GO terms are only significant in the interaction effect. Their genes show a bimodal behaviour as if the functional module was composed of two sub-units (pattern *b13*, *b24*; see methods). For instance, GO terms *phosphoinositide-mediated signaling* and *response to reactive oxygen species* have a positive interaction coefficient, which means that some of the genes of the module are being coordinately over-expressed in both diseases while the remaining genes in the GO term are under-expressed also in both diseases. In a symmetric way, *negative regulation of lymphocyte proliferation* (and the parent process *negative regulation of mononuclear cell proliferation*) shows a negative interaction. Part of the genes in these modules increase their expression in dermatitis but decrease it in psoriasis while the rest of them present the opposite behaviour. The reduced cutaneous IFNalpha2 transcription which has been described as a differential characteristic of dermatitis with respect to psoriasis [32] could be causing this effect detectable in the analysis when the two variables are included in the model. All this bimodal terms highlight antagonistic effect, detectable only trough the combined analysis of both diseases.

Combined analysis of several genomic measurements: a case study with genotyping, gene expression and copy number alterations in breast cancer

It is known that mutations or alteration in copy number are related to cancer and tumour development [33,34]. Current microarray technologies allow for the measurement of SNP variation and copy number estimation at the same time [35,36] and have been used to gain insights into breast cancer [37,38,39], among other diseases.

Table 2. Significant GO terms when differential expression and splicing index are taken into account in the model.

	Log odds ratio (model coefficients)			Adjusted p-value				
GO id	**splicing**	**diff.exp**	**inter**	**splicing**	**diff.exp**	**inter**	**pattern**	**GO name**
GO:0006767	0.15	−0.15	0.14	1	0.61	0.04	b13	water-soluble vitamin metabolic process
GO:0045216	0.29	−0.04	0.17	1	0.95	0.02	b13	cell-cell junction assembly and maintenance
GO:0007043	0.38	−0.03	0.18	1	0.97	0.02	b13	cell-cell junction assembly
GO:0048706	0.2	0.08	0.17	1	0.89	0.03	b13	embryonic skeletal development
GO:0007034	0.32	−0.18	0.17	1	0.65	0.02	b13	vacuolar transport
GO:0007041	0.32	−0.1	0.18	1	0.86	0.01	b13	lysosomal transport
GO:0048704	0.23	0.12	0.19	1	0.84	0.02	b13	embryonic skeletal morphogenesis
GO:0048705	0.17	0.1	0.17	1	0.85	0.02	b13	skeletal morphogenesis
GO:0016197	0.08	0.1	0.15	1	0.79	0.02	b13	endosome transport
GO:0030155	0.01	−0.16	−0.15	1	0.43	0.01	b24	regulation of cell adhesion
GO:0045785	−0.04	0.06	−0.18	1	0.94	0.02	b24	positive regulation of cell adhesion
GO:0030032	−0.16	−0.17	−0.18	1	0.72	0.03	b24	lamellipodium biogenesis

A total of 12 GO terms were found as significant in the interaction at FDR-adjusted $p<0.05$.
doi:10.1371/journal.pone.0010348.t002

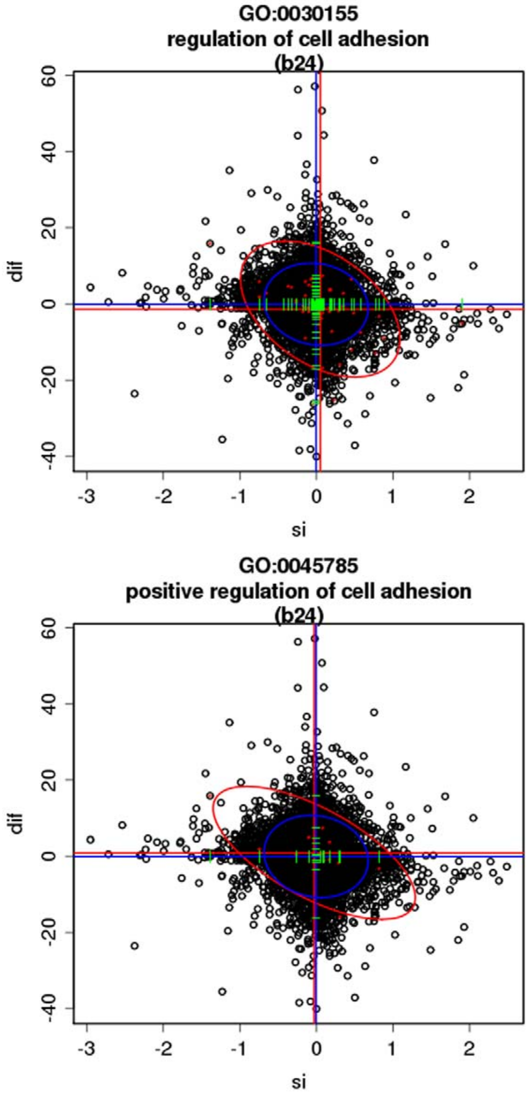


Figure 2. Combined analysis of differential gene expression and splicing index with the logistic model. Differential expression is represented in vertical axis and splicing index is represented in the horizontal axis for GO terms positive regulation of cell adhesion (bottom) and its parent processes regulation of cell adhesion (top). Blue lines intersect in the mean of the distribution of all the values and red lines intersect in the mean of the distribution of values of the genes corresponding to the GO term. Blue ellipse delimits the confidence interval for all the values and red ellipse delimits the confidence interval for the GO term analysed. The red ellipse marks the trend of the relationship between both variables. MD-GSA assigns a bimodal pattern b24 to these functional modules.
doi:10.1371/journal.pone.0010348.g002

Using the multidimensional logistic model proposed we have re-analyzed here data from several separated studies previously collected by us in an integrative analysis of breast cancer disease [38]. In particular we provide a combined description of GO and KEGG relationship to different parameters such as SNP association, copy number alteration and differential gene expression in connection to disease outcome (all the data were taken from the additional information of the above mentioned study, see methods).

When analyzing SNP association data and copy number in luminal B tumours by the proposed MD-GSA, *basal cell carcinoma* KEGG pathway raised up (File S4B) showing a bimodal pattern towards quadrants 1 and 3 (*b13*, see methods). This indicates that the genes in the pathway highly associated to disease are also increased in their copy number, and that genes not associated to disease do not have an increased copy number (they may even have a reduced copy number what would fit with the no association or even protection of the SNPs to disease). Most probably, the SNPs are markers associated either to regions undergoing copy number alterations or to other mutations that affect the *basal cell carcinoma* pathway, which obviously underlies breast cancer disease. The same analysis using the GO reported some negative bimodal terms (Table 4 and File S4B) like *L-amino acid transport* which is known to be involved proliferation processes [40]. A similar analysis with GO terms can be found in File S4A. Figure S4 displays the GO terms in Table 4 within the GO hierarchy.

We also applied the MD-GSA to the variables prognosis and differential expression in tumours. In the representation (File S5A), high values in the differential expression dimension indicate under-expression in tumour while low values indicate over-expression. Conversely, high values in the prognosis dimension indicate bad prognosis (if the gene is expressed) while low values in the prognosis dimension indicate good prognosis (if the gene is expressed).

Table 5 (more details in File S5A) show results obtained from the application of the MD-GSA using modules defined with GO terms. The relationships among them within the GO hierarchy are depicted in Figure S5. Most of the GO terms related to *cell division* and *cell cycle* show a *q2i* pattern (see methods) indicating a significant convergence of their genes in the prognosis and differential expression dimensions. From the relatively high prognosis value associated to the genes annotated to this GO terms we know that, if over expressed they indicate bad prognosis. From the low values in the t-statistic we know these GO terms are enriched in the tumours samples. Hence the multivariate logistic model is pointing out those modules which are dangerous to the patient if they are activated, and, that are certainly know to be activated in luminal B tumours. This extended functional analysis provides the researcher not only with a quick an easy interpretation of the combined data but also with the additional information of the interaction term in the model. It is worth pointing out here that better and more detailed results are obtained by combining both datasets under the proposed methodology than by applying independently the univariant methodology to any of the datasets and summing up the results obtained. The equivalent MD-GSA for KEGG pathways can be found in File S5B.

Advantages and limitations of the logistic regression methodology

The major advantage of the logistic regression methodology is it flexibility. It can be used in any genomic context in which certain biological characteristic of a gene is measured using a numerical scale. This numerical scale may be a continuous "ranking statistic" as described previously [41] or in this paper, but it may also be a categorical variable [42].

Moreover, many modifications of the logistic model with potential applications in biology are already statistically developed and can be used straight forward. Here, for instance we showed how to extend the methodology to study not one but two gene characteristics at a time. It is also straightforward to include the interaction in the model as we showed here. A unidimensional binary logistic model can be used instead the conventional 2×2 contingency table alternative because the logistic model easily

Table 3. Significant GO terms when differential expression of dermatitis and psoriasis are taken into account in the model.

	Log odds ratio (model coefficients)			Adjusted p-value				
GO id	**dermatitis**	**psoriasis**	**inter**	**dermatitis**	**psoriasis**	**inter**	**pattern**	**GO name**
GO:0022403	−0.13	0.36	0.11	0.11	0	0.01	q1i	cell cycle phase
GO:0000279	−0.06	0.37	0.12	0.55	0	0.03	q1i	M phase
GO:0051301	−0.1	0.25	0.15	0.36	0	0	q1i	cell division
GO:0000087	−0.11	0.4	0.12	0.32	0	0.05	q1i	M phase of mitotic cell cycle
GO:0048015	0.08	0.07	0.16	0.72	0.68	0.05	b13	phosphoinositide-mediated signaling
GO:0000302	0.24	−0.06	0.29	0.59	0.85	0	b13	response to reactive oxygen species
GO:0032945	0.43	0.33	−0.79	0.26	0.39	0	b24	negative regulation of mononuclear cell proliferation
GO:0050672	0.43	0.33	−0.79	0.26	0.39	0	b24	negative regulation of lymphocyte proliferation
GO:0048589	−0.19	−0.06	−0.59	0.53	0.91	0.04	b24	developmental growth
GO:0007028	0.21	−0.11	−0.75	0.47	0.83	0	b24	cytoplasm organization and biogenesis
GO:0007043	0.07	−0.5	−0.91	0.86	0.22	0	b24	cell-cell junction assembly
GO:0045216	0.12	−0.26	−0.86	0.75	0.59	0	b24	cell-cell junction assembly and maintenance

A total of 12 GO terms were found as significant in the interaction at FDR-adjusted $p<0.05$.
doi:10.1371/journal.pone.0010348.t003

allows for weighting genes [42]. This simplicity of extension is not at all intrinsic to most other GSA approaches, what makes the logistic model worth to be explored.

Another advantage of the method is that it does not start from the original observed data set (gene expression matrix for instance) but from a ranking statistic that already summarizes the relevant characteristic under study. This makes the methodology useful in many genomic contexts beyond the microarray paradigm. One example of ranking statistic we have discussed is the classical t-test which, perhaps with some modification, is underneath most GSA methodologies. For each gene, this statistic measures the biological characteristic of "how much" the gene is differentially expressed in a particular biological experiment. But we also exemplified how the ranking statistic can be a hazard ratio form a Cox model or other gene-wise variable[19]. In the case of the hazard ratio, the biological characteristic measured for each gene by the statistic is the association of expression and risk disease. The GSA for this second example can be directly carried out using the logistic methodology and software. On the contrary, most GSA approaches will require major modifications of their methods and software to be applied in a case other than differential gene expression in a class comparison experiment.

Virtually any gene-wise variable can be explored from a GSA perspective using the logistic regression model. In this paper we presented examples for the analysis of transcription rates, mRNA stabilities, splicing, SNP association to disease and copy number estimation. The analysis of other measurements is possible, including the evolutionary selective pressure in the human genome or a study of gene connectivity in the interactome [19]. Other publications also discuss on the advantage of a methodology that starts form a single ranking statistic and not from the whole expression data matrix [42,43].

Having said that, some remarks and warnings should be given related mainly with the null hypothesis that underpin the method and p-value computation.

In Sator's logistic regression approach [41] and in the extension we are proposing here, the distribution of the ranking statistic within each module is compared to that of its complement. Thus,

Table 4. Significant GO terms when copy number and gene association to the disease (see text) are taken into account in the model.

	Log odds ratio (model coefficients)			Adjusted p-value				
GO id	**association**	**copy number**	**inter**	**association**	**copy number**	**inter**	**pattern**	**GO name**
GO:0015807	−0.09	−0.85	−0.59	0.98	0.46	0.04	b24	L-amino acid transport
GO:0032228	−0.63	−1.21	−0.68	0.65	0.24	0.01	b24	regulation of synaptic transmission, GABAergic
GO:0050805	−0.94	−1.24	−0.63	0.22	0.24	0.04	b24	negative regulation of synaptic transmission
GO:0051932	−0.82	−1.35	−0.67	0.49	0.17	0.02	b24	synaptic transmission, GABAergic
GO:0042398	−0.77	−0.02	0.12	0.04	0.99	1	xl	amino acid derivative biosynthetic process
GO:0042401	−0.93	0.12	0.2	0.01	0.98	1	xl	biogenic amine biosynthetic process
GO:0030216	0.2	0.41	−0.03	0.8	0.03	1	yh	keratinocyte differentiation
GO:0031424	0.29	0.59	−0.01	0.81	0	1	yh	keratinization

A total of 8 GO terms were found as significant at FDR-adjusted $p<0.05$.
doi:10.1371/journal.pone.0010348.t004

following Goeman's nomenclature they are "competitive" tests [10]. Also, the way p-values are computed in the logistic model make of this approach a "gene sampling model" methodology [10].

It has been shown that, in general contexts of gene expression, where gene measurements are correlated within modules, GSA approaches that test "competitive" hypothesis based on "gene sampling models" are anticonservative [10]. This undesirable property also applies to the main effects of the bivariate logistic model as we could confirm in simulation studies (only in the case of internal correlation in the gene sets, which is the case of gene expression but not of the rest of the measurements used in this study). Interestingly, the consequence of gene correlation over the interaction effect, which is the main contribution of the proposed methodology, was the opposite and makes the method more conservative (see File S6). One way to avoid the bias of the particular context of gene expression would be to compute p-values based on a subject sampling permutation.

Care should be taken also when interpreting p-values from the method proposed here due to its "competitive" nature and the fact that it starts from a ranking statistic instead of the original data. Consequently, p-values test whether the distribution of the ranking statistic within each module is different to that of the whole genome. Therefore p-values do not extrapolate directly to the individual level class comparison which was done in order to compute the ranking statistic.

Discussion

Functional annotations, such as GO or KEGG pathways, have been used for the definition of modules of genes, carrying out common functional roles, in functional profiling methods [1,2]. All these methods, including the most recent versions, such as the GSA, can only deal with data that have been preselected or arranged by a unique variable (e.g. differential gene expression between cases and controls, etc.) The approach we are presenting here constitutes a novel and conceptually different proposal for the functional analysis of genomic experiments. It allows the simultaneous analysis of several variables, which can account for different properties of the genes. This approach can detect interactions between these variables that account for functional roles dependent on several genomic properties or measurements.

Table 5. Significant GO terms when differential expression and prognosis are taken into account in the model.

	Log odds ratio (model coefficients)			Adjusted p-value				
GO id	**diff.exp**	**prognosis**	**inter**	**diff.exp**	**prognosis**	**inter**	**pattern**	**GO name**
GO:0000087	−0.45	−0.08	−0.42	0.01	0.81	0	q2i	M phase of mitotic cell cycle
GO:0000279	−0.53	−0.07	−0.38	0.04	0.85	0	q2i	M phase
GO:0000910	−0.27	−0.09	−0.57	0.01	0.95	0	q2i	cytokinesis
GO:0007067	−0.47	−0.07	−0.4	0.04	0.9	0	q2i	mitosis
GO:0022618	−0.22	−0.33	−0.42	0.03	0.21	0	q2i	ribonucleoprotein complex assembly
GO:0051301	−0.38	0	−0.38	0.01	0.99	0	q2i	cell division
GO:0051726	−0.01	0.05	−0.22	0.03	0.91	0.01	q2i	regulation of cell cycle
GO:0045638	0.09	−0.35	−0.6	0.01	0.65	0.04	q4i	negative regulation of myeloid cell differentiation
GO:0000226	−0.08	0.16	−0.31	0.11	0.47	0.02	b24	microtubule cytoskeleton organization and biogenesis
GO:0000278	−0.34	0.04	−0.28	0.11	0.94	0	b24	mitotic cell cycle
GO:0007346	−0.3	−0.08	−0.39	0.07	0.9	0	b24	regulation of mitotic cell cycle
GO:0022403	−0.42	0	−0.31	0.09	0.99	0	b24	cell cycle phase
GO:0042254	−0.4	−0.45	−0.42	0.19	0.1	0.01	b24	ribosome biogenesis
GO:0006412	0.06	−0.28	−0.2	0.02	0.01	0.07	q4f	translation
GO:0006414	0.45	−1.12	−0.43	0	0	0.28	q4f	translational elongation
GO:0042312	0.45	0.08	−0.51	0.03	0.97	0.22	xh	regulation of vasodilation
GO:0000209	−0.25	0.55	0.13	0.94	0.01	1	yh	protein polyubiquitination
GO:0006066	0.08	0.2	−0.02	0.97	0.02	1	yh	alcohol metabolic process
GO:0010033	0.05	0.29	0	0.99	0.02	1	yh	response to organic substance
GO:0032944	−0.17	−0.7	0.06	0.97	0.02	1	yl	regulation of mononuclear cell proliferation
GO:0042098	−0.18	−0.61	0.08	0.95	0.04	1	yl	T cell proliferation
GO:0042110	0.03	−0.38	0.14	0.75	0.03	0.86	yl	T cell activation
GO:0042129	−0.33	−0.74	−0.02	0.99	0.05	1	yl	regulation of T cell proliferation
GO:0045321	−0.04	−0.28	0.06	0.92	0.03	1	yl	leukocyte activation
GO:0046649	−0.06	−0.33	0.07	0.89	0.02	1	yl	lymphocyte activation
GO:0046651	−0.19	−0.49	−0.05	0.99	0.05	1	yl	lymphocyte proliferation
GO:0050670	−0.17	−0.7	0.06	0.97	0.02	1	yl	regulation of lymphocyte proliferation
GO:0051249	−0.06	−0.44	0.24	0.52	0.04	0.71	yl	regulation of lymphocyte activation

Terms were significant at FDR-adjusted $p<0.05$.

doi:10.1371/journal.pone.0010348.t005

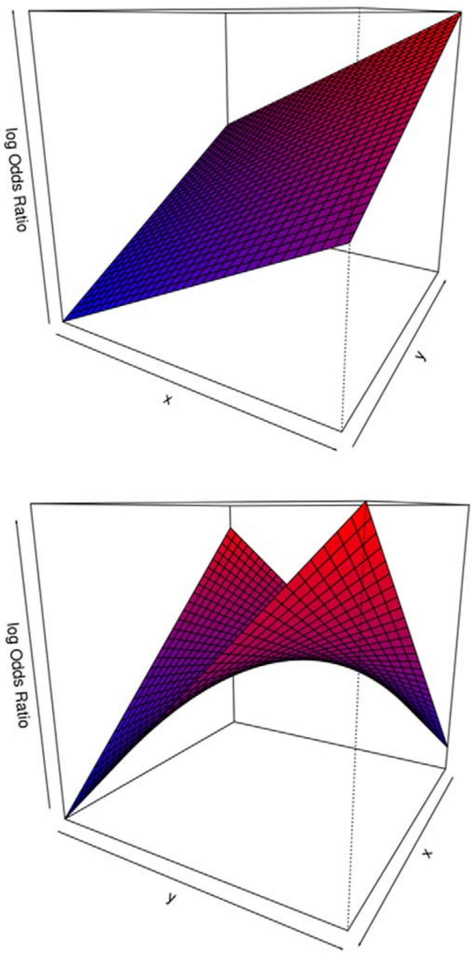


Figure 3. Surfaces described by the logistic model. The surface described by the logistic model is a plane when the interaction term (γ) is 0 (top) and a hyperbolic paraboloid when the interaction term (γ) is not zero (bottom).
doi:10.1371/journal.pone.0010348.g003

We have used for this purpose a logistic model. It has recently been shown that the application of the logistic model to one single variable (differential gene expression in this case) produces results conceptually similar to the outcome of any conventional GSA method [41]. The aim here is not to improve the one dimensional detection of gene modules related to the measurement, but to look for gene modules that have complex dependences on several genomic variables or measurements. Thus, in the first example we show how some functional GO categories depend on particular combinations of their transcription rates and mRNA stabilities. Different strategies can be used by the cellular machinery to ensure, for example, a rapid activation or a long lasting of a particular team of genes that cannot be explained with only one variable. Thus, combinations of several variables (e.g. a rapid transcription rate and a low mRNA stability can be useful for a rapid release and a rapid deactivation of a transient function) are on the root of many biological processes. The variables used can be properties of the genes or can be also measurements of behaviours such as their expression in a given condition. In the second case example we have analyzed a combination of gene property (splicing index) and gene behaviour (differential gene expression). The MD-GSA was able of detecting biological processes that depend on combinations of both variables and would remain undetected if the variables were independently analyzed. Finally, we applied the same concept to the same type of measurement (differential gene expression) in two different but related scenarios: a case control of dermatitis and another case-control of psoriasis. In this example we were able of finding common and distinctive altered functionalities of both related diseases that remained otherwise undetected with the conventional one-dimensional GSA. The combination of measurements that can be studied under this framework and their biological relevance is unimaginable. Thus the relation of biological roles to combinations of different parameters of different types, such as gene intrinsic properties (e.g. mRNA stability), gene behaviours (e.g. level of expression) or gene states (e.g. methylation, SNPs, copy number), etc., can be easily be studied using this approach.

Summarizing, MD-GSA constitutes a novel approach to the functional profiling of genome scale experiments that paves the way for a higher level understanding of the behaviour of functional modules in the cell.

Materials and Methods

Datasets and data preprocessing

Transcription rates and mRNA stabilities in yeast. Genome-wide values for the transcription rates (TR) and mRNA stabilities (RS) of the genes of yeast used in the first sub-section of results can be found in the supplementary material of the manuscript by Garcia-Martinez et al. [23].

Gene expression and splicing index. Okoniewski & Miller [44] used exon arrays to compare breast cancer cell line MCF7 (fetal calf serum) to non tumorgenic breast epithelial cell line MCF10A (horse serum). They estimated differential gene expression using standard t-statistics and alternative splicing using the splicing index described in [30]. Since the splicing index is defined for each exon, we have used here median values to provide splicing measurements at a gene level. Thus, we have two numerical variables recorded for each gene in the study. The first one assesses the variation in the general expression level. The second one quantifies the change in splicing pattern of the gene, independently of its expression levels.

Differential expression in psoriasis and dermatitis. Expression data from two separated case control experiments where combined in this analysis. The first experiment consisted of the comparison of lessional and non lessional skin samples in atopic dermatitis patients [32] (data were obtained from the GEO database, accession: GSE5667). The second experiment compared affected and unaffected skin in psoriatic patients [31] (GEO database, accession: GSE6710). Separated gene expression analyses of these two datasets were performed using standard methods: RMA algorithm [45] was used to normalize data within each of the experiments. The limma package [46] from Bioconductor [47] was used to estimate, separately for each of the studies, differential gene expression between diseased and non-diseased skin. Hence, two experimental measurements (limma t-statistics) where generated for each gene and used in the analysis: a first measurement of differential gene expression in dermatitis and a second measurement of differential gene expression in psoriasis.

Combined analysis of several breast cancer genomic measurements. Data used in the combined analysis of genomic measurements, in the results section, were taken from the supplementary material of [38]. SNP association to disease was measured using Odds Ratio (OR) of their corresponding minor allele frequencies. Then, the magnitude of the association of each gene to the disease was obtained as the value of association of the

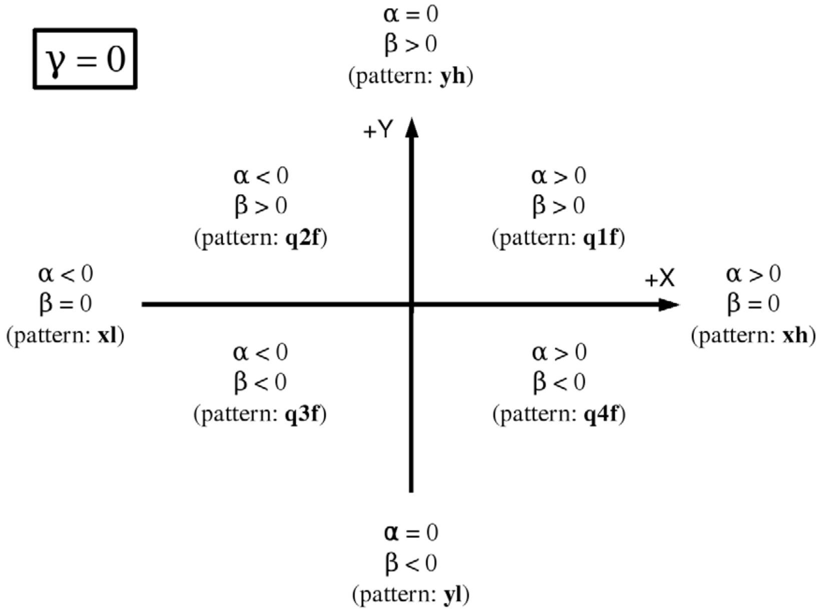


Figure 4. Location of the areas where genes are more likely to be annotated to the function F depending on the coefficients of the fitted model. When $\gamma = 0$ the fitted surface is a plane which slope grows towards the area.
doi:10.1371/journal.pone.0010348.g004

SNP more associated to the disease among all the SNPs mapping in the gene (or near the gene and being in linkage disequilibrium) [21,38]. Differences in gene expression between tumour and normal breast tissues where estimated using t-statistics. Cox regression models where used to correlate survival time and gene expression, yielding a "prognosis" value for each gene (genes with "high" hazard ratios in the Cox model are associated to poor prognosis; genes with "low" hazard ratios associated to good prognosis). Another genomic measurements used was the average copy number for each gene in luminal B tumours, obtained from the hybridization intensity of the probesets corresponding to each gene (taken from the additional material of our study [38]).

Annotation Data. Functional modules are defined according the annotations of the GO [4] and the KEGG Pathway [48] repositories. Functional modules of more than 500 genes where considered to be too general to be informative so they where filtered out. Functional modules having less than 10 genes annotated to them where considered to be too small to be properly fitted by the multivariate logistic model and where also discarded.

Multi dimensional GSA (MD-GSA) using a logistic model that considers more than one variable

Logistic regression is a well established statistical methodology used to model the probability of occurrence of a binary event as a function of some other independent variables [49]. In the context of genomic studies, univariate logistic models have been shown to be suitable to perform gene set enrichment analysis [41].

Modelling functional class membership in terms of some measurement, X, of differential gene expression between two conditions as follows:

$$\ln\left(\frac{P(g\in F)}{P(g\notin F)}\right) = K + \alpha X \qquad (1)$$

we can call the gene set F enriched in one of the conditions a significant estimate of the α coefficient is obtained [41].

In this paper we extend the use logistic models to perform a multidimensional gene set enrichment analysis. Our model describes the probability of a gene belonging to a functional class as a function of not one, but several experimental measurements. For two of those measurements the model will be as follows:

$$\ln\left(\frac{P(g\in F)}{P(g\notin F)}\right) = K + \alpha X + \beta Y + \gamma XY \qquad (2)$$

where α and β are the main effects and γ is the interaction effect.

In a case-control study measuring, for instance, gene expression and genotype, we could model the probability of genes being annotated to a GO term as a function of both, differential gene expression (X) and allelic association to disease (Y).

Modelling not only the additive effects but also the interaction term, we accurately describe how the genes in a gene set are related to both measurements X and Y together, allowing for the detection of enrichment patterns which will remain unnoticed in two independent univariate analyses.

The model in equation (2) describes the log odds ratio of a gene being annotated to functional module F as a function of two variables, X and Y. The shape of this surface when embedded in a 3D space is that of a plane if the interaction coefficient γ is zero (Figure 3, top), or a hyperbolic paraboloid, also called saddle surface, when the estimate of γ is different from zero (Figure 3, bottom). Hence, from the sign and significance of the fitted coefficients, we can find the direction in the two dimensional space XY in which the genes annotated to the function F are more likely to be found.

When γ is zero the sign of the coefficients α and β describe the slopes of the plane and therefore, the direction towards which the probability of genes being annotated is greater. Figure 4 describes the areas where genes belonging to a functional module are more likely to be found, depending on the estimated α and β coefficients of the logistic model (2) and provided that the estimate of γ is not significantly different from zero.

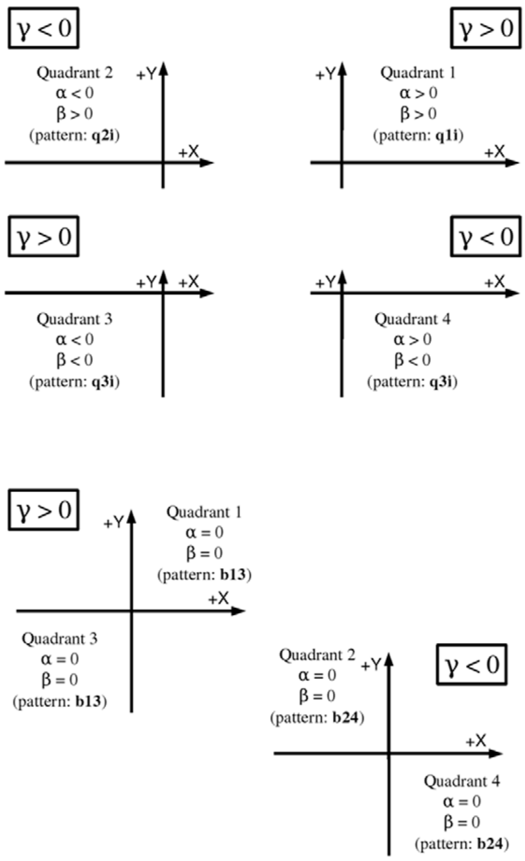


Figure 5. Location of the areas where genes are more likely to be annotated to the function F depending on the coefficients of the fitted model. If $\gamma \neq 0$ the fitted surface is a hyperbolic paraboloid, when $\alpha \neq 0$ and $\beta \neq 0$ (top part) the most likely area to find genes annotated to F is the quadrant opposite to the saddle point of the surface. When $\alpha = 0$ and $\beta = 0$ (bottom part) the saddle point of the surface is in the (0,0) and the genes annotated to the function F are more likely to be found in two opposite quadrants, reflecting the bimodality of the function F.
doi:10.1371/journal.pone.0010348.g005

When γ is different from zero the interaction dominates the growth of the log odds ratio while the saddle point in the surface has the coordinates $(-\beta/\gamma, -\alpha/\gamma)$. If for instance, for a particular functional module F, all estimated coefficients are positive, then, the saddle point of the hyperbolic paraboloid will be in the third quadrant and the surface will grow to the infinite in the first quadrant. As the surface represents how likely we are to find genes annotated to module F in the plane XY, we will conclude that the module F is located towards the firs quadrant. Moreover, as the interaction effect is positive we know that the evidence of this localization is greater than the one we will get from separated analysis of each one of the dimensions X and Y on their own (following equation 1). Then, biological interpretation can be done recalling the meaning of the X and Y quantities. Figure 5 (top) describes the areas where genes belonging to a functional module are more likely to be found, depending on the estimates of α, β and γ and when γ is estimated to be different from zero.

If it was the case that just the interaction coefficient γ would be different from zero, then the saddle point will be the (0, 0) and the genes annotated to functional module F will be allocated to opposite quadrants of the XY space; the first and the third quadrant if $\gamma > 0$; the second and the fourth quadrants if $\gamma < 0$. In this latest case we will call the functional module F bimodal and the biological interpretation will be that, genes in F are effectively spited up in two groups of opposite patterns. Figure 5 (bottom) describes the areas where genes belonging to a functional module are more likely to be found, if the estimates of α and β are zero.

Table 6 shows how to interpret all possible combinations of α, β and γ estimates.

Wald statistics to test the main effect coefficients and the interaction effects [41]. Other approaches like likelihood ratio tests could also have been used.

As one logistic regression model needs to be fit for each functional module in the analysis, multiple testing occurs and p-value correction must be performed. In this paper we use Benjamini and Hochberg [50] approach to correct all p-values of the same parameter of the model α, β or γ.

Implementation

The proposed algorithm has been implemented as an R library available at http://bioinfo.cipf.es/supplementary/multidimensional_GSA, released under the GPL license.

Supporting Information

Figure S1 GO terms significantly associated to the interaction between transcription rate and mRNA stability in yeast. Octagons represent terms with p-values<0.05, after adjustment for multiple testing using the popular FDR [48]. White squares represent non-significant terms connecting the significant terms found. The picture has been obtained using the GOGraphViewer option of the Babelomics package [49].
Found at: doi:10.1371/journal.pone.0010348.s001 (1.79 MB JPG)

Figure S2 GO terms significantly associated to the interaction between gene expression and splicing index. Octagons represent terms with p-values<0.05, after adjustment for multiple testing using the popular FDR [48]. White squares represent non-significant terms connecting the significant terms found. The picture has been obtained using the GOGraphViewer option of the Babelomics package [49].
Found at: doi:10.1371/journal.pone.0010348.s002 (1.07 MB JPG)

Figure S3 GO terms significantly associated to the interaction between differential gene expression in psoriasis and dermatitis. Octagons represent terms with p-values<0.05, after adjustment for multiple testing using the popular FDR [48]. White squares represent non-significant terms connecting the significant terms found. The picture has been obtained using the GOGraphViewer option of the Babelomics package [49].
Found at: doi:10.1371/journal.pone.0010348.s003 (1.66 MB JPG)

Figure S4 GO terms significantly associated to the interaction between copy number and gene association to breast cancer (see text). Octagons represent terms with p-values<0.05, after adjustment for multiple testing using the popular FDR [48]. White squares represent non-significant terms connecting the significant terms found. The picture has been obtained using the GOGraphViewer option of the Babelomics package [49].
Found at: doi:10.1371/journal.pone.0010348.s004 (1.12 MB JPG)

Figure S5 GO terms significantly associated to the interaction between differential expression and prognosis of breast cancer. Octagons represent terms with p-values<0.05, after adjustment for multiple testing using the popular FDR [48]. White squares represent non-significant terms connecting the significant terms found. The picture has been obtained using the GOGraphViewer option of the Babelomics package [49].

Table 6. Interpretation of all relevant combinations of α, β and γ estimates.

α	β	γ	pattern identifier	pattern	description
+	+	+	q1i	Quadrant 1 with interaction	F is allocated towards one of the quadrants and the evidence is greater than just the additive evidences from the univariate analysis.
+	0	+			
0	+	+			
−	−	+	q3i	Quadrant 3 with interaction	
−	0	+			
0	−	+			
−	+	−	q2i	Quadrant 2 with interaction	
−	0	−			
0	+	−			
+	−	−	q4i	Quadrant 4 with interaction	
+	0	−			
0	−	−			
0	0	+	b13	Bimodal + (quadrants 1 and 3)	F is split in two opposite quadrants.
0	0	+	b24	Bimodal − (quadrants 2 and 4)	
+	+	0	q1f	Quadrant 1 flat	F is allocated towards one of the quadrants and the evidence is similar to the additive evidences from the univariate analysis.
−	−	0	q3f	Quadrant 3 flat	
−	+	0	q2f	Quadrant 2 flat	
+	−	0	q4f	Quadrant 4 flat	
+	0	0	xh	X high (+) values	F is enriched just in the first condition.
−	0	0	xl	X low (−) values	
0	+	0	yh	Y high (+) values	F is enriched just in the second condition.
0	−	0	yl	Y low (−) values	

doi:10.1371/journal.pone.0010348.t006

Found at: doi:10.1371/journal.pone.0010348.s005 (0.76 MB JPG)

Table S1 Excel file containing significant GO terms obtained upon the application of the logistic model to the mRNA stability (RS) and to the transcription rate (TR) variables independently.
Found at: doi:10.1371/journal.pone.0010348.s006 (0.19 MB XLS)

File S1 A) GO Biological Process terms and B) KEGG pathways, significant for Transcription Rate (TR), RNA Stability (RS) and their interaction, along with the corresponding graphical representations. In the plots blue lines intersect in the mean of the distribution of all the values and red lines intersect in the mean of the distribution of values of the genes corresponding to the GO term analysed. Blue ellipse delimits the confidence interval for all the values and red ellipse delimits the confidence interval for the GO term analysed. The red ellipse marks the trend of the relationship between both variables.
Found at: doi:10.1371/journal.pone.0010348.s007 (9.04 MB PDF)

File S2 A) GO Biological Process terms and B) KEGG pathways, significant for alternative splicing and differential gene expression and their interaction, along with the corresponding graphical representations. In the plots blue lines intersect in the mean of the distribution of all the values and red lines intersect in the mean of the distribution of values of the genes corresponding to the term analysed. Blue ellipse delimits the confidence interval for all the values and red ellipse delimits the confidence interval for the term analysed. The red ellipse marks the trend of the relationship between both variables.
Found at: doi:10.1371/journal.pone.0010348.s008 (9.22 MB PDF)

File S3 A) GO Biological Process terms, and B) KEGG pathways, significant for differential gene expression in dermatitis and psoriasis case-control studies and their interaction, along with the corresponding graphical representations. In the plots blue lines intersect in the mean of the distribution of all the values and red lines intersect in the mean of the distribution of values of the genes corresponding to the term analysed. Blue ellipse delimits the confidence interval for all the values and red ellipse delimits the confidence interval for the term analysed. The red ellipse marks the trend of the relationship between both variables.
Found at: doi:10.1371/journal.pone.0010348.s009 (30.60 MB ZIP)

File S4 A) GO Biological Process terms, and B) KEGG pathways, significant for gene association (derived from genotyping, see text) association data and genomic copy number in breast cancer and their interaction, along with the corresponding graphical representations. In the plots blue lines intersect in the mean of the distribution of all the values and red lines intersect in the mean of the distribution of values of the genes corresponding to the term analysed. Blue ellipse delimits the confidence interval for all the values and red ellipse delimits the confidence interval for

the term analysed. The red ellipse marks the trend of the relationship between both variables.
Found at: doi:10.1371/journal.pone.0010348.s010 (0.80 MB PDF)

File S5 A) GO Biological Process terms, and B) KEGG pathways, significant for prognosis and differential expression in a case-control study of breast cancer and their interaction, along with the corresponding graphical representations. In the plots blue lines intersect in the mean of the distribution of all the values and red lines intersect in the mean of the distribution of values of the genes corresponding to the term analysed. Blue ellipse delimits the confidence interval for all the values and red ellipse delimits the confidence interval for the term analysed. The red ellipse marks the trend of the relationship between both variables.
Found at: doi:10.1371/journal.pone.0010348.s011 (2.22 MB PDF)

File S6 Interaction simulation study. A simulation study of the bias in p-value estimates for the interaction term of the bivariate logistic model.
Found at: doi:10.1371/journal.pone.0010348.s012 (0.15 MB DOC)

Author Contributions

Conceived and designed the experiments: DM JD. Performed the experiments: DM. Analyzed the data: DM. Wrote the paper: JD.

References

1. Dopazo J (2009) Formulating and testing hypotheses in functional genomics. Artif Intell Med 45: 97–107.
2. Huang DW, Sherman BT, Lempicki RA (2008) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1–13.
3. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402: C47–52.
4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
5. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277–280.
6. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, et al. (2007) Reactome: a knowledge base of biologic pathways and processes. Genome Biol 8: R39.
7. Dopazo J (2006) Functional interpretation of microarray experiments. Omics 10: 398–410.
8. Khatri P, Draghici S (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 21: 3587–3595.
9. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34: 267–273.
10. Goeman JJ, Buhlmann P (2007) Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23: 980–987.
11. Al-Shahrour F, Diaz-Uriarte R, Dopazo J (2005) Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics 21: 2988–2993.
12. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20: 93–99.
13. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550.
14. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203–218.
15. Kim SY, Volsky DJ (2005) PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics 6: 144.
16. Kitano H (2004) Cancer as a robust system: implications for anticancer therapy. Nat Rev Cancer 4: 227–235.
17. Bentink S, Wessendorf S, Schwaenen C, Rosolowski M, Klapper W, et al. (2008) Pathway activation patterns in diffuse large B-cell lymphomas. Leukemia 22: 1746–1754.
18. Bardelli A, Velculescu VE (2005) Mutational analysis of gene families in human cancer. Curr Opin Genet Dev 15: 5–12.
19. Al-Shahrour F, Arbiza L, Dopazo H, Huerta-Cepas J, Minguez P, et al. (2007) From genes to functional classes in the study of biological systems. BMC Bioinformatics 8: 114.
20. Wu C, Delano DL, Mitro N, Su SV, Janes J, et al. (2008) Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. PLoS Genet 4: e1000070.
21. Medina I, Montaner D, Bonifaci N, Pujana MA, Carbonell J, et al. (2009) Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. Nucleic Acids Res 37: W340–344.
22. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, et al. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature.
23. Garcia-Martinez J, Gonzalez-Candelas F, Perez-Ortin JE (2007) Common gene expression strategies revealed by genome-wide analysis in yeast. Genome Biol 8: R222.
24. Perez-Ortin JE (2007) Genomics of mRNA turnover. Brief Funct Genomic Proteomic 6: 282–291.
25. Al-Shahrour F, Diaz-Uriarte R, Dopazo J (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20: 578–580.
26. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, et al. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302: 2141–2144.
27. Faustino NA, Cooper TA (2003) Pre-mRNA splicing and human disease. Genes Dev 17: 419–437.
28. Srinivasan K, Shiue L, Hayes JD, Centers R, Fitzwater S, et al. (2005) Detection and measurement of alternative splicing using splicing-sensitive microarrays. Methods 37: 345–359.
29. Bitton DA, Okoniewski MJ, Connolly Y, Miller CJ (2008) Exon level integration of proteomics and microarray data. BMC Bioinformatics 9: 118.
30. Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, et al. (2007) Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biol 8: R64.
31. Reischl J, Schwenke S, Beekman JM, Mrowietz U, Sturzebecher S, et al. (2007) Increased expression of Wnt5a in psoriatic plaques. J Invest Dermatol 127: 163–169.
32. Plager DA, Leontovich AA, Henke SA, Davis MD, McEvoy MT, et al. (2007) Early cutaneous gene transcription changes in adult atopic dermatitis and potential clinical implications. Exp Dermatol 16: 28–36.
33. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, et al. (2007) The genomic landscapes of human breast and colorectal cancers. Science 318: 1108–1113.
34. Pinkel D, Albertson DG (2005) Array comparative genomic hybridization and its applications in cancer. Nat Genet 37 Suppl: S11–17.
35. Bignell GR, Huang J, Greshock J, Watt S, Butler A, et al. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res 14: 287–295.
36. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, et al. (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16: 1136–1148.
37. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447: 1087–1093.
38. Bonifaci N, Berenguer A, Diez J, Reina O, Medina I, et al. (2008) Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes. BMC Med Genomics 1: 62.
39. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39: 870–874.
40. Singh RK, Rinehart CA, Kim JP, Tolleson-Rinehart S, Lawing LF, et al. (1996) Tumor cell invasion of basement membrane in vitro is regulated by amino acids. Cancer Invest 14: 6–18.
41. Sartor MA, Leikauf GD, Medvedovic M (2008) LRpath: A logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics 25: 211–217.
42. Montaner D, Minguez P, Al-Shahrour F, Dopazo J (2009) Gene set internal coherence in the context of functional profiling. BMC Genomics 10: 197.
43. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E (2004) Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. Neurochem Res 29: 1213–1222.
44. Okoniewski MJ, Miller CJ (2008) Comprehensive analysis of affymetrix exon arrays using BioConductor. PLoS Comput Biol 4: e6.
45. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249–264.
46. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: Article3.

47. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
48. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36: D480–484.
49. Agresti A (2002) Categorical data analysis. HobokenNew Jersey: John Wiley and Sons.
50. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57: 289–300.

Chapter 6

Summary and Conclusions

In this work we have introduced several bioinformatic tools and methods for the analysis of genomic data. We have specially focused in some of our developments relating *gene set analysis* algorithms. We have argued how, despite being an already established step in most high throughput experiments, current *functional profiling* methods are oversimplified and cannot efficiently address complex experimental designs. We have presented GEPAS, a general purpose analysis resource for the analysis of microarray data. We have illustrated how it is closely tied to the *functional profiling* suite Babelomics, up to the point that both tools have been fused in their latest release. After the exposition of such general pipeline for the analysis of genomic data, we have highlighted how it lacked some desirable tools or parts of the work flow. The most technical developments carried out in this thesis aimed to fill some of those missing characteristics. Thus, we have contributed to this general analysis framework with particular applications as for instance the module for the *multidimensional gene set* analysis, or the utility that allows for the *weighting* of genes. But we have also provided several conceptual advances in the field, as for instance the avoidance of cutoffs in the methodology set up by FatiScan, the understanding of *gene sets* as non discrete entities, or the depiction of the modular approach to carry out data analysis. Finally, we have also done some additions to the biological science as it is the empirical estimation of the internal coherence of GOs and KEGGs.

Thus, as a results of the accomplished research, we can particularly conclude that:

1. *Logistic regression* may be successfully applied in genomic *functional profiling*. This common statistical model behaves as previously used approaches; keeps their suitable properties while overcomes some of their major drawbacks and extends the scope of experiments in which *gene set analysis* may be applied.

2. Jointly analyzing the results of several kinds of genomic measurements in terms of *gene sets* is feasible and biologically meaningful. Moreover, to combine the different sources of genomic information within the *functional profiling* step of the analysis, may be an effective strategy for data integration.

3. Data base defined functional blocks are ultimately heterogeneous entities and thus, not all genes annotated to them will show a coherent pattern in their measurements. Despite of that, the internal homogeneity of *gene sets* may be quantified from previous experiments. Such quantification may spot the *annotations* that truly describe functional blocs of genes. This may simply be used to filter out databases before a meaningful *gene set analysis* is carried out but it may be as well included into the analysis itself, taking advantage of approaches similar to our *weighting schema*.

4. The KEGG pathways and the Gene Ontology terms are internally *less coherent* in their expression levels than expected. This is likely to occur to most other biological databases.

5. In any experiment carried out at a genome level there is an underlying "universe" of observed genes. This background of genomic features is implicit in many steps of the analysis, from data *normalization*, to *gen set analysis*, passing through p-value correction. Care should always be taken for it not to bias our results or their interpretation.

6. Not always the genes or genomic features showing the greatest differences across biological conditions are the most relevant ones in an experiment. Small but coordinated changes in genes that act together (in a *gene set* for instance) may be responsible for many changes in phenotype. Contrary to *functional enrichment analysis* approaches, *gen set analysis* methodologies are able to detect such small but coordinated changes. It is up to the researcher to decide which paradigm is more suitable to interpret its experiment. In any case, *logistic regression* models are general enough to encompass both analytical schemes.

7. Functional profiling methods provide more power to the analysis first because they incorporate the *extra* information of the *annotation*. Second because they pool and combine the information from the several genes conforming the *gene sets*. And third because *shifting in the observational unit* from the gene to the *gene sets*, implies a data reduction that results in weaker multiple testing correction. Additionally, the data reduction that occurs due to this *shift in the unit of observation* is more meaningful than other available procedures like for instance PCA[1], because the data reduction is done attending to a predefined biological criteria and not to a general purpose statistical considerations.

8. Taking a modular approach to investigate genomic data, separating algorithmically the conceptual steps of the analysis, makes easier the biological interpretation of results. Such a modular approach simplifies software implementation and maintenance and facilitates the replacement or improvement of parts of the analytical pipeline. It also makes possible to reuse ideas, code or software in upcoming research contexts. The modular approach is particularly useful when the analysis procedure needs to make a transition between different levels of the biological organization as it occurs in the *gene set analysis* due to the *shift in the observational unit*.

[1] Principal Component Analysis.

9. Interestingly from the statistical perspective, in this modular approach, the outcome of one step of the analysis is not treated as an evidence to accept or reject a hypothesis but as an *index* which accounts for some genomic property of our biological sample, and that needs to be reanalyzed gain in the following milestone of the pipeline. Following some of our examples, the p-values resulting from a differential expression test wont be taken as the probabilities of a *type I errors*, but as direct indices accounting for the differences in expression at genomic level. Such p-values will constitute the new sample to be explored in the subsequent step of the analysis pipeline.

10. Well designed presentation of results is crucial for its interpretation of genomic experiments due to the complexity of the results themselves. But not less important is to use the appropriated data structures during the analysis process. This will highlight the biologically relevant characteristics in our data and will help conducting the analysis.

11. When creating a software for data analysis, including several available methods to perform the same task can be helpful for users, but more importantly, it will help them intuitively understanding how complex analyses work.

12. Generating web interfaces is an efficient way to bring algorithms and methods closer to a wide range of users, from experts in the field to casual analysts.

I hope the above summary points out our main contributions to a research field that is anything but completed: that of the *functional profiling of genomic experiments*. The growing importance of genomics, the increasing facility to collect data and the continuous development of databases, allows us foreseeing an increasing demand for accurate *gene set analysis* methods.

Besides the always recurring need to speed up computation and algorithms, the future development of the *gene set* methodologies will have to address major challenges. Effectively dealing with more tan two genomic dimensions at a time, will remain a hot topic as long as new experimental techniques enter in the scene, allowing for the observation of novel biological features. Developing *set*[1] tests for such novel features will necessarily require the extrapolation of information form one genomic dimension to the newer one. GO annotation, for instance, currently predicated of genes, will need to be somehow extended to the regulatory elements of the genome before functional profiling can be carried out over their measurements. And of course the new *set* methods to come will need to take such extrapolation into account to perform unbiased analyses. Finlay, the growing complexity of genomic knowledge stored in biological databases, is describing the *topology* of the *genomic feature sets* increasingly better. The smart modeling of such internal structure of the *sets*, besides the consideration of the relationships among them, will surely constitute the new paradigm of *functional profiling* methods, bring the field closer to what is currently referred to as *systems biology*.

[1] Note the disappearance of the word *gene*. In the immediate future we shall be talking of *genomic feature sets*,

Appendix A

New trends in the analysis of functional genomic data.

In this appendix you can fin the original writing of my document “New trends in the analysis of functional genomic data.” This was first published as a chapter of the book “Progress in Industrial Mathematics at ECMI 2006.” edited by Luis L. Bonilla (Bonilla, 2008) in Springer.

New Trends in the Analysis of Functional Genomic Data

David Montaner[1,2], Fatima Al-Shahrour[1], and Joaquin Dopazo[1,2]

[1] Bioinformatics Department, Centro de Investigacin Prncipe Felipe (CIPF) Autopista del Saler 16, E-46013, Valencia, Spain. dmontaner@cipf.es

[2] Functional Genomics Node, INB, CIPF. Autopista del Saler 16, E-46013, Valencia, Spain

1 Replications of the same statistical test

Most analyses carried out using high throughput data consist of the repetition of the same statistical test for all genes in the dataset. As a result of such replicated analysis we get, for each gene, several estimates of statistical parameters: statistics, p-values or confidence intervals. Being aware that most statistical methods were developed to test for a single hypothesis, researchers will usually correct p-values for multiple testing before choosing a cut-off that will indicate the rejection of the null hypotheses, whichever it is. Once chosen the genes with alternative pattern (meaning different form the one stated in the null hypothesis) the next step is to biologically interpret such departure from hypothesis. Different repositories of functionally relevant biological information such as Gene Ontology [1], KEGG [2] or Interpro [3] are available and can be used for the functional annotation of genome-scale experiments. Thus the functional properties of the selected genes can be analysed.

The trouble of this approach is that, by discarding genes with p-values above the cut-off, we loose most of our information. Not only we loose the measurements taken over the genes but also the functional annotation that could be linked to them from repositories, making it difficult the biological interpretation of results.

2 Blocks of functional genes

Aiming to prevent such waste of information, some authors have recently proposed to directly analyse the behaviour of blocks of functionally related genes in a whole-genome context. The Gene Set Enrichment Analysis (GSEA) [4,5], the FatiScan [6,7] or the Global Test [8,9] constitute examples of this type of approach inspired from systems biology. This three methodologies address the issue of whether the general expression pattern of a group of genes, for example a GO term or a KEGG pathway, changes across biological conditions. Here we will discus just some particular aspects of these methods but a more general view of this and similar methods can be found in Dopazo's revision of 2006 [10].

The Global Test uses generalised linear models to study the relationship between the expression of the genes of the block of interest and a characteristic associated to each biological sample. Such characteristic may be a categorical condition, like the class of the microarray in the context of differential gene expression, or a continuous variable such as a level of a metabolite. In this approach we can see a change in the philosophy of the analysis. The unit of interest is not any more a single gene but a block of genes with a common biological meaning. This new way of looking at the data provides, among others, obvious advantages for the biological interpretation of results and for the p-value adjustment. We just need to correct by the number of blocks, usually smaller than the number of genes.

3 The overall approach

The block of genes is also the unit of interest of the GSEA and the FatiScan. These two methods are similar to the Global Test in that they are also used to discover groups of genes which overall expression pattern changes across biological conditions. Nevertheless, GSEA and FatiScan consider all genes in the data when analysing each of the blocks. They compare the pattern of the genes of one block with the general pattern of the genes in the whole dataset. GSEA is particularly designed for the two class comparison context while FatiScan may be applied in a wider range of studies.

The rationale underlying both methodologies is that, if a property of genes can be described using a continuous index, then the statistical distribution of such index within a functional block of genes can be compared to the general distribution of the index across all genes in the data. We can therefore asses whether the property described by the index depends on the characteristic that defines the block of genes

As said before GSEA is developed for the two class comparison. In this methodology, a signal-to-noise ratio comparing mean expression across classes is computed for each gene in the dataset. This statistic can be seen as a continuous index that ranks the genes according to their differential expression, from those more expressed in one of the biological conditions to those more expressed the second condition, passing through those genes non differentially expressed. Then, given a block of genes, for instance a functional class that we may be interested in, we can compare the distribution of the signal-to-noise ratio of the genes in the block to the distribution of the same statistic in the remaining genes. If the values of the signal-to-noise ratio are, for instance, systematically higher in the genes of the block compared to the genes in the whole dataset, we will conclude that, as a block, the genes of the functional class of interest are overexpressed in one of the biological conditions. GSEA uses a modification of the Kolmogorov-Smirnov test to asses differences between the signal-to-noise ratio in the class of interest and in the rest of the genes. Significance of the modified Kolmogorov-Smirnov statistic is computed in GSEA using permutations of the expression data. The original expression data is permuted several times, the signal-to-noise ratios are calculated over each permuted expression dataset and the modified Kolmogorov-Smirnov statistic is computed over each new distribution of the signal-to-noise ratio. Thus GSEA can estimate the random variability of the Kolmogorov-Smirnov statistic and test its significance in the original data.

4 Detaching concepts and algorithms

FatiScan follows the same analytical philosophy than GSEA but with a more general and flexible approach. FatiScan implements a segmentation test which checks for asymmetrical distributions of biological labels associated to genes ranked by any index. The main difference is that FatiScan does not implement a permutation test to asses such asymmetry. Therefore, the algorithm that computes the index and the algorithm that analyses the distribution of the index are completely separated so the calculations can be done in two different steps. This means that FatiScan can be used to study the relationship between biological labels associated to genes and any type of experiment whose outcome is a sorted list of genes or a variable that can be used to rank genes according to some characteristic of interest. Block of genes sorted by differential expression between two experimental conditions can be studied as it would be done using GSEA. But with FatiScan we can also consider many other gene properties or characteristics.

We can easily explore the correlation between gene expression and a clinical continuous variable such as the level of a metabolite. First, for each gene we will compute the correlation between its expression measurements and the levels of the metabolite. Thus we can range the genes from those which expression is more positively correlated to the levels of the metabolite to those inversely correlated, passing by genes which expression does not correlate with the clinical variable. In a second step, FatiScan explores the distribution of such correlation measurements, testing whether the distribution of correlations within a block of genes is different from the overall distribution of correlation in the dataset.

We can fit a Cox proportional hazard model to each gene in our data in order to study the relationship between gene expression and survival times. The estimates of the slope coefficients may be used as an index that ranks genes from those which increased expression is associated with long time survival to those which increased expression is associated to an early death. After computing this rank-index, FatiScan will find those blocks of genes for which the distribution of the slopes differs from the global distribution of the slopes.

The complete separation of the two steps in FatiScan analysis is the key point which provides its flexibility to the method. Such flexibility makes possible to handle many different sources of information, not only microarray gene expression data. Any lists of genes ranked by any other experimental or theoretical criteria can be studied. Genes can be for example arranged by physico-chemical properties, mutability, structural parameters and so on. In order to understand whether there is some biological feature, characterised by the blocks of genes, which is related to the experimental parameter studied.

5 Coda

The three methodologies here mentioned illustrate two of the main new conceptual trends in the analysis of functional genomic data.

The first one is the change of the descriptive unit used to address biological studies, shifting from gene to functional class. Gene still remains the unit of measured information, as what we record at the end is gene expression. But the conceptual entity over which biological interpretation is done, is the functional class of genes. New analytical strategies, like those above mentioned, should consider this fact in order to use the available information in the most efficient an meaningful way

The second one is probably more subtle but not less important. Usual genomic studies follow the classical statistical approach in which one or several hypotheses are stated, estimate statistics and p-values are computed from data and finally, hypotheses are accepted or rejected depending on such estimated values. The analytical approach explicit in FatiScan an implicit in GSEA shows how estimated values provided by one first statistical analysis are not directly interpreted in terms of acceptance or rejection of hypotheses. Instead they are treated as variables quantifying some characteristic of the genes under study. This new variables may then be analysed using statistical methodologies. Thus, statistical results of one step of the analysis become themselves a new dataset which needs to be explored in a second analytical step. As we see, modular implementations of complex data analysis strategies like FatiScan, seem to be both, conceptually useful for the analysis of biological data and computationally advantageous, calling for the development of the theoretical framework within which combinations of statistical methods can be properly done.

References

[1] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet., 25, 25-29 (2000)

[2] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M.; The KEGG resource for deciphering the genome. Nucleic Acids Res., 32, D277-D280 (2004)

[3] Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., et al.; InterPro, progress and status in 2005. Nucleic Acids Res., 33, D201-D205 (2005)

[4] Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al; PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genet., 34, 267-273 (2003)

[5] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al; Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA, 102, 15545-15550 (2005)

[6] Al-Shahrour, F., Diaz-Uriarte, R., Dopazo, J.; Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information Bioinformatics, 21, 2988-2993 (2005)

[7] Al-Shahrour F., Minguez P., Trraga J., Montaner D., Alloza E., Vaquerizas J.M., Conde L., Blaschke C., Vera J. and Dopazo J.; BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucl Acids Res., 34, W472-W476 (2006)

[8] Goeman, J.J., van de Geer, S.A., de Kort, F., van Houwelingen, H.C.; A global test for groups of genes: testing association with a clinical outcome. Bioinformatics, 20, 93-99 (2004)

[9] Goeman J.J., Oosting J., Cleton-Jansen A.M., Anninga J.K., van Houwelingen H.C.; Testing association of a pathway with survival using gene expression data. Bioinformatics, 21, 1950-1957 (2005)

[10] Dopazo, J.; Functional Interpretation of Microarray Experiments. OMICS: A Journal of Integrative Biology, 10, 398-410 (2006)

Appendix B

Multidimensional Gene Set Paper Supplementary materials.

In this appendix you can find the supplementary materials from **Montaner** and Dopazo 2010.

Multi Dimensional Gene Set Analysis

David Montaner and Joaquín Dopazo.

Supplementary material

A simulation study of the bias in p-value estimates for the interaction term of the bivariate logistic model.

Montaner et all. (2009) collected microarray data for 3034 human samples measured under the most diverse biological conditions. They combined all information into a huge data matrix of gene expressions and used it to estimate internal correlation of Gene Ontology Biological Process terms.

We use this same data set to simulate microarray case-control experiments where differentially expressed genes do not exist but the real biological correlation structure between genes is preserved. We applied our Multi Dimensional Gene Set Analysis to these simulated data sets in order to estimate false positive rates for the interaction term and their relationship with the internal correlation of the tested GO term.

500 data sets where randomly sampled (with no replacement) from the 3034 available microarrays. Each of this 500 datasets was constituted by:

- 10 arrays randomly labeled as *cases* within the condition **A**
- 10 different arrays randomly labeled as *controls* within the condition **A**
- 10 different arrays randomly labeled as *cases* within the condition **B**
- 10 different arrays randomly labeled as *controls* within the condition **B**

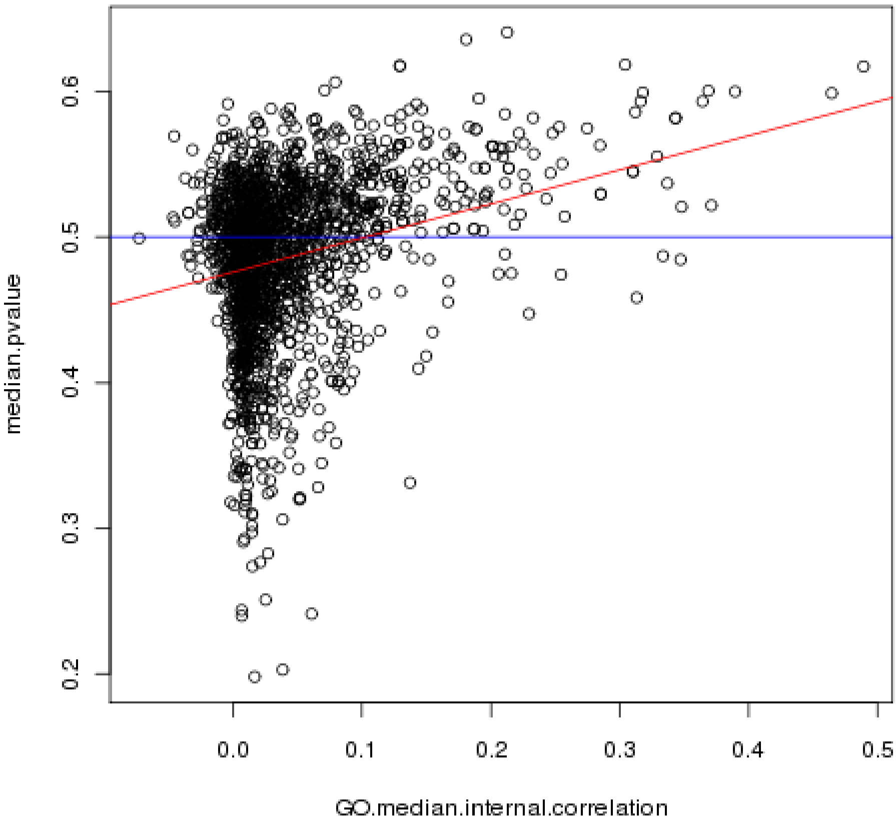
Two t-test statistics where computed for each data set, one comparing cases to controls within condition A and the second one comparing cases to controls within condition B. Hence, we simulated 500 datasets of bidimensional ranking statistics.

As no consistent biological differences are expected in the simulated data sets, no gene is expected to be differentially expressed and no gene set is expected to be enriched in any of the comparisons. However, as the data come form real biological samples, the true correlation structure is kept in all 500 datasets and hence should be dragged to the computed t-statistics.

Real GO Biological Process annotation of the genes in the data set was collected form Ensembl. and filtered by size (as represented within the 10866 transcripts available in the data) 1870 Biological Process of sizes between 10 and 500 where kept for this study. Internal correlation of each GO Biological Process term was estimated by the median correlation between all pairs of genes within the GO term.

All 1870 GO terms where tested over the 500 simulated ranking statistics using our bivariate logistic model. Then, for each GO term p-values of the interaction where summarized using the median values across the 500 simulations. Therefore we obtained a median p-value estimate for each Biological Process.

Median p-value of the interaction term is plotted against median internal correlation for each of the 1870 GO terms in the following graph.



It can be appreciated in the graph first, the already described bias of p-values being smaller than it would be expected in a random experiment with no enriched Gene Sets. Second, and not so much expected, that the bias in p-values decreases as the internal correlation of the GO terms increases. This is just the opposite pattern described in Goeman (2007) and is most probably due to the fact that in general the interaction term is a correction of the estimated main effects. In general higher p-values in the interaction term correspond to lower p-values of the main effects an those will be associated to higher internal correlation of the GO term, as described in Goeman (2007).

Appendix C

Multidimensional Gene Set R package manual.

In this appendix you can find the help pages from the R package implementing the algorithms in **Montaner** and Dopazo 2010.

Multi-Dimensional Gene Set Analysis

David Montaner

February 9, 2010

Lets first load the library and the data

```
> library(mdgsa)
> data(breast)
```

Now we have a couple of objects in our session:

```
> ls()
```

```
[1] "annot"   "ranking"
```

For each gene, the ranking statistics according to **prognosis** and **differential expression**

```
> head(ranking)
```

```
        dif.exp prognosis
MKI67   -11.545     1.242
CENPE   -11.451     1.036
COL10A1 -11.370     1.055
CKS2    -11.303     1.174
RACGAP1 -11.229     1.651
CD97    -10.542     1.095
```

and the Gene Ontology annotation of those genes for the Biological Process

```
> head(annot)
```

```
    gene         GO
1   LIG4 GO:0000726
2   MLH1 GO:0000726
3 MRE11A GO:0000726
4  NHEJ1 GO:0000726
5  PRKDC GO:0000726
6  UBE2N GO:0000726
```

Just 13 Gen Sets (GOs) are included in this example.

```
> length(unique(annot[, 2]))
```

```
[1] 13
```

To run a Multi-Dimensional Gene Set Analysis for this data we use the function mdGsa.

```
> res <- mdGsa(ranking, annot)
```

And generally we want to standardize the results using standardizeMdGsa.

```
> res <- standardizeMdGsa(res, rankstat = ranking)
```

We get a data.frame with the following columns:

```
> colnames(res)
```

```
 [1] "size"          "conv"          "error"         "LOR.dif.exp"
 [5] "LOR.prognosis" "LOR.I"         "sd.dif.exp"    "sd.prognosis"
 [9] "sd.I"          "z.dif.exp"     "z.prognosis"   "z.I"
[13] "p.dif.exp"     "p.prognosis"   "p.I"           "adj.dif.exp"
[17] "adj.prognosis" "adj.I"
```

and a row for each of the Gene Sets tested.

```
> round(res[, c("LOR.dif.exp", "LOR.prognosis", "LOR.I", "adj.dif.exp",
+     "adj.prognosis", "adj.I")], 3)
```

```
           LOR.dif.exp LOR.prognosis  LOR.I adj.dif.exp adj.prognosis adj.I
GO:0000726      -0.348         0.110 -0.097       1.000         1.000 0.997
GO:0003015       0.089         0.096  0.025       1.000         0.991 0.997
GO:0006414       0.452        -1.120 -0.435       0.000         0.000 0.019
GO:0009888       0.162         0.005  0.046       1.000         1.000 0.997
GO:0015698      -0.097         0.116 -0.028       1.000         0.991 0.997
GO:0016525       0.129         0.129  0.038       1.000         0.991 0.997
GO:0042110       0.029        -0.383  0.145       0.618         0.001 0.606
GO:0043408       0.049         0.051  0.012       1.000         1.000 0.997
GO:0046330       0.155         0.000  0.041       1.000         1.000 0.997
GO:0048729       0.121        -0.005  0.035       1.000         1.000 0.997
GO:0050821      -0.003        -0.157  0.001       1.000         0.991 0.997
GO:0051301      -0.385        -0.003 -0.382       0.000         1.000 0.000
GO:0060047       0.089         0.096  0.025       1.000         0.991 0.997
```

The function classifyMdGsaPattern help us to classify the pattern of each Gene Set in the bidimensional space of differential expression and prognosis.

```
> pat <- classifyMdGsaPattern(res, cutoff = 0.01)
> pat
```

```
GO:0000726 GO:0003015 GO:0006414 GO:0009888 GO:0015698 GO:0016525 GO:0042110
      "NS"       "NS"      "q4f"       "NS"       "NS"       "NS"       "yl"
GO:0043408 GO:0046330 GO:0048729 GO:0050821 GO:0051301 GO:0060047
      "NS"       "NS"       "NS"       "NS"      "q2i"       "NS"
```

So we get that:

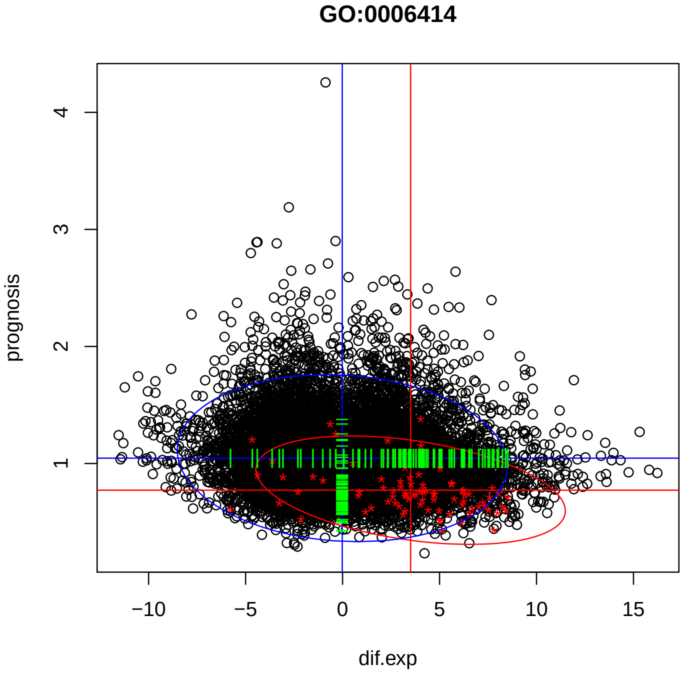
```
> cbind(round(res[, c("LOR.dif.exp", "LOR.prognosis", "LOR.I",
+     "adj.dif.exp", "adj.prognosis", "adj.I")], 3), pat)
```

```
           LOR.dif.exp LOR.prognosis  LOR.I adj.dif.exp adj.prognosis adj.I pat
GO:0000726      -0.348         0.110 -0.097       1.000         1.000 0.997  NS
GO:0003015       0.089         0.096  0.025       1.000         0.991 0.997  NS
GO:0006414       0.452        -1.120 -0.435       0.000         0.000 0.019 q4f
GO:0009888       0.162         0.005  0.046       1.000         1.000 0.997  NS
GO:0015698      -0.097         0.116 -0.028       1.000         0.991 0.997  NS
GO:0016525       0.129         0.129  0.038       1.000         0.991 0.997  NS
GO:0042110       0.029        -0.383  0.145       0.618         0.001 0.606  yl
GO:0043408       0.049         0.051  0.012       1.000         1.000 0.997  NS
GO:0046330       0.155         0.000  0.041       1.000         1.000 0.997  NS
GO:0048729       0.121        -0.005  0.035       1.000         1.000 0.997  NS
GO:0050821      -0.003        -0.157  0.001       1.000         0.991 0.997  NS
GO:0051301      -0.385        -0.003 -0.382       0.000         1.000 0.000 q2i
GO:0060047       0.089         0.096  0.025       1.000         0.991 0.997  NS
```

and hence, ”GO:0006414” is enriched in both, differential expression and prognosis in an independent way (q4f).

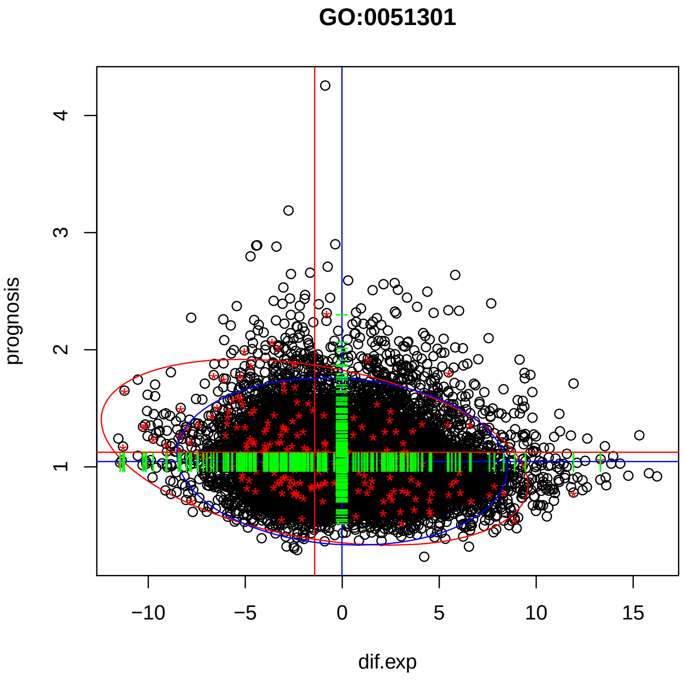
We can get a graphical display using the function plotMdGsa.

```
> plotMdGsa("GO:0006414", ranking, annot)
```



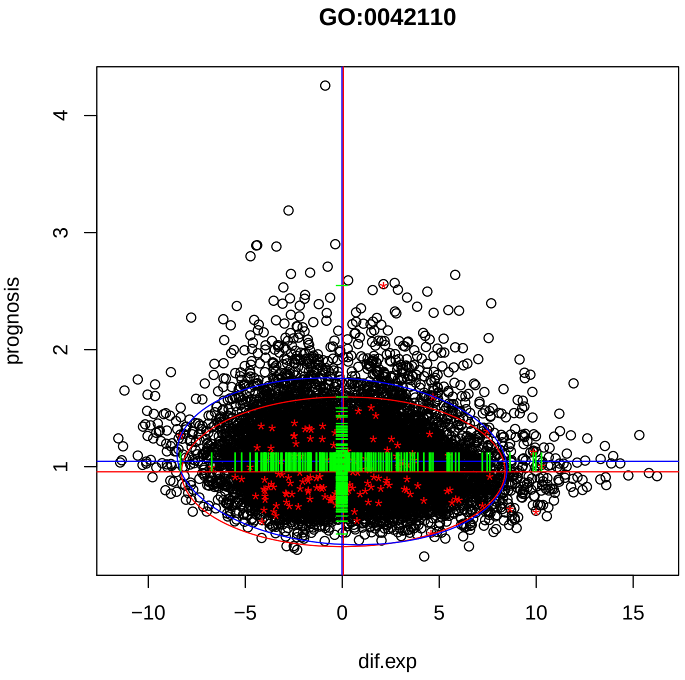
The GO term ”GO:0051301” shows a q2i pattern meaning that it is enriched in both differential expression and prognosis but strongly than the univariate analyses will show.

```
> plotMdGsa("GO:0051301", ranking, annot)
```



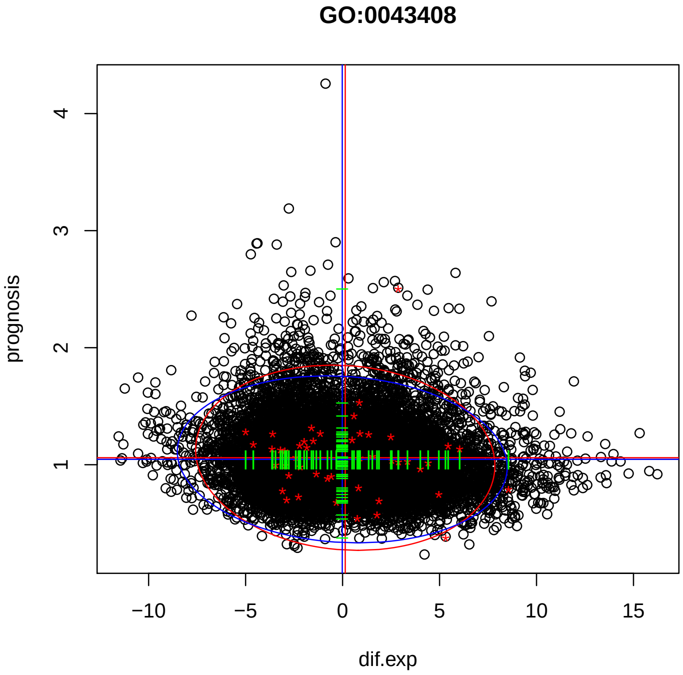
The GO term ”GO:0042110” is just enriched in the prognosis dimension (yl).

```
> plotMdGsa("GO:0042110", ranking, annot)
```



And ”GO:0043408” is not enriched in any of the effects (NS).

```
> plotMdGsa("GO:0043408", ranking, annot)
```



Bibliography

A. Agresti. *Categorical Data Analysis.* Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd edition, 2002. 24, 60

F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, Mar 2004. [DOI:10.1093/bioinformatics/btg455] [PubMed:14990455]. 4, 10, 52, 59

F. Al-Shahrour, P. Minguez, J. Tarraga, **D. Montaner**, E. Alloza, J. M. Vaquerizas, L. Conde, C. Blaschke, J. Vera, and J. Dopazo. BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res.*, 34(Web Server issue):W472–476, Jul 2006. [PubMed Central:PMC1538844] [DOI:10.1093/nar/gkl172] [PubMed:16845052]. 11, 12

F. Al-Shahrour, L. Arbiza, H. Dopazo, J. Huerta-Cepas, P. Minguez, **D. Montaner**, and J. Dopazo. From genes to functional classes in the study of biological systems. *BMC Bioinformatics*, 8:114, 2007a. [PubMed Central:PMC1853114] [DOI:10.1186/1471-2105-8-114] [PubMed:17407596]. 54, 55, 59

F. Al-Shahrour, P. Minguez, J. Tarraga, I. Medina, E. Alloza, **D. Montaner**, and J. Dopazo. FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res.*, 35(Web Server issue):W91–96, Jul 2007b. [PubMed Central:PMC1933151] [DOI:10.1093/nar/gkm260] [PubMed:17478504]. 4, 10, 52, 59

F. Al-Shahrour, J. Carbonell, P. Minguez, S. Goetz, A. Conesa, J. Tarraga, I. Medina, E. Alloza, **D. Montaner**, and J. Dopazo. Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments. *Nucleic Acids Res.*, 36(Web Server issue):W341–346, Jul 2008. [PubMed Central:PMC2447758] [DOI:10.1093/nar/gkn318] [PubMed:18515841]. 12

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29, May 2000. 11, 12, 52

Yoav Benjamini and Yosef Hochberg. 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, 1995. ISSN 00359246. [DOI:10.2307/2346101]. 3, 19, 52

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):pp. 1165–1188, 2001. ISSN 00905364. [DOI:10.1214/aos/1013699998]. 19, 52

L.L. Bonilla. *Progress in Industrial Mathematics at ECMI 2006*. The European Consortium for Mathematics in Industry. Springer London, Limited, 2008. ISBN 9783540719922. URL http://books.google.es/books?id=tlLkhSNROOEC. 99

M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines.

Proc. Natl. Acad. Sci. U.S.A., 97(1):262–267, Jan 2000. [PubMed Central:PMC26651] [PubMed:10618406]. 61

A. Carvajal-Rodriguez and J. de Una-Alvarez. Assessing significance in high-throughput experiments by sequential goodness of fit and q-value estimation. *PLoS ONE*, 6(9):e24700, 2011. [PubMed Central:PMC3170371] [DOI:10.1371/journal.pone.0024700] [PubMed:21931819]. 3, 52

G. M. Church. Genomes for all. *Sci. Am.*, 294(1):46–54, Jan 2006. [PubMed:16468433]. 3

L. Conde, J. M. Vaquerizas, J. Santoyo, F. Al-Shahrour, S. Ruiz-Llorente, M. Robledo, and J. Dopazo. PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res.*, 32(Web Server issue):W242–248, Jul 2004. [PubMed Central:PMC441576] [DOI:10.1093/nar/gkh438] [PubMed:15215388]. 9

L. Conde, J. M. Vaquerizas, H. Dopazo, L. Arbiza, J. Reumers, F. Rousseau, J. Schymkowitz, and J. Dopazo. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Res.*, 34(Web Server issue):W621–625, Jul 2006. [PubMed Central:PMC1538854] [DOI:10.1093/nar/gkl071] [PubMed:16845085]. 9

L. Conde, **D. Montaner**, J. Burguet-Castell, J. Tarraga, I. Medina, F. Al-Shahrour, and J. Dopazo. ISACGH: a web-based environment for the analysis of Array CGH and gene expression which includes functional profiling. *Nucleic Acids Res.*, 35(Web Server issue):W81–85, Jul 2007. [PubMed Central:PMC1933149] [DOI:10.1093/nar/gkm257] [PubMed:17468499]. 9

I. Dinu, J. D. Potter, T. Mueller, Q. Liu, A. J. Adewale, G. S. Jhangri, G. Einecke, K. S. Famulski, P. Halloran, and Y. Yasui. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, 8:

242, 2007. [PubMed Central:PMC1931607] [DOI:10.1186/1471-2105-8-242] [PubMed:17612399]. 55

I. Dinu, J. D. Potter, T. Mueller, Q. Liu, A. J. Adewale, G. S. Jhangri, G. Einecke, K. S. Famulski, P. Halloran, and Y. Yasui. Gene-set analysis and reduction. *Brief. Bioinformatics*, 10(1):24–34, Jan 2009. [PubMed Central:PMC2638622] [DOI:10.1093/bib/bbn042] [PubMed:18836208]. 37, 55

J. Dopazo. Formulating and testing hypotheses in functional genomics. *Artif Intell Med*, 45(2-3):97–107, 2009. [DOI:10.1016/j.artmed.2008.08.003] [PubMed:18789659]. 3, 20, 53

J. Dopazo. *Functional Profiling Methods in Cancer*, volume 576 of *Methods in Molecular Biology*, pages 363–374. Humana Press, 2010. ISBN 978-1-934115-76-3. [DOI:10.1007/978-1-59745-545-9_19]. 18, 19

B. L. Fridley, G. D. Jenkins, and J. M. Biernacka. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS ONE*, 5(9), 2010. [PubMed Central:PMC2941449] [DOI:10.1371/journal.pone.0012693] [PubMed:20862301]. 56

F. Garcia-Alcalde, F. Garcia-Lopez, J. Dopazo, and A. Conesa. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics*, 27(1):137–139, Jan 2011. [PubMed Central:PMC3008637] [DOI:10.1093/bioinformatics/btq594] [PubMed:21098431]. 10

A. M. Glas, A. Floore, L. J. Delahaye, A. T. Witteveen, R. C. Pover, N. Bakx, J. S. Lahti-Domenici, T. J. Bruinsma, M. O. Warmoes, R. Bernards, L. F. Wessels, and L. J. Van't Veer. Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics*, 7:278, 2006. [PubMed Central:PMC1636049] [DOI:10.1186/1471-2164-7-278] [PubMed:17074082]. 3

J. J. Goeman and P. Buhlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, Apr 2007. [DOI:10.1093/bioinformatics/btm051] [PubMed:17303618]. 54, 55, 58

J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, Jan 2004. [PubMed:14693814]. 54, 55, 56

J. J. Goeman, J. Oosting, A. M. Cleton-Jansen, J. K. Anninga, and H. C. van Houwelingen. Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21(9):1950–1957, May 2005. [DOI:10.1093/bioinformatics/bti267] [PubMed:15657105]. 56

N. Hall. Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.*, 210(Pt 9):1518–1525, May 2007. [DOI:10.1242/jeb.001370] [PubMed:17449817]. 3

J. Herrero, F. Al-Shahrour, R. Diaz-Uriarte, A. Mateos, J. M. Vaquerizas, J. Santoyo, and J. Dopazo. GEPAS: A web-based resource for microarray gene expression data analysis. *Nucleic Acids Res.*, 31(13):3461–3467, Jul 2003. [PubMed Central:PMC168997] [PubMed:12824345]. ix, 4, 5

J. Herrero, J. M. Vaquerizas, F. Al-Shahrour, L. Conde, A. Mateos, J. S. Diaz-Uriarte, and J. Dopazo. New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res.*, 32(Web Server issue):W485–491, Jul 2004. [PubMed Central:PMC441559] [DOI:10.1093/nar/gkh421] [PubMed:15215434]. ix, 5

J. Huerta-Cepas, J. Dopazo, and T. Gabaldon. ETE: a python Environment for Tree Exploration. *BMC Bioinformatics*, 11:24, 2010. [PubMed Central:PMC2820433] [DOI:10.1186/1471-2105-11-24] [PubMed:20070885]. 8

R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, Apr 2003. [DOI:10.1093/biostatistics/4.2.249] [PubMed:12925520]. 4

E. Jantus Lewintre, C. Reinoso Martin, **D. Montaner**, M. Marin, M. Jose Terol, R. Farras, I. Benet, J. J. Calvete, J. Dopazo, and J. Garcia-Conde. Analysis of chronic lymphotic leukemia transcriptomic profile: differences between molecular subgroups. *Leuk. Lymphoma*, 50(1):68–79, Jan 2009. [DOI:10.1080/10428190802541807] [PubMed:19127482]. 9

G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, 33(Database issue):D428–432, Jan 2005. [PubMed Central:PMC540026] [DOI:10.1093/nar/gki072] [PubMed:15608231]. 13

M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, Jan 2000. [PubMed Central:PMC102409] [PubMed:10592173]. 11, 52

M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, 32 (Database issue):D277–280, Jan 2004. 13

P. Khatri, M. Sirota, and A. J. Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, 8(2):e1002375, Feb 2012. [PubMed Central:PMC3285573] [DOI:10.1371/journal.pcbi.1002375] [PubMed:22383865]. 52, 54

S. M. Lin, P. Du, W. Huber, and W. A. Kibbe. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic*

Acids Res., 36(2):e11, Feb 2008. [PubMed Central:PMC2241869] [DOI:10.1093/nar/gkm1075] [PubMed:18178591]. 4

D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, 14(13):1675–1680, Dec 1996. [DOI:10.1038/nbt1296-1675] [PubMed:9634850]. 2

U. Mansmann and R. Meister. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf Med*, 44(3):449–453, 2005. [DOI:10.1267/METH05030449] [PubMed:16113772]. 55

K. K. Mantripragada, P. G. Buckley, T. Diaz de Stahl, and J. P. Dumanski. Genomic microarrays in the spotlight. *Trends Genet.*, 20(2): 87–94, Feb 2004. [DOI:10.1016/j.tig.2003.12.008] [PubMed:14746990]. 9

A. Mateos, J. Dopazo, R. Jansen, Y. Tu, M. Gerstein, and G. Stolovitzky. Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Res.*, 12(11):1703–1715, Nov 2002. 61

McKusick-Nathans. Online mendelian inheritance in man, OMIM®. URL `http://omim.org/`. 13

I. Medina, **D. Montaner**, J. Tarraga, and J. Dopazo. Prophet, a web-based tool for class prediction using microarray data. *Bioinformatics*, 23(3):390–391, Feb 2007. [DOI:10.1093/bioinformatics/btl602] [PubMed:17138587]. 4

I. Medina, **D. Montaner**, N. Bonifaci, M. A. Pujana, J. Carbonell, J. Tarraga, F. Al-Shahrour, and J. Dopazo. Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic Acids Res.*, 37(Web

Server issue):W340–344, Jul 2009. [PubMed Central:PMC2703970] [DOI:10.1093/nar/gkp481] [PubMed:19502494]. 10

I. Medina, J. Carbonell, L. Pulido, S. C. Madeira, S. Goetz, A. Conesa, J. Tarraga, A. Pascual-Montano, R. Nogales-Cadenas, J. Santoyo, F. Garcia, M. Marba, **D. Montaner**, and J. Dopazo. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res.*, 38(Web Server issue):W210–213, Jul 2010. [PubMed Central:PMC2896184] [DOI:10.1093/nar/gkq388] [PubMed:20478823]. 12, 78

P. Minguez, S. Gotz, **D. Montaner**, F. Al-Shahrour, and J. Dopazo. SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. *Nucleic Acids Res.*, 37(Web Server issue):W109–114, Jul 2009. [PubMed Central:PMC2703972] [DOI:10.1093/nar/gkp402] [PubMed:19454602]. 10

C. Montero-Conde, J. M. Martin-Campos, E. Lerma, G. Gimenez, J. L. Martinez-Guitarte, N. Combalia, **D. Montaner**, X. Matias-Guiu, J. Dopazo, A. de Leiva, M. Robledo, and D. Mauricio. Molecular profiling related to poor prognosis in thyroid carcinoma. Combining gene expression data and biological information. *Oncogene*, 27(11):1554–1561, Mar 2008. [DOI:10.1038/sj.onc.1210792] [PubMed:17873908]. 9

V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, 34(3):267–273, Jul 2003. xiii, 10, 21, 54

E. R. Morrissey and R. Diaz-Uriarte. Pomelo II: finding differentially expressed genes. *Nucleic Acids Res.*, 37(Web Server issue):W581–586, Jul 2009. [PubMed Central:PMC2703955] [DOI:10.1093/nar/gkp366] [PubMed:19435879]. 6

N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Buillard, L. Cerutti, R. Copley, E. Courcelle, U. Das, L. Daugherty, M. Dibley, R. Finn, W. Fleischmann, J. Gough, D. Haft, N. Hulo, S. Hunter, D. Kahn, A. Kanapin, A. Kejariwal, A. Labarga, P. S. Langendijk-Genevaux, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, A. N. Nikolskaya, S. Orchard, C. Orengo, R. Petryszak, J. D. Selengut, C. J. Sigrist, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. New developments in the InterPro database. *Nucleic Acids Res.*, 35(Database issue):D224–228, Jan 2007. [PubMed Central:PMC1899100] [DOI:10.1093/nar/gkl841] [PubMed:17202162]. 52

M. A. Sartor, G. D. Leikauf, and M. Medvedovic. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25(2):211–217, Jan 2009. [PubMed Central:PMC2639007] [DOI:10.1093/bioinformatics/btn592] [PubMed:19038984]. 55, 56

M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. U.S.A.*, 93(20):10614–10619, Oct 1996. [PubMed Central:PMC38202] [PubMed:8855227]. 2

L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner,

R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T. M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X. H. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q. Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, 24(9):1151–1161, Sep 2006. [PubMed Central:PMC3272078] [DOI:10.1038/nbt1239] [PubMed:16964229]. 3

L. Shi, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, Z. Su, T. M. Chu, F. M. Goodsaid, L. Pusztai, J. D. Shaughnessy, A. Oberthuer, R. S. Thomas, R. S. Paules, M. Fielden, B. Barlogie, W. Chen, P. Du, M. Fischer, C. Furlanello, B. D. Gallas, X. Ge, D. B. Megherbi, W. F. Symmans, M. D. Wang, J. Zhang, H. Bitter, B. Brors, P. R. Bushel, M. Bylesjo, M. Chen, J. Cheng, J. Cheng, J. Chou, T. S. Davison, M. Delorenzi, Y. Deng, V. Devanarayan, D. J. Dix, J. Dopazo, K. C. Dorff, F. Elloumi, J. Fan, S. Fan, X. Fan, H. Fang, N. Gonzaludo, K. R. Hess, H. Hong, J. Huan, R. A. Irizarry, R. Judson, D. Juraeva, S. Lababidi, C. G. Lambert, L. Li, Y. Li, Z. Li, S. M. Lin, G. Liu, E. K. Lobenhofer, J. Luo, W. Luo, M. N. McCall, Y. Nikolsky, G. A. Pennello, R. G. Perkins, R. Philip, V. Popovici,

N. D. Price, F. Qian, A. Scherer, T. Shi, W. Shi, J. Sung, D. Thierry-Mieg, J. Thierry-Mieg, V. Thodima, J. Trygg, L. Vishnuvajjala, S. J. Wang, J. Wu, Y. Wu, Q. Xie, W. A. Yousef, L. Zhang, X. Zhang, S. Zhong, Y. Zhou, S. Zhu, D. Arasappan, W. Bao, A. B. Lucas, F. Berthold, R. J. Brennan, A. Buness, J. G. Catalano, C. Chang, R. Chen, Y. Cheng, J. Cui, W. Czika, F. Demichelis, X. Deng, D. Dosymbekov, R. Eils, Y. Feng, J. Fostel, S. Fulmer-Smentek, J. C. Fuscoe, L. Gatto, W. Ge, D. R. Goldstein, L. Guo, D. N. Halbert, J. Han, S. C. Harris, C. Hatzis, D. Herman, J. Huang, R. V. Jensen, R. Jiang, C. D. Johnson, G. Jurman, Y. Kahlert, S. A. Khuder, M. Kohl, J. Li, L. Li, M. Li, Q. Z. Li, S. Li, Z. Li, J. Liu, Y. Liu, Z. Liu, L. Meng, M. Madera, F. Martinez-Murillo, I. Medina, J. Meehan, K. Miclaus, R. A. Moffitt, **D. Montaner**, P. Mukherjee, G. J. Mulligan, P. Neville, T. Nikolskaya, B. Ning, G. P. Page, J. Parker, R. M. Parry, X. Peng, R. L. Peterson, J. H. Phan, B. Quanz, Y. Ren, S. Riccadonna, A. H. Roter, F. W. Samuelson, M. M. Schumacher, J. D. Shambaugh, Q. Shi, R. Shippy, S. Si, A. Smalter, C. Sotiriou, M. Soukup, F. Staedtler, G. Steiner, T. H. Stokes, Q. Sun, P. Y. Tan, R. Tang, Z. Tezak, B. Thorn, M. Tsyganova, Y. Turpaz, S. C. Vega, R. Visintainer, J. von Frese, C. Wang, E. Wang, J. Wang, W. Wang, F. Westermann, J. C. Willey, M. Woods, S. Wu, N. Xiao, J. Xu, L. Xu, L. Yang, X. Zeng, J. Zhang, L. Zhang, M. Zhang, C. Zhao, R. K. Puri, U. Scherf, W. Tong, and R. D. Wolfinger. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.*, 28(8):827–838, Aug 2010. [PubMed Central:PMC3315840] [DOI:10.1038/nbt.1665] [PubMed:20676074]. 3

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 102(43):15545–15550, Oct 2005. [PubMed Central:PMC1239896] [DOI:10.1073/pnas.0506580102]

[PubMed:16199517]. xiii, 10, 21, 55

J. Tarraga, I. Medina, J. Carbonell, J. Huerta-Cepas, P. Minguez, E. Alloza, F. Al-Shahrour, S. Vegas-Azcarate, S. Goetz, P. Escobar, F. Garcia-Garcia, A. Conesa, **D. Montaner**, and J. Dopazo. GEPAS, a web-based tool for microarray data analysis and interpretation. *Nucleic Acids Res.*, 36(Web Server issue):W308–314, Jul 2008. [PubMed Central:PMC2447723] [DOI:10.1093/nar/gkn303] [PubMed:18508806]. 12

D. Montaner and J. Dopazo. Multidimensional gene set analysis of genomic data. *PLoS ONE*, 5(4):e10348, 2010. [PubMed Central:PMC2860497] [DOI:10.1371/journal.pone.0010348] [PubMed:20436964]. 1, 2, 12, 25, 30, 32, 40, 41, 51, 55, 56, 77, 78, 105, 109, 132

D. Montaner, J. Tarraga, J. Huerta-Cepas, J. Burguet, J. M. Vaquerizas, L. Conde, P. Minguez, J. Vera, S. Mukherjee, J. Valls, M. A. Pujana, E. Alloza, J. Herrero, F. Al-Shahrour, and J. Dopazo. Next station in microarray data analysis: GEPAS. *Nucleic Acids Res.*, 34(Web Server issue):W486–491, Jul 2006. [PubMed Central:PMC1538867] [DOI:10.1093/nar/gkl197] [PubMed:16845056]. ix, 1, 2, 6, 8, 9, 11, 12, 43, 44, 132

D. Montaner, F. Al-Shahrour, and J. Dopazo. New trends in the analysis of functional genomic data. In L.L. Bonilla, M. Moscoso, G. Platero, and J.M. Vega, editors, *Progress in Industrial Mathematics at ECMI 2006*, volume 12 of *Mathematics in Industry*, pages 576–580. Springer, 2008. ISBN 978-3-540-71991-5. [DOI:10.1007/978-3-540-71992-2_94]. 51

D. Montaner, P. Minguez, F. Al-Shahrour, and J. Dopazo. Gene set internal coherence in the context of functional profiling. *BMC Genomics*, 10:197, 2009. [PubMed Central:PMC2680416] [DOI:10.1186/1471-2164-10-197] [PubMed:19397819]. 1, 2, 12, 17, 25, 30, 32, 34, 36, 37, 51, 55, 56, 61, 62, 132

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, Jun 2001. [PubMed:11395428]. 4

J. Valls, M. Grau, X. Sole, P. Hernandez, **D. Montaner**, J. Dopazo, M. A. Peinado, G. Capella, V. Moreno, and M. A. Pujana. CLEAR-test: combining inference for differential expression and variability in microarray data analysis. *J Biomed Inform*, 41(1):33–45, Feb 2008. [DOI:10.1016/j.jbi.2007.05.005] [PubMed:17597009]. 7

J. H. van Heerden, A. Conesa, D. J. Stein, **D. Montaner**, V. Russell, and N. Illing. Parallel changes in gene expression in peripheral blood mononuclear cells and the brain after maternal separation in the mouse. *BMC Res Notes*, 2:195, 2009. [PubMed Central:PMC2759952] [DOI:10.1186/1756-0500-2-195] [PubMed:19781058]. 9

J. M. Vaquerizas, J. Dopazo, and R. Diaz-Uriarte. DNMAD: web-based diagnosis and normalization for microarray data. *Bioinformatics*, 20(18):3656–3658, Dec 2004. [DOI:10.1093/bioinformatics/bth401] [PubMed:15247094]. 6

J. M. Vaquerizas, L. Conde, P. Yankilevich, A. Cabezon, P. Minguez, R. Diaz-Uriarte, F. Al-Shahrour, J. Herrero, and J. Dopazo. GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. *Nucleic Acids Res.*, 33(Web Server issue):W616–620, Jul 2005. [PubMed Central:PMC1160260] [DOI:10.1093/nar/gki500] [PubMed:15980548]. ix, 5, 6

I. Vastrik, P. D'Eustachio, E. Schmidt, G. Joshi-Tope, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, L. Matthews, G. Wu, E. Birney, and L. Stein. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, 8(3):R39, 2007. [PubMed Central:PMC1868929] [DOI:10.1186/gb-2007-8-3-r39] [PubMed:17367534]. 52

Resumen en Castellano

Objetivos

En cualquier contexto de análisis de datos experimentales hay un primer paso en el que los datos numéricos son procesados en sentido estadístico. En esta etapa, la información se resume en unos pocos índices o valores numéricos que pueden ser luego fácilmente interpretados por los investigadores. Generalmente, este segundo paso de interpretación de resultados y extracción de conclusiones no es técnicamente complicado ni requiere de herramientas específicas para ser llevado a cabo. La interpretación de los estadísticos, p-valores o demás índices que resultan del análisis estadístico se realiza de forma casi intuitiva y la extracción de conclusiones se realiza en función de lo que el investigador *sabe (de memoria)* acerca del proceso estudiado. En los contextos de análisis de datos genómicos, el primer paso del análisis estadístico no cambia, aunque el volumen de datos suele ser considerablemente mayor que en un estudio biológico o genético convencional. Esto complica su manejo en términos informáticos, pero sin embargo, conceptualmente, el análisis de los datos genómicos experimentales es igual al de los datos genéticos[1].

Como contrapartida, la segunda etapa de interpretación de los resultados estadísticos cambia radicalmente en el contexto del análisis de datos genómicos. En este nuevo paradigma no hay uno o unos pocos valo-

[1] Usamos aquí el término *genético* para referirnos a los estudios que analizan uno o unos pocos genes y el término *genómico* para los estudios que involucran medidas de todos los genes del genoma de la especie bajo observación.

res numéricos de resumen de la información sino, generalmente, uno para cada gen, lo que significa miles de estadísticos, p-valores o simplemente índices numéricos que deben ser interpretados a la hora de extraer conclusiones del estudio. En este nuevo escenario la intuición del investigador no vale, y es frecuente la extracción de conclusiones sesgadas cuando la interpretación de los resultados se se hace simplemente en función de lo que el investigador *sabe*. Si se quiere evitar este tipo de sesgos, la interpretación de los resultados de experimentos genómicos debe hacerse de forma sistemática, utilizando herramientas computacionales. Además de esto, ningún investigador puede manejar de memoria la gran cantidad de información que se tiene en la actualidad sobre el genoma y que está almacenada en cientos de bases de datos dispersas en la web. Son necesarias herramientas bioinformáticas para recopilar esta información que representa el conocimiento biológico establecido y que debe ser superpuesta a los resultados experimentales para su correcta interpretación así como para la extracción de conclusiones.

En la jerga del análisis de datos genómicos se identifica como métodos de *análisis funcional* a todos aquellos algoritmos encaminados a la combinación de la información puramente experimental, y por lo tanto nueva para el investigador, con la información ya disponible y validada que existe en las bases de datos biológicos, es decir, el conocimiento ya establecido por la comunidad científica.

Las ventajas de combinar la información disponible en bases de datos con la información experimental son claras: por una parte el *análisis funcional* permite incrementar la cantidad de información incluida en la discusión del experimento. No sólo la información nueva se tiene en cuenta sino que se pone en el contexto de lo ya conocido sobre el proceso biológico estudiado. Además, el *análisis funcional* facilita el resumen sistemático de la información recogida. Las miles de medidas registradas para los genes se agrupan en bloques de funcionalidad biológica reconocida a priori. Este resumen facilita directamente la interpretación de los resultados ya que, disertar sobre funciones o procesos biológicos es, en sí mismo, más informativo que discutir sobre genes aislados. Por último, el *análisis funcional* combina la evidencia experimental de los genes anotados conjuntamente en las bases de datos, consiguiendo así un efecto estadístico similar al que resultaría de incrementar el tamaño muestral del experimento.

En los últimos años se han realizado esfuerzos considerables en el campo de la bioinformática con el objetivo de establecer metodologías apropiadas de *análisis funcional*. A pesar de ello, la rápida evolución de las tecnologías de toma de medidas genómicas hace que, por el momento, el *análisis funcional* sea un campo abierto de la investigación y el desarrollo bioinformático. En esta tesis proponemos y desarrollamos algoritmos de *análisis funcional* que abordan problemas no tratados con anterioridad, pero cuya resolución es indispensable para el correcto análisis de los experimentos genómicos más avanzados. Entre otros, proponemos soluciones para el análisis combinado de varias características genómicas. También desarrollamos la posibilidad de incluir pesos asociados a los genes para matizar la relevancia de cada uno de ellos en el propio *análisis funcional*, haciéndolo más adecuado al contexto biológico estudiado.

La mayor parte de los algoritmos de *análisis funcional* se desarrollaron en el contexto del análisis de expresión diferencial. Posteriormente fueron surgiendo modificaciones metodológicas que permitían aplicarlos en estudios de variabilidad genética, basadas por ejemplo en SNPs, datos de alteraciones en el número de copias de los genes, etcétera. Uno de los principales desarrollos de esta tesis es el de proponer una metodología general que funciona independientemente del tipo de estudio genético al que se aplica; sin importar incluso del tipo de *test* estadístico utilizado en el estudio de los datos experimentales a nivel de gen. Nuestro algoritmo puede por ejemplo aplicarse en cualquier tipo de contexto de expresión de genes, no sólo en el caso de expresión diferencial sino también en contextos de regresión, análisis de supervivencia, series temporales . . . Pero además, es directamente aplicable en datos de variación genómica medidos en SNPs, alteraciones del número de copias, metilación y casi en cualquier otro contexto de estudios genómicos incluyendo por ejemplo estudios evolutivos. Esta flexibilidad del método hace que sea aplicable en casos en los que otras metodologías de *análisis funcional* no pueden ser utilizadas.

Sin embargo la mayor utilidad de nuestra metodología no viene dada por el hecho de poder analizar cualquier tipo de experimento genómico sino porque nos permite además analizar varios tipos de datos genómicos a la vez. Nos permite por ejemplo analizar funcionalmente un conjunto de datos en el que se han tomado medidas de expresión y de alteraciones en el número de copias de los genes. Nos permite combinar en un único *análisis funcional* por ejemplo datos de metilación y de variantes o SNPs. De esta forma, nuestra nueva metodología permite encontrar bloques funcionales o procesos biológicos que se activan o regulan por la interacción de varias características genómicas. Podemos por ejemplo realizar comparativas caso control y detectar rutas metabólicas alteradas como consecuencia de cambios de expresión y alteraciones genómicas de forma combinada. Rutas que de otra forma pasarían desapercibidas en el *análisis funcional* convencional de cada una de las características genómicas por separado.

Metodología

La metodología propuesta para la consecución de los objetivos arriba mencionados fue desarrollada en detalle en los tres artículos que se compendian en esta tesis: **Montaner** et al. 2006, **Montaner** et al. 2009 y **Montaner** and Dopazo 2010. La idea principal es la de utilizar *métodos de regresión logística* para realizar los análisis funcionales. Estos modelos tienen gran flexibilidad y pueden ser aplicados en contextos experimentales muy diversos, permitiendo adaptar el *análisis funcional* a todo tipo de estudios genómicos.

El análisis de casi cualquier conjunto de datos genómicos devuelve como resultado un índice numérico asociado a cada gen. Este índice puede ser, por ejemplo, un p-valor que evalúa la expresión diferencial de cada gen, un estadístico que mide la asociación de la variabilidad del gen con cierto fenotipo, o una tasa evolutiva asociada a cada uno de los genes estudiados. Este índice numérico es en sí un indicador o medida de la característica biológica estudiada. En el caso paradigmático de la expresión diferencial, el p-valor nos indica cuanto aumenta la expresión del gen en casos relativa a controles, es decir, cuanto aumenta la expresión de los genes como respuesta, por ejemplo, a un determinado tratamiento. Indirectamente entonces, el p-valor mide el grado de activación o sobre expresión que el tratamiento produce en cada uno de los genes. Hasta este punto en el que se deriva el índice asociado a cada gen, toda la información que se ha utilizado en el estudio es puramente experimental. Es información nueva en el sentido de que probablemente nadie antes ha observado la expresión, por ejemplo, de los genes en ese mismo escenario. En caso contrario, seguramente, no sería necesario plantear la investigación.

El siguiente paso en el proceso es el del *análisis funcional.* Este comienza seleccionando una base de datos que aporte conocimientos relevantes para el investigador y recuperando la información asociada a los genes involucrados en el estudio. Generalmente se usa el término *anotación* para referirse a esta información. Hay cientos de bases de datos disponibles vía web en múltiples paginas repartidas por todo el mundo. La del consorcio *Gene Ontology*, por ejemplo, es una de las más utilizadas para extraer la información o anotación de los procesos biológicos en los que están involucrados los genes. La base de datos KEGG describe las rutas de señalización y procesos metabólicos en los que los distintos genes participan. En general la información extraída de estas bases de datos se estructura, de forma simplificada, como etiquetas asociadas a los genes. Un gen por ejemplo estará asociado con una determinada ruta metabólica o con un proceso biológico concreto si se sabe por estudios previos que el gen toma parte en el desarrollo de dicho metabolismo o proceso. Es precisamente debido a esta estructura de datos formada por etiquetas asociadas a los genes que, en el ámbito del análisis de datos genómicos, nos referimos a esta información recuperada de las bases de datos como *anotación* de los genes. Pero más allá de la nomenclatura, lo que nos interesa resaltar en lo referente a la información recuperada de las bases de datos es su carácter dicotómico o binario. Para cada proceso biológico descrito en las bases de datos sabemos si un gen está o no involucrado en él. Para cada ruta metabólica podemos conocer qué genes realizan el metabolismo correspondiente. Volviendo a la terminología de los análisis genómicos, dada la etiqueta de un proceso biológico podemos decir si un gen está o no anotado con ella. En sentido estadístico esta información dicotómica de la anotación o no anotación de los genes bajo cierta etiqueta se recoge en una variable binaria: un valor numérico que puede ser *uno* o *cero* y que está asociado con cada gen del genoma estudiado.

Así, después de recuperar la información disponible en las bases de datos y combinarla con el índice numérico extraído de los datos experimentales obtenemos dos valores asociados con cada gen, uno continuo y otro binario. El siguiente paso del *análisis funcional* es el de explorar la relación global que existe entre estas dos variables. Asociaciones o correlaciones entre las dos variables nos indicaran la existencia de una relación entre la información ya conocida y registrada en las bases de datos, y la información nueva, resumida en el índice resultante del análisis de los genes. En el caso de la expresión diferencial, por ejemplo, valores bajos del p-valor indican una mayor diferencia de expresión o, como se ha descrito más arriba, un mayor efecto del tratamiento sobre esos genes. Al incorporar la información de un determinado proceso biológico tendremos una segunda variable binaria que tomara valor *uno* para los genes involucrados en dicho proceso. Si al relacionar las dos variables observamos que los genes con valor *uno* en la variable de anotación tienen p-valores relativamente bajos, inferiremos que los genes del bloque tienden a estar más diferencial mente expresados entre casos y controles. Podremos entonces decir que el tratamiento no sólo provoca un incremento de la expresión de algunos genes sino que además esos genes constituyen las piezas de una determinada maquinaria biológica. Concluiremos entonces que el tratamiento debe activar el proceso biológico o la ruta metabólica correspondiente. Con esto, la interpretación de los resultados de nuestro experimento será mucho más clara puesto que ya no hablaremos de la activación de varios genes con consecuencias desconocidas sino que hablaremos de la activación de un proceso biológico que está bien descrito incluso ya antes de plantear nuestro experimento.

Los *métodos de regresión logística* se desarrollaron precisamente para estudiar la dependencia entre una variable binaria y otra variable continua. Por ello pueden ser directamente utilizados para llevar a cabo un análisis *análisis funcional* clásico. Pero además de eso, los *modelos de regresión logística* pueden incorporar no sólo una sino múltiples variables continuas y estudiar la configuración de sus valores con respecto de la variable discreta. Esto permite incluir en el *análisis funcional* no sólo los valores numéricos resumen de una única característica genómica medida experimentalmente, sino de varias. Podemos plantear entonces un experimento en el que además de las medias de expresión de los genes se tome, por ejemplo, datos de su metilación. Las diferencias de expresión se resumirán en un p-valor o índice y las diferencias en la metilación de los genes se resumirán en un segundo índice. La *regresión logística* nos permitirá entonces ver si la anotación funcional de los genes está relacionada con la expresión diferencial, con la metilación, o con ambas. Concluiremos entonces que, cambios de expresión y metilación combinados son los que activan o desactivan la maquinaria biológica correspondiente.

Por otra parte los *métodos de regresión* permiten la incorporación directa de pesos que representan la importancia de los genes en cada análisis. En este trabajo mostramos como se pueden utilizar dichos pesos para modelizar información adicional derivada de la estructura interna de cada uno de los bloques funcionales o anotaciones. Introducimos así un nuevo concepto de análisis funcional en el que la estructura interna de los bloques no es homogénea sino difusa y mostramos como este tipo de modelización puede reflejar de forma más adecuada la la biología subyacente.

Conclusiones

Después del trabajo realizado podemos concluir, de forma muy resumida que:

- Los *modelos de regresión logística* son una herramienta apropiada para realizar el *análisis funcional* de datos genómicos.
- Es viable realizar el *análisis funcional* no sólo de una característica genómica experimental sino de varias en conjunto.
- Además, este tipo de análisis nos permite descubrir características biológicas relevantes que pasarían inadvertidas en los análisis funcionales tradicionales de cada una de las *dimensiones* genómicas por separado.
- La incorporación de pesos asociados a los genes en el *análisis funcional* puede proporcionar un mejor modelado de los procesos biológicos estudiados.
- Parte de la información de la estructura interna de los bloques funcionales descritos en las bases de datos puede incorporarse de forma efectiva en los pasos del *análisis funcional*, permitiendo que el modelo de análisis refleje mejor la realidad biológica estudiada.