Multivariate methods for the integrative analysis of multi-omics datasets

TCGAome

Universitat Oberta de Catalunya
Author: Pablo Riesgo Ferreiro (priesgo@uoc.edu)
Director: Alexandre Sánchez Plà
1/13/2016

Introduction

The dramatic growth of Omics information available both in the biomedical research and in the healthcare system poses a challenge when it comes to analyze and visualize the data and consequently to obtain reliable and relevant results. The challenge has shifted from collecting enough data to providing with advanced tools to uncover the value of the available data. Furthermore, there is a strong need to translate the newly acquired biological knowledge into the clinical use and thus the integration of Omics and phenotypic information. This is well illustrated by the so called Big Data problem and its 3 V's – *volume*, *velocity* and *variety* –, extended to 4 – including *veracity* – or even to 5 – also including *value*. *Variety* and *value* are two requirements highly tied to data analysis; dealing with them needs a general improvement in our data analysis ability [1,2].

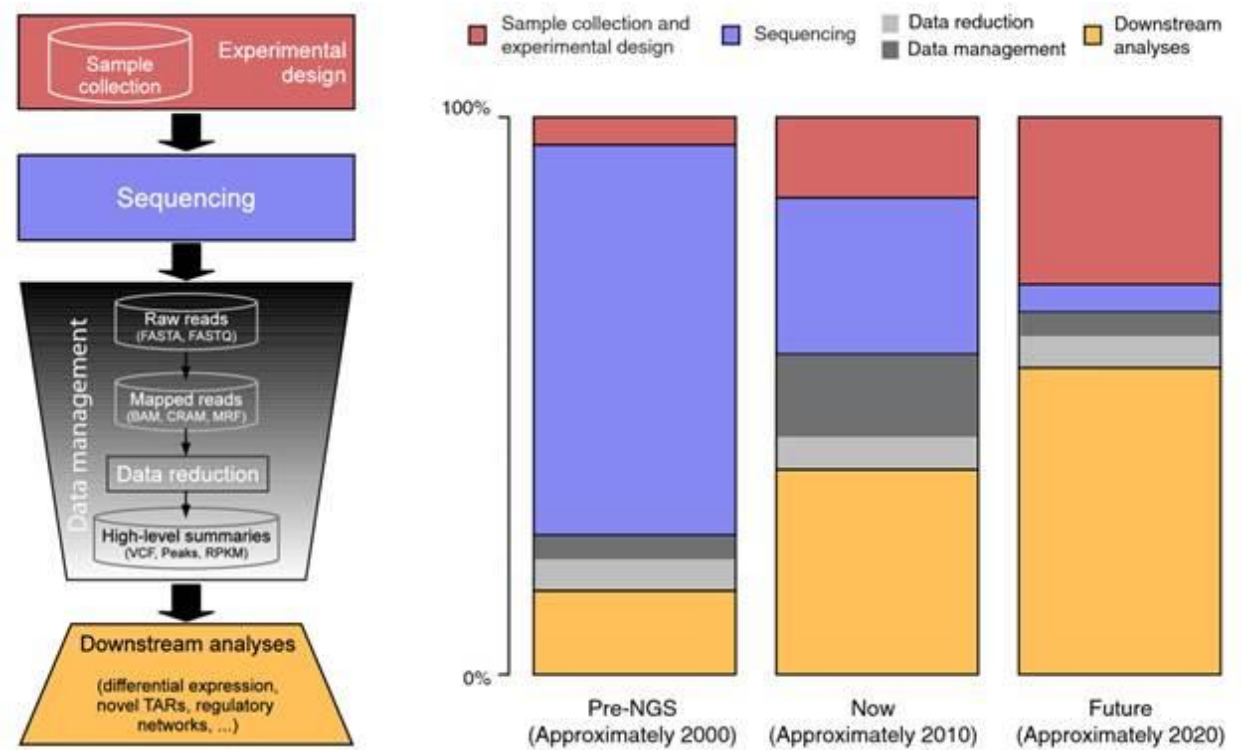


Figure 1: the cost of sequencing is decreasing at a higher rate than Moore's law, while our computational capabilities follow the mentioned law steadily, thus the cost distribution has and will change [1]

Gaining insights on cellular processes and disease etiology by discovering associations of phenotype and genotype from multiple Omics data is the main driver of biomedical data analysis. Large-scale consortia projects are making their data public like The Cancer Genome Atlas (TCGA) [3], the 1000 Genomes Project [4] and the Encyclopedia of DNA Elements Project (ENCODE) [5]; but providing tools that allow an explorative data analysis by (1) integrating heterogeneous data and (2) visualizing results in a meaningful way is a computational and analytical challenge [6]. Some of the current approaches to data integration are based in Co-Inertia Analysis (CIA) [7], Canonical Correlation Analysis (CCA) [8] and Partial Least Squares (PLS) adapted in some cases to sparse data analysis [9]. It has been acknowledged that to fully understand a biological system the prior biological knowledge will need to be included, but using this information without creating a bias in the analysis results is a concern [10].

The integration of the transcriptome, the metabolome and the proteome is of special interest for us due to three reasons: there is a known relationship between them, there is no physical interaction that require a specific modeling and they are all quantitative measurements. We will make use of the public TCGA data to test the different multivariate approaches to this problem. Unfortunately, TCGA lacks metabolomics data and hence we will be just focusing on the transcriptome and the proteome.

	RNA-Seq	ncRNA	ChIP-Seq Histone	ChIP-Seq TF	CpG DNA Methylation	DNase-Seq	Complete DNA sequencing	Exome sequencing	Proteomics	Metabolomics	Chromatin Conformation	Clinical Data	Co-morbidities	Other
RNA-Seq		29.6%	24.8%	29.6%	32.8%	16.0%	21.6%	22.4%	36.8%	21.6%	14.4%	28.0%	10.4%	0.0%
ncRNA	6.4%		8.0%	7.2%	10.4%	4.0%	6.4%	8.0%	5.6%	4.0%	1.6%	10.4%	4.0%	0.0%
ChIP-Seq Histone	6.4%	0.8%		16.0%	16.0%	11.2%	3.2%	4.8%	7.2%	4.0%	8.8%	5.6%	2.4%	0.0%
ChIP-Seq TF	6.4%	0.8%	0.8%		12.0%	16.0%	5.6%	7.2%	9.6%	4.0%	10.4%	7.2%	2.4%	0.0%
CpG DNA Methylation	11.2%	2.4%	3.2%	2.4%		8.8%	9.6%	7.2%	6.4%	4.0%	9.6%	12.0%	4.8%	0.0%
DNase-Seq	4.0%	0.8%	1.6%	2.4%	4.8%		4.0%	5.6%	4.8%	4.0%	10.4%	9.6%	2.4%	0.0%
Complete DNA sequencing	8.8%	1.6%	1.6%	1.6%	2.4%	4.0%		10.4%	13.6%	10.4%	2.4%	20.0%	5.6%	0.0%
Exome sequencing	17.6%	0.8%	1.6%	0.8%	2.4%	0.8%	6.4%		12.0%	8.8%	0.0%	20.0%	7.2%	0.0%
Proteomics	15.2%	1.6%	0.8%	0.8%	1.6%	2.4%	4.8%	8.0%		27.2%	5.6%	16.8%	5.6%	1.6%
Metabolomics	16.8%	2.4%	2.4%	1.6%	3.2%	2.4%	6.4%	4.8%	10.4%		2.4%	17.6%	6.4%	0.8%
Chromatin Conformation	0.8%	0.0%	2.4%	2.4%	0.8%	0.0%	0.8%	0.0%	0.0%	0.8%		4.0%	2.4%	0.0%
Clinical Data	31.2%	8.0%	7.2%	9.6%	15.2%	9.6%	20.0%	21.6%	16.8%	20.0%	4.0%		14.4%	3.2%
Co-morbidities	8.8%	4.0%	3.2%	5.6%	6.4%	4.8%	7.2%	5.6%	2.4%	5.6%	0.8%	16.0%		1.6%
Other	0.8%	0.0%	0.0%	0.0%	0.8%	0.0%	0.8%	0.0%	0.0%	0.0%	0.0%	2.4%	0.8%	

Table 1: the interest in the community for integrative analysis of specific pairwise omics dimensions as a survey in [6], showing an interest in respondents of 36.8% for transcriptomics-proteomics, 21.6% for transcriptomics-metabolomics and 27.2% proteomics-metabolomics.

The secondary objective will be the interpretation of results using the prior biological knowledge. For this we will reuse much of the existing methodology in gene enrichment for the interpretation of gene lists derived from from differential gene expression analysis. The enrichment analysis on gene lists is usually performed on biological pathways – using sources like Reactome or Panther – or Gene Ontology (GO) terms. The later has become quite extended in the community, even if GO terms are not pathways which are well understood by biologists. Instead GO is formed by an Acyclic Directed Graph (ADG) forming a complex hierarchy on three different spaces: biological process, molecular function and cellular component. The main resource that associate GO terms to genes is the UniProt Gene Ontology Annotation (UniProt-GOA) (http://www.ebi.ac.uk/GOA). The GO underlying structure and the associations to genes in GOA will be the two sources of biological knowledge that we will be using.

Being GO organized as a Semantic Web ontology helps employing it programmatically. The use of biomedical ontologies is an active field of research fueled by initiatives like the Open Biomedical Ontologies (http://www.obofoundry.org/), being some of the most popular the Gene Ontology, the Sequence Ontology or the Disease Ontology. We could think of extending the enrichment approach to other ontologies or "jump" between ontologies to extend the analysis into different areas of knowledge (drugs, diseases, etc.). A rich ontology annotation environment might facilitate in the future the interpretation of a more heterogeneous multi-omics dataset not only using genes as the biological entities connecting the knowledge silos.

Several approaches exist to facilitate the interpretation of the enrichment analysis results based on MDS and clustering. Revigo [11] is a very accessible implementation to this technique available via web . Other main actor is the standard de facto for network representation Cytoscape. Even if Cytoscape is a more general tool for network visualization, plugins exist for this specific problem like BinGO [12] and EnrichmentMap [13].

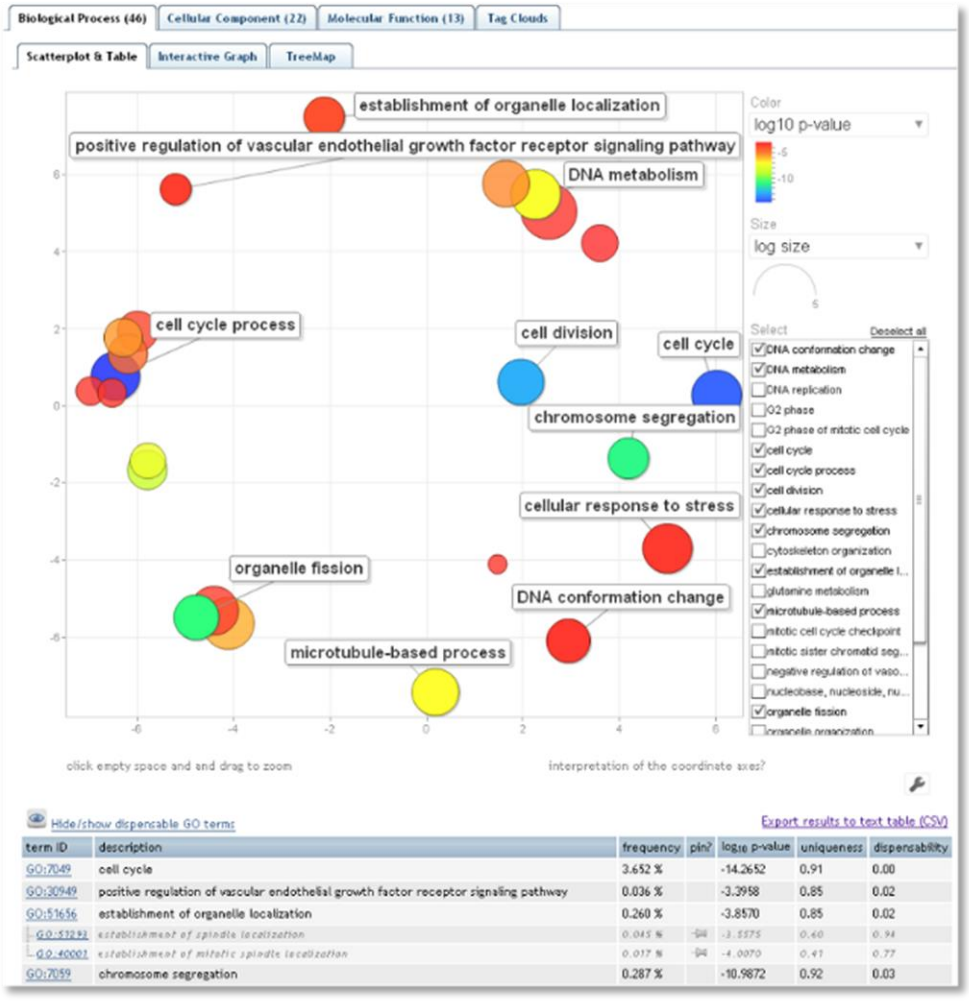


Figure 2: Revigo plot of biological process GO terms after MDS and clustering. The color represents the significance of the enrichment result and the size of each node represents the size of the GO term defined as the number of genes that are associated it.

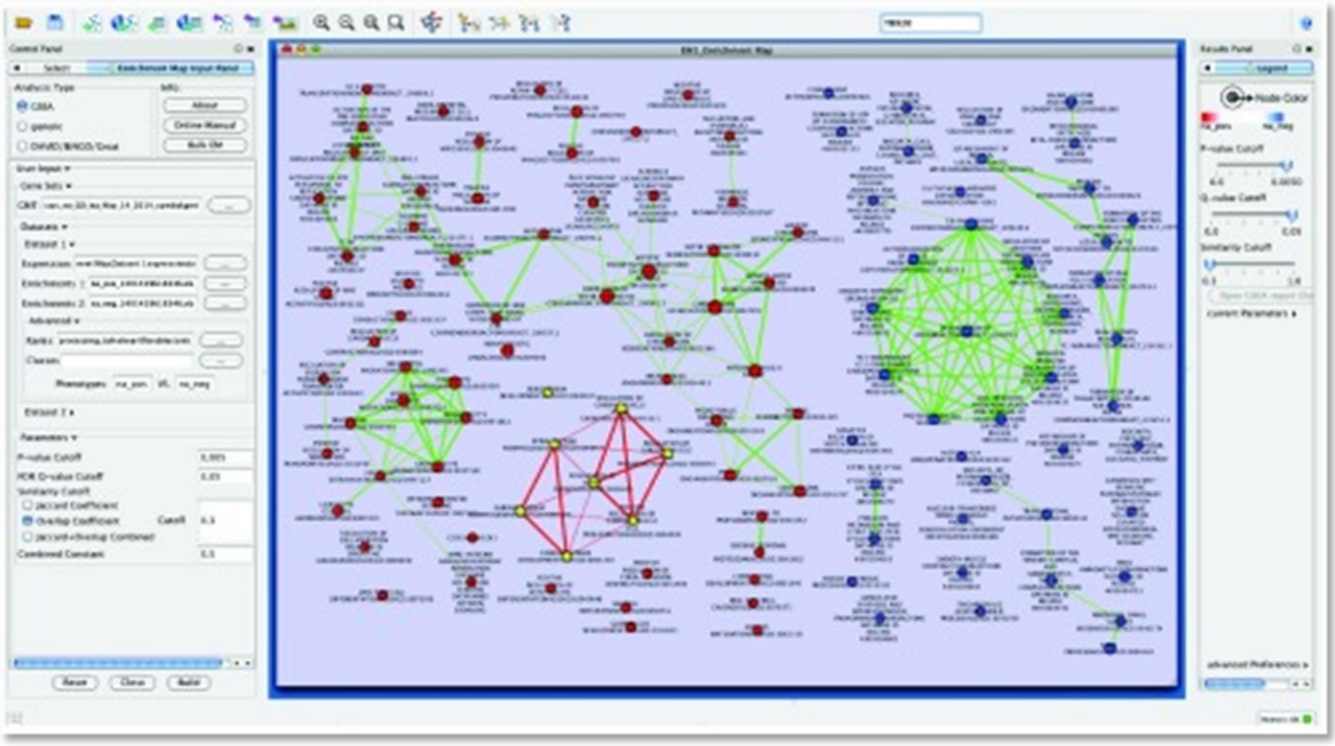


Figure 3: Visualization of Enrichment Map in Cytoscape after some user naming and classification of results.

The objectives of this project are:

- Evaluate the existing state of the art on a selected public dataset
- Analyze alternatives to include prior biological knowledge in the analysis
- Explore different visualization approaches to integrative analysis results

The project will be organized as described in the Gantt diagram on Figure 4.

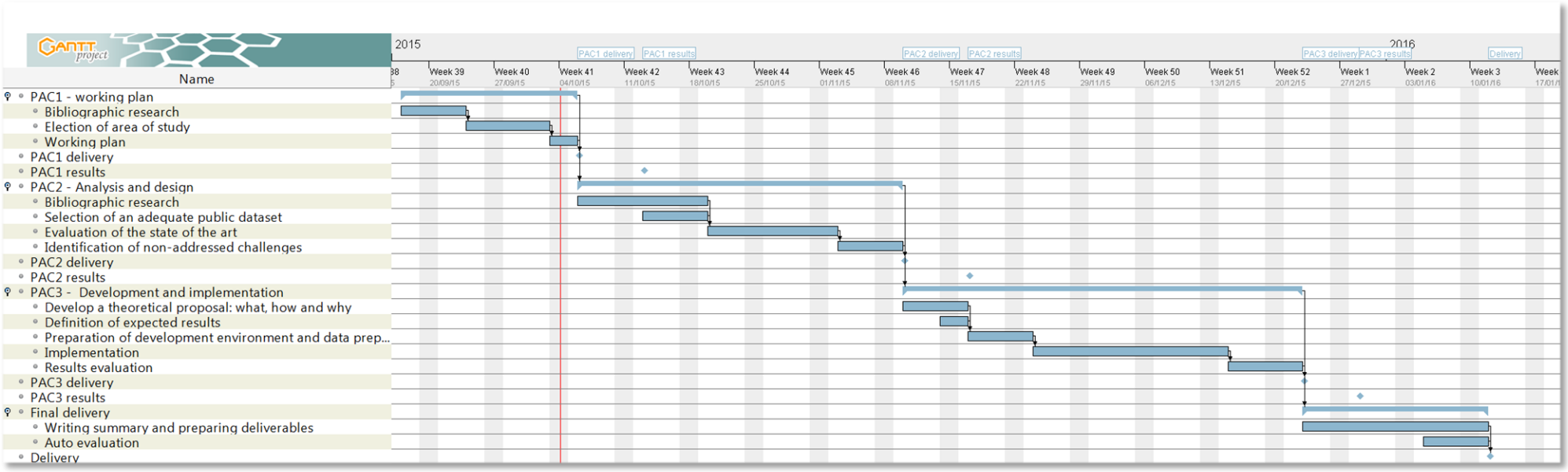


Figure 4: Gantt diagram with the working plan defined for the project.

Methodology

Selection of an adequate public dataset

It exists an unexplained variance in cancer allegedly due to technical reasons for which we fail to identify driver genes in tumor samples and/or the existence of epi-driver genes [14], but the biological knowledge is not complete and thus there is interest in combining different types of data to elucidate underlying mechanisms. The historic NCI60 cell lines have been employed to generate several datasets used as a test dataset for multi-omics data analysis [15]. Nevertheless, NCI60 dataset is a "difficult" dataset as it has only 59 samples while it has several tumor types and the number of variables might be of many thousand depending on the type of data under analysis.

In this case it is of special interest for us the TCGA dataset. The Broad Institute set up GDAC Firehose with the aim of systematizing analysis from The Cancer Genome Atlas pilot data. Several standard de facto pipelines are run on TCGA data to provide the end user with normalized data on the different types of data (i.e.: copy number variants, methylation, etc.). Firehose provides gene expression from RNAseq normalized with RNAseq by Expectation Maximization (RSEM) and protein expression from Reverse Phase Protein Arrays (RPPA) normalized data. As an "easy" dataset we will select only two tumor types: breast cancer and ovary cancer and methylation data will not be included as TCGA does not contain this type of data. Downloading Firehose datasets is facilitated by using the "RTCGAToolbox" [16] package in Bioconductor.

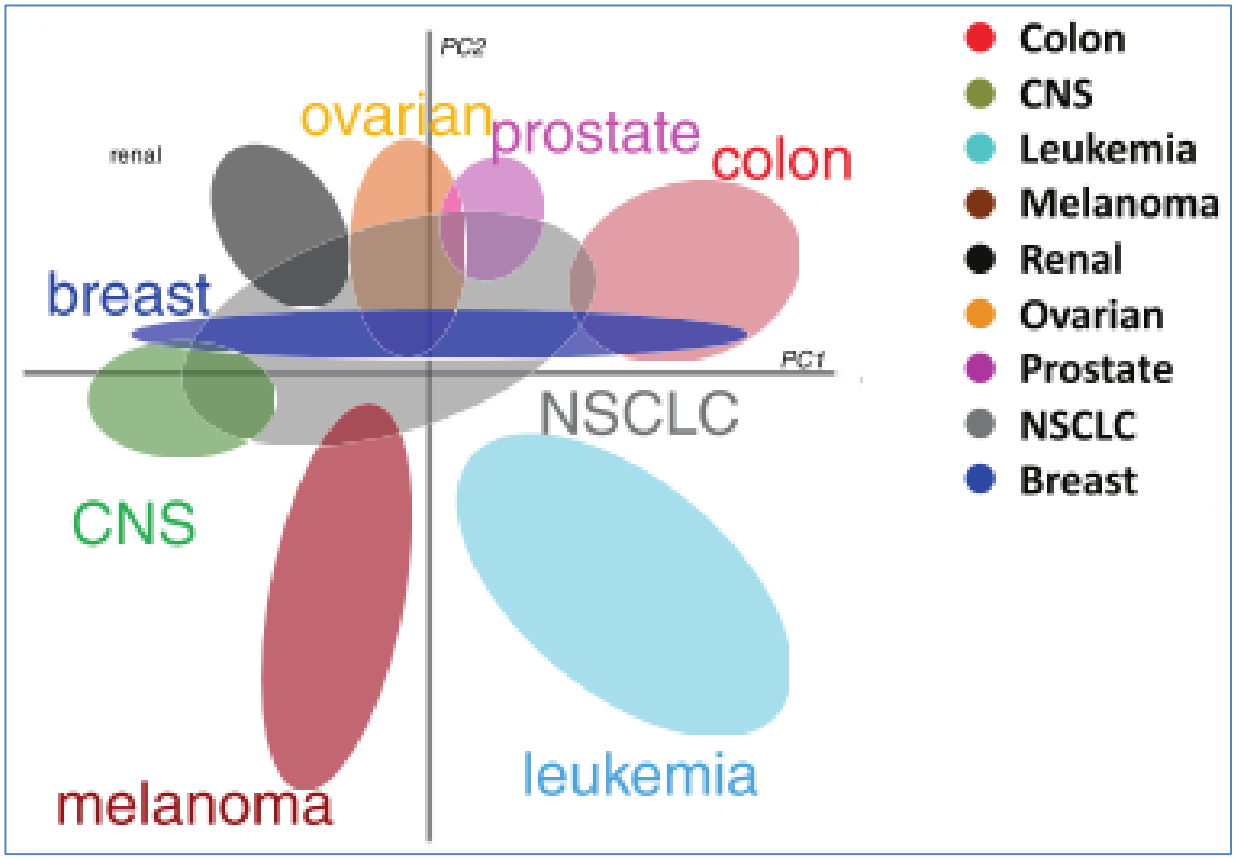


Figure 5: Relation between samples of different tumor types on a reduced dimension space.

Preprocessing

The initial datasets of breast cancer and ovary cancer contain respectively 1212 RNAseq and 410 RPPA, and 265 RNAseq and 412 RPPA samples; corresponding to 20501 RNAseq variables and 142 and 165 RPPA variables. Data preprocessing is needed in order to remove missing values across variables and across samples. It is also important a normalization step on the identifiers provided by Firehose, we use Biomart and its corresponding R library [17] to check that all identifiers match to HGNC gene names. Antibodies from RPPA are also matched to gene names and checked with Biomart. In the case that several RPPA variables match the same gene only the variable with the maximum absolute expression is kept.

In order to reduce dimensions on the dataset we removed zero and near zero variance variables. Variables with a correlation over 0.7 are also removed. Finally, variables are scaled so they are comparable. After preprocessing we have respectively **407 and 201 samples of breast and ovary cancer**; and **12880 RNAseq variables and only 55 RPPA variables**.

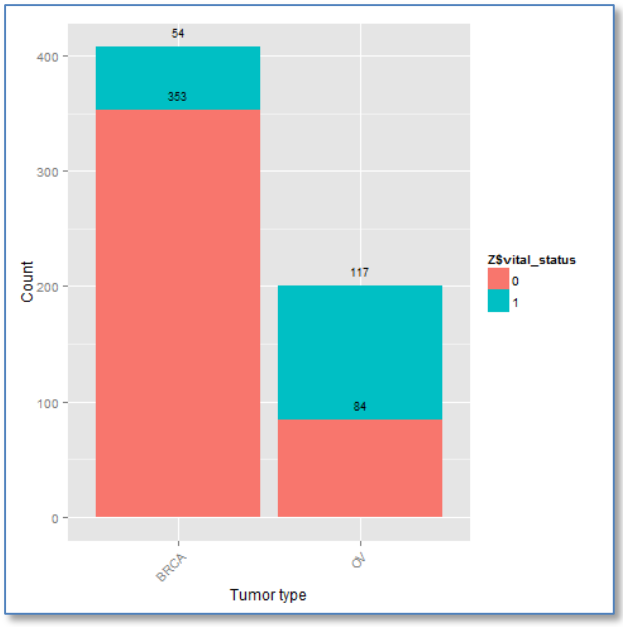


Figure 6: Distribution of samples by tumor type and vital status after preprocessing. All the samples have data for RNAseq and RPPA variables.

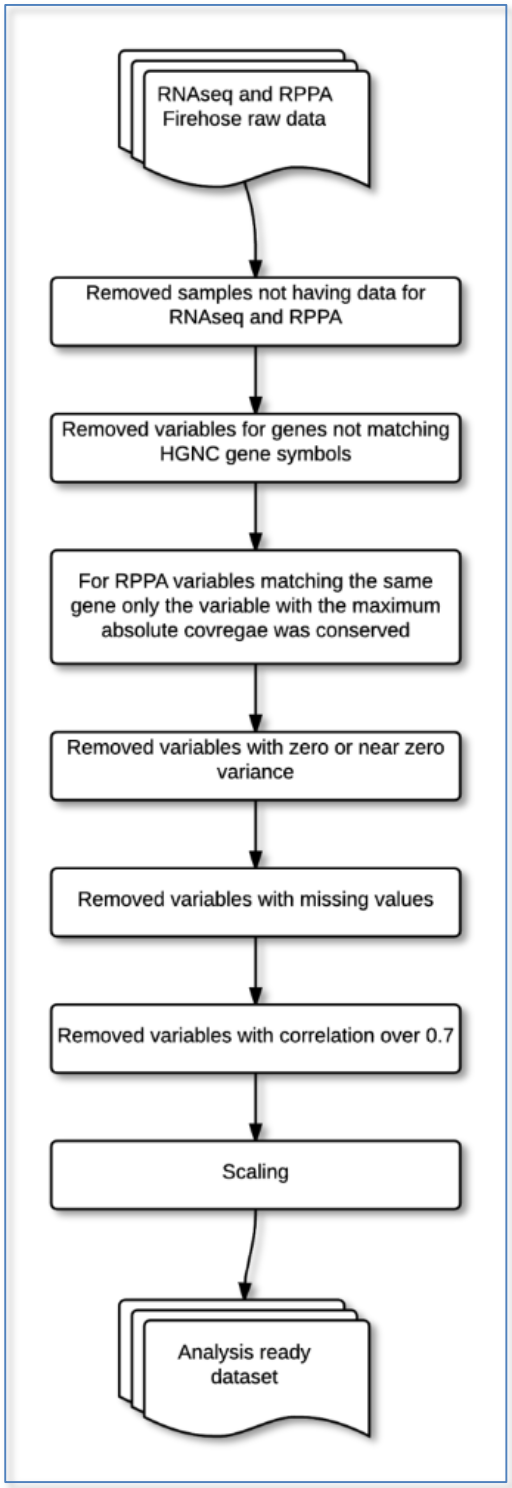


Figure 7: Data preprocessing pipeline

Principal Components Analysis

Principal Components Analysis is a dimensionality reduction technique that allows visualizing high dimensional data in a low dimensional space maximizing variance. We perform a PCA on three sets of data: (1) the RNAseq data alone, (2) the RPPA data alone and (3) the joint dataset RNAseq+RPPA. In all three analyses we observe that the two tumor types are well differentiated. The percentage of explained variances is for (1) 8.8% and 4.1%, for (2) 26.7% and 10.6% and for (3) 8.9% and 4.2%. It is expected that components in (2) explain considerably more variance as we only have 55 variables; but in (2) there is a perfect separation by tumor type along PC1 which is masked when we are using the RNAseq data. This will be an inherent problem of this dataset, the RPPA information will be masked by noise in the RNAseq dataset. The conclusion is that RPPA data allows to better separate the type of cancer, on the other hand, the explained variance by PCA method is just 36.7%.

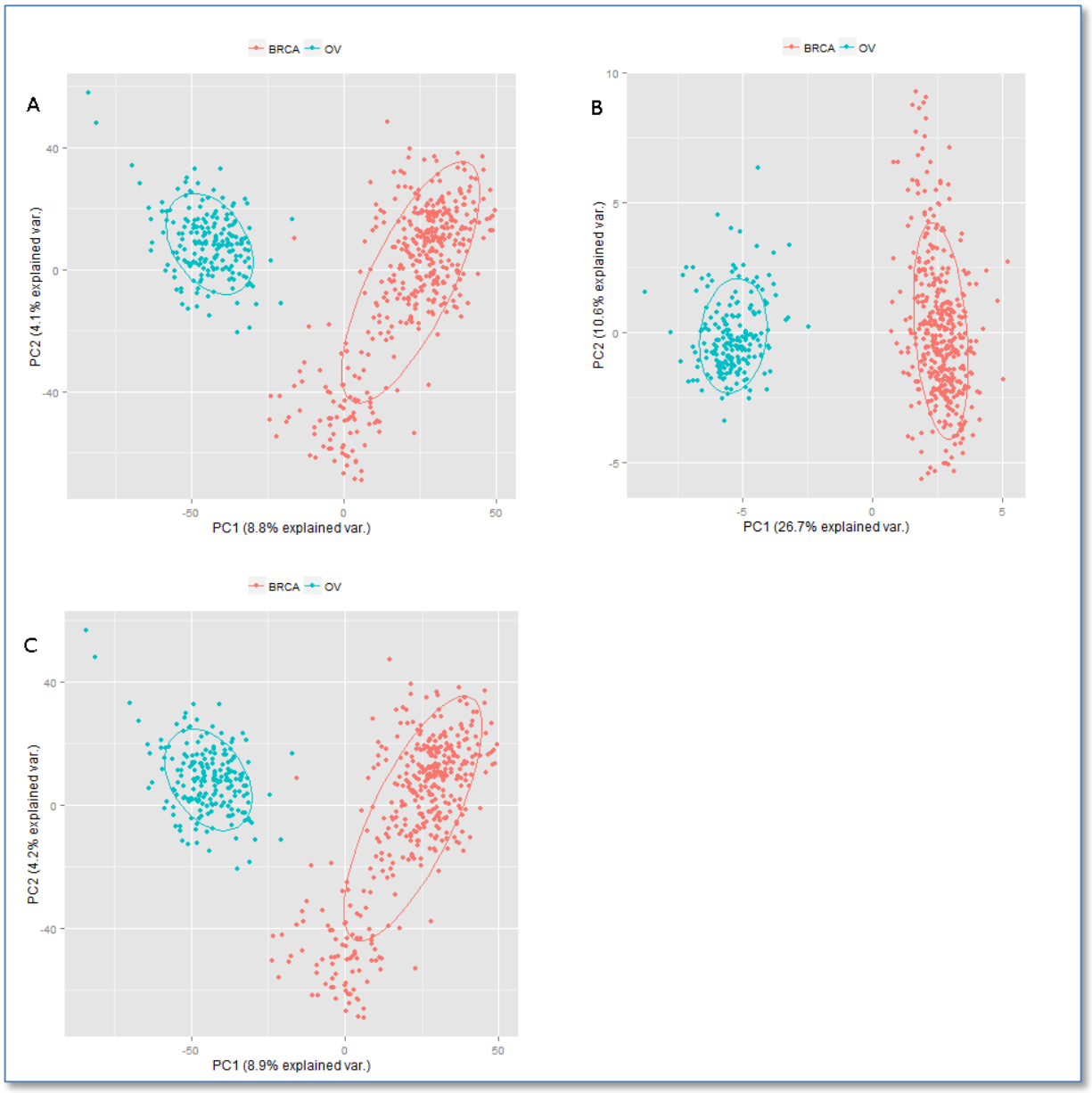


Figure 8: Samples plot on two first principal components (A) on the RNAseq matrix, (B) on the RPPA matrix and (C) on the joint matrix of the two previous (BRCA on red; ovary cancer on green).

Hierarchical clustering analysis

Hierarchical clustering is a clustering technique based on a distance metric and a linkage criterion (i.e.: distance definition for groups of more than two elements). In this case we used the defaults Euclidean distance and complete linkage method. We performed a hierarchical clustering on three sets of data: (1) the RNAseq data alone, (2) the RPPA data alone and (3) the joint dataset RNAseq + RPPA. The results are coherent with the PCA results. The maximum variance is observed in (2) where we can observe a clear separation in two groups and a third small subgroup; analyses including RNAseq data (1) and (3) show minimal variance and the separation between tumor types is less clear.

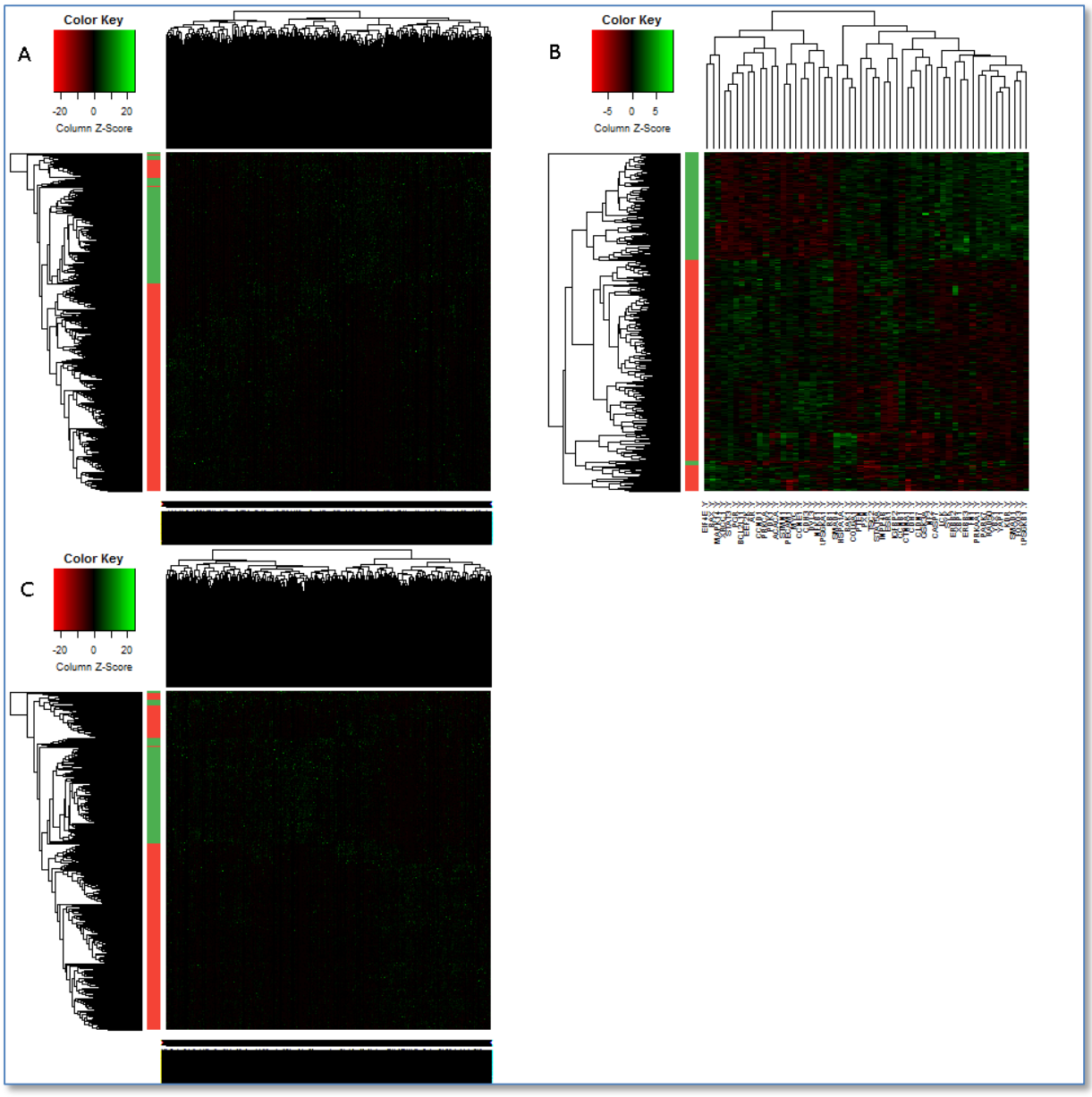


Figure 9: Hierarchical clustering (A) on the RNAseq matrix, (B) on the RPPA matrix and (C) on the joint matrix of the two previous. The samples are on the Y-axis and the variables are on the X-axis; the color band on the Y-axis represents the tumor type (BRCA in red, ovary cancer in green).

Multiple Co-Inertia Analysis

Multiple co-inertia analysis [9] is an extension of the co-inertia analysis [7] for more than two datasets. As CIA it is performed in two steps (1) on each dataset separately an ordination method is applied (e.g.: PCA or COA) and (2) maximize the squared covariance between eigenvectors for each dataset. This results in a transformation of data into a new space where samples and variables are represented in the same coordinates that maximize covariance. The proximity of two elements in the new space represents a high covariance between them, while the distance to the origin represents the magnitude of the variance.

We can observe on Figure 10.A that MCIA results segregate both tumor types as expected. The first component separate tumor type, while the second component represents the intra-tumor variance. The explained covariance is of 34.05% and 11.55% respectively, which is higher than the explained variance obtained with PCA for the joint analysis. The selected variables in Figure 11 A and B are those that maximize covariance in any of the first three components for each dataset. We could imagine a tool that allows the user to select manually the variables of interest before further processing; for instance, those variables separating tumor types on the first component might be of special interest. Furthermore, we can associate specific variables with samples as they are both plotted in the same space, so we can identify those variables that maximize covariance – and thus variance – for a specific tumor type. The RNAseq expression for gene ZNF638 (in Figure 11.A) is located in the far right of the space variance for breast cancer samples (samples in red in Figure 10.A) is 1.92 times that of ovary cancer samples (samples in green in Figure 10.A); while for gene ZNF837 located at the far left the variance for breast cancer samples is 0.11 times the variance for ovary samples. When taking a gene with an extreme value in the second component, like KLHDC1, we verify that variance is more stable between tumor types, just 1.03 times.

In our automated variable selection implementation, we select those variables with highest variance in the three first components. 5 variables for each extreme of each component of each dataset, that is a maximum of 60 variables considering that there might be overlaps between space components and datasets. In our test dataset we selected 56 variables, without overlap between the datasets, being 30 variables selected from RNAseq data and the other 26 from RPPA data.

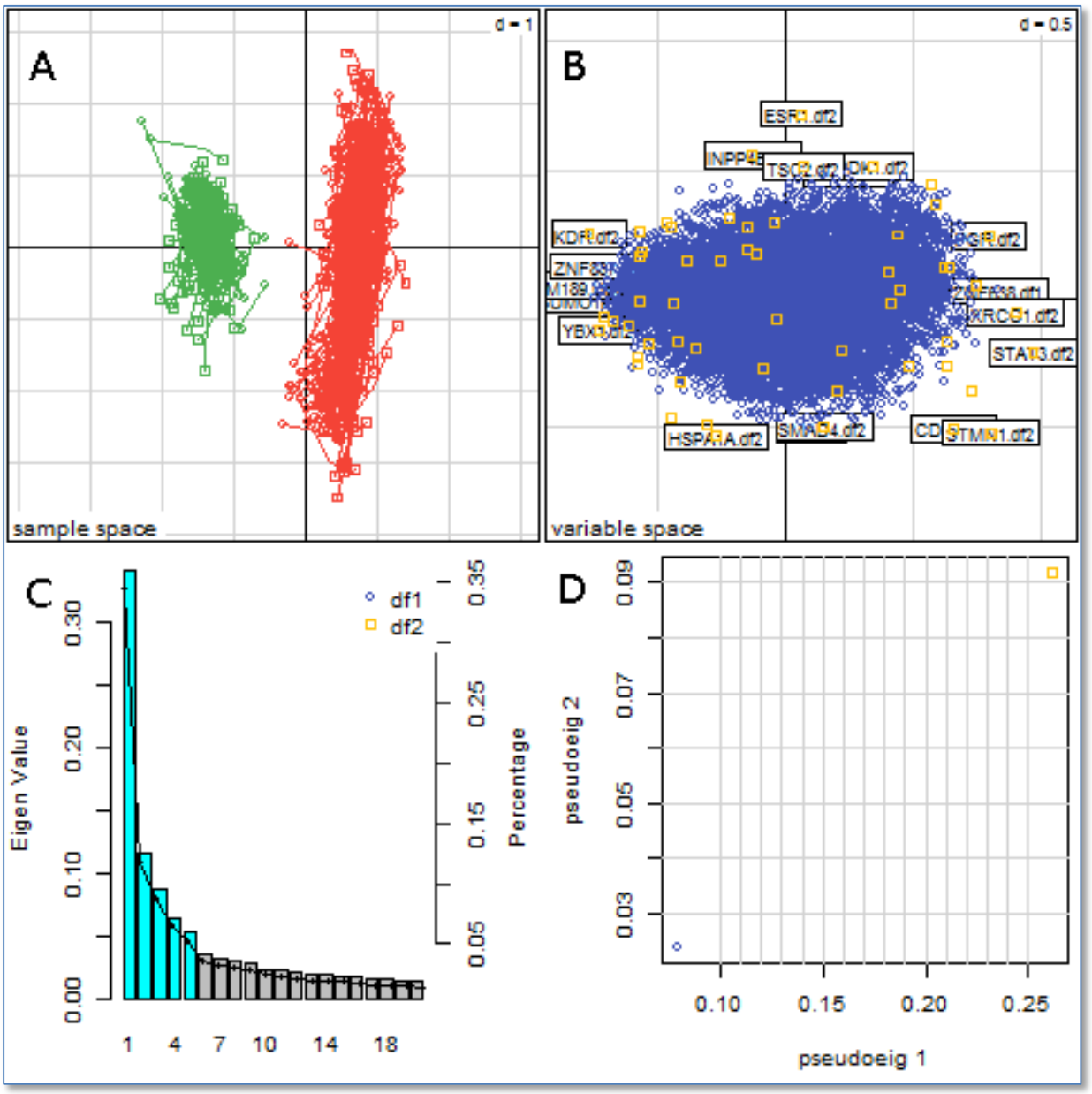


Figure 10: (A) samples represented in the two first components of the new space by tumor type (green ovary cancer; red breast cancer), each sample is represented twice for RNAseq and RPPA data and connected by lines, the further apart the lower the covariance, (B) variables represented in the two first components of the new space by technology (blue RNAseq; yellow RPPA), (C)variance explained by each of the Eigen vectors, (D)…

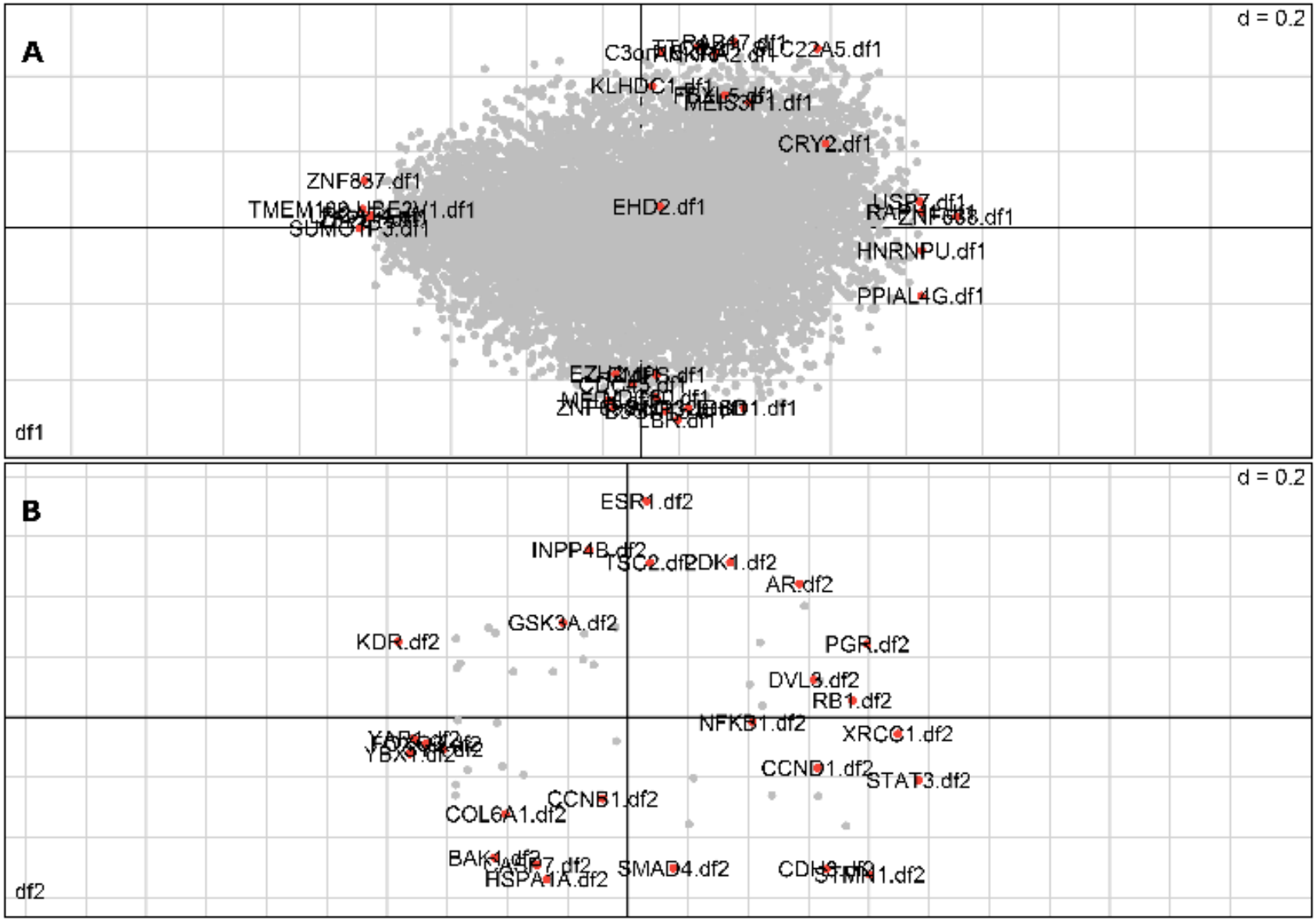


Figure 11: Selected variables for (A) RNAseq variables on X matrix and (B) RPPA variables on Y matrix. The selected variables are those that maximize covariance in any of the 3 first components. As the third component is not visible in this graph we are seeing variables around the center of the plot.

Sparse Partial Least Squares analysis

Sparse Partial Least Square [8] is a technique based on Partial Least Squares. PLS is based on the decomposition of the two datasets (explained and explanatory variables) on latent variables that maximize covariance. The sparse approach applies a penalization to loading vectors such that datasets with number of variables >> number of samples can be computed using the regression or the canonical model, here we only analyzed the regression, further analysis on the canonical mode results is needed. Using sPLS requires the introduction of a mathematical artifact: the need to define an explanatory and an explained dataset. In our case we defined the RNAseq dataset as explanatory and the RPPA dataset as explained.

The correlation plot shown in Figure 12.A helps identifying those variants with a highest correlation as highly correlated variables cluster together, the angle between variables is determined by their correlation as explained in Figure 12.B.b retrieved from [18]. We can observe four clusters in each extreme of the first two components, this is expected as these two components maximize covariance. Unlike MCIA we can not map between variables and a set of specific samples, but on the other hand we have the ability to identify correlations between variables on different datasets. It is important to note that sPLS approach is restricted to two datasets.

Our variable selection approach is similar to that followed with MCIA results, we select those variables having higher loadings positive and negative of each of the three first components of each dataset. Like in the correlation plot this corresponds to those variables having a higher distance from the origin in the three-dimensional correlation plot. In our test dataset we selected 58 variables, with an overlap between the datasets of two genes. Being 30 variables selected from RNAseq data and other 30 from RPPA data.

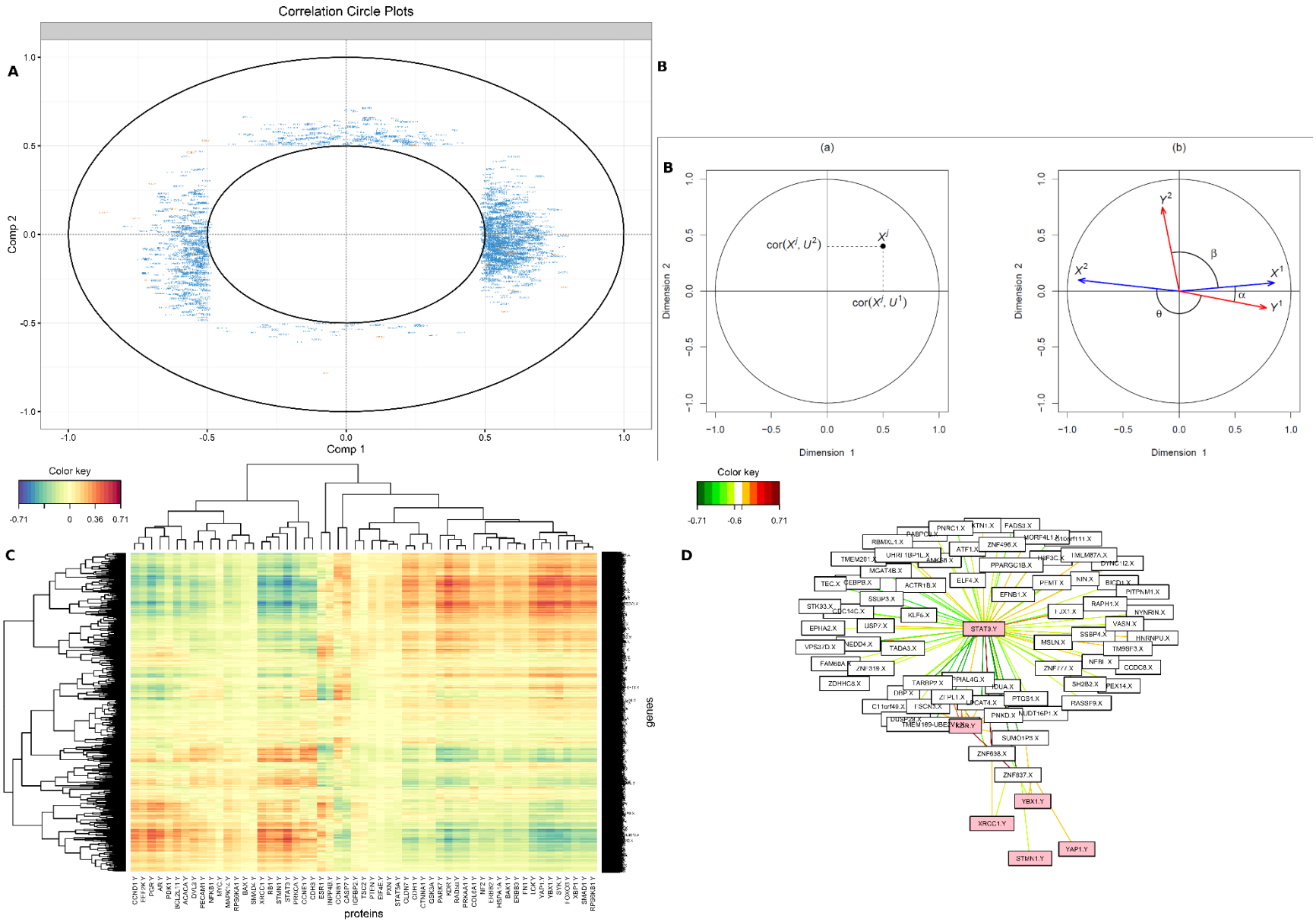


Figure 12: (A) correlation plot, RNAseq variables in blue and RPPA variables in orange; (B) a) Coordinates of the X-variables on the plane defined by the first two variates U1 and U2. b) The correlation between two variables is positive if the angle is sharp cos(α)>0, negative if the angle is obtuse cos(ϑ)<0, and null if the vectors are perpendicular cos(θ)≈0 [18]. (C) heatmap of covariance between RNAseq variables (Y-axis) and RPPA variables (X-axis); (D) relevance network of variables with a correlation higher than 0.7, in red genes inferred from RPPA dataset, in white genes inferred from RNAseq dataset, the nodes represent the strength of the association. An association can only exist between an explanatory and an explained variable.

Other methods

Other methods exist that target similar problems that will not be studied:

- Regularized Canonical Correlation Analysis (rCCA)
- Regularized Generalized Canonical Correlation Analysis (rgCCA)
- Canonical Correlation Analysis with Elastic Net penalization
- Procrustes

Some of this methods lack a public implementation and others are not fitted to compute on high-dimensional dataset. For instance, rCCA implies a prior step of regularization that uses cross-validation and requires intensive computing. Nevertheless, some of the outputs for rCCA are interesting and similar to those provided by sPLS. rCCA and rgCCA have been integrated into our package TCGAome in experimental mode and disabled by default.

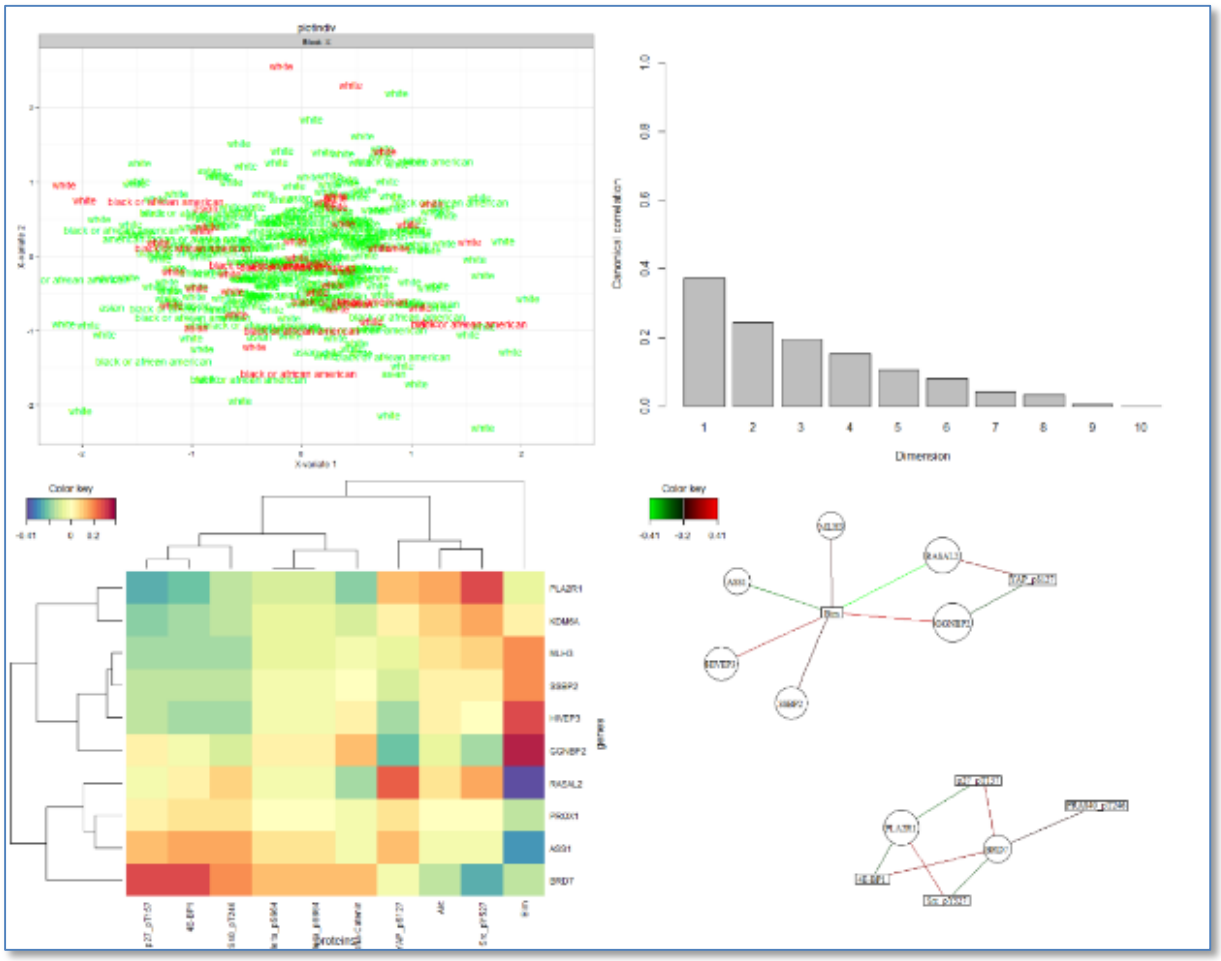


Figure 13: Visual output for rCCA run on a reduced dataset only for BRCA samples with 10 RNAseq variables and 10 RPPA variables randomly selected.

Gene Ontology enrichment analysis

We perform an enrichment analysis using a Fisher test that computes for each GO term the exact Fisher test on the expected associated genes and the observed associated genes. This enrichment relies on the Uniprot-GOA annotations as the source of expected information.

In order to represent GO terms enrichment results visually we need to (1) know how to plot those terms spatially and (2) be able to reduce the number of terms to plot. A distance definition is required; in the ontologies domain these are called semantic similarity measures and they use the information implicit in the ontology to determine how close two terms are. Several semantic similarities measures have been proposed in the domain of text processing and artificial intelligence for the WordNet ontology [19] [20]and these have been later applied in the biomedical domain [21]. Semantic similarity approaches can be divided into edge-based and node-based. The former rely on the shortest path in the ontology between nodes, but as ontologies are not well structured having for instance terms at the same level different degrees of specificity these metrics are not reliable. The later can be based on the Information Content (IC) of each node or on the shared ancestors. The IC is related to the probability of occurrence of a given term and in the GO it is usually estimated as the frequency of annotation in a resource like UniProt-GOA. This can be applied in two flavors the Most Informative Common Ancestor (MICA) and the Disjoint Common Ancestors (DCA). Some of these methods are: Resnik, Lin and Jiang and Conrath. Surprisingly, these methods do not make use of the ontology structure, neither the specific annotations of each GO term relying solely on the probabilistic distribution of annotations. Nevertheless, they are the most commonly used semantic similarity measures [22].

The methods that use shared ancestors, like Wang method, are not making use of the shared genes associated to each term. Other methods exist that use the shared annotations between elements in the ontology [19], but as far as we know it has not been applied to the Gene Ontology. We will call this measures functional similarity measures as we are not using the semantic information in the ontology, but only the functional association information in UniProt-GOA. Our main question is: could we improve the separation and prioritization of GO terms in a multidimensional space by using the functional information of shared genes?

In order to answer this question, we will compare state of the art semantic similarity measures: Wang, Resnik, Lin, Schlicker (rel) and Jiang methods; with several functional similarity measures. These functional similarity measures intend to make use of the information present in Uniprot-GOA that associate GO terms and genes. They are based on comparing those genes associated to each GO term as if they were binary vectors, 1 meaning association and 0 no association. There is a wide literature that evaluates the different binary similarity metrics and their characteristics, we will study a limited subset for our specific problem [23]: binary, Union-Intersection (UI), Bray-Curtis and Ochiai (equivalent to cosine in the binary case).

$$D_{binary}(GO_A,GO_B) = \sum (XOR\ (GO_A,GO_B))$$

$$D_{UI}(GO_A,GO_B) = \sum (\cap(GO_A,GO_B)) / \sum (U(GO_A,GO_B))$$

$$D_{Bray\text{-}Curtis}(GO_A,GO_B) = \sum (2*\cap(GO_A,GO_B)) / \sum (\sum GO_A + \sum GO_B)$$

$$D_{Ochiai}(GO_A,GO_B) = \sum (\cap(GO_A,GO_B)) / \sum (\sqrt{} (\sum GO_A * \sum GO_B)))$$

Each of these metrics intend to provide a “universal” pairwise similarity metric as determined by the information in UniProt-GOA, just as the other metrics described. But, if we restrict the search universe to the input gene list employed to calculate the enrichment analysis we will obtain a similarity specific to the data under analysis. We found that limiting the data to the gene list resulted in an over fit on our data. Similarly, we could extend the search universe to other species than human, using all available annotations in GOA which include several other species. This analysis has not been done. We expect that the functional similarity measures will be missing part of the picture as they ignore the ontology structure, but hopefully a combination of semantic and functional similarity measure might be more informative than current state of the art measures.

Having defined a pairwise similarity measure we perform a clustering on the enriched GO terms based on Partitioning Around Medoids (PAM) and silhouette analysis to select the optimal number of clusters in the range of 2 to 30 clusters. For each cluster we select a representative GO term, being the most significant GO term having a frequency of annotation in GOA under 5% if any. In order to represent the clusters visually we perform a Multi-Dimensional Scaling (MDS) on the data to obtain the most

informative axis. Having that information we are able to create the graphical output: (1) a scatter plot of the cluster representatives showing the enrichment significance represented by its color, the frequency of annotation represented by its size and their relative position to each other in two-dimensions according to the similarity measure employed and the MDS results (Figure 14); (2) a treemap where we show how clusters are formed together with the frequency of annotation of each term represented by its size (Figure 15); and (3) a graph representing the ontology hierarchy of the representative GO terms and their ancestors up to the root of the biological process ontology (Figure 16).

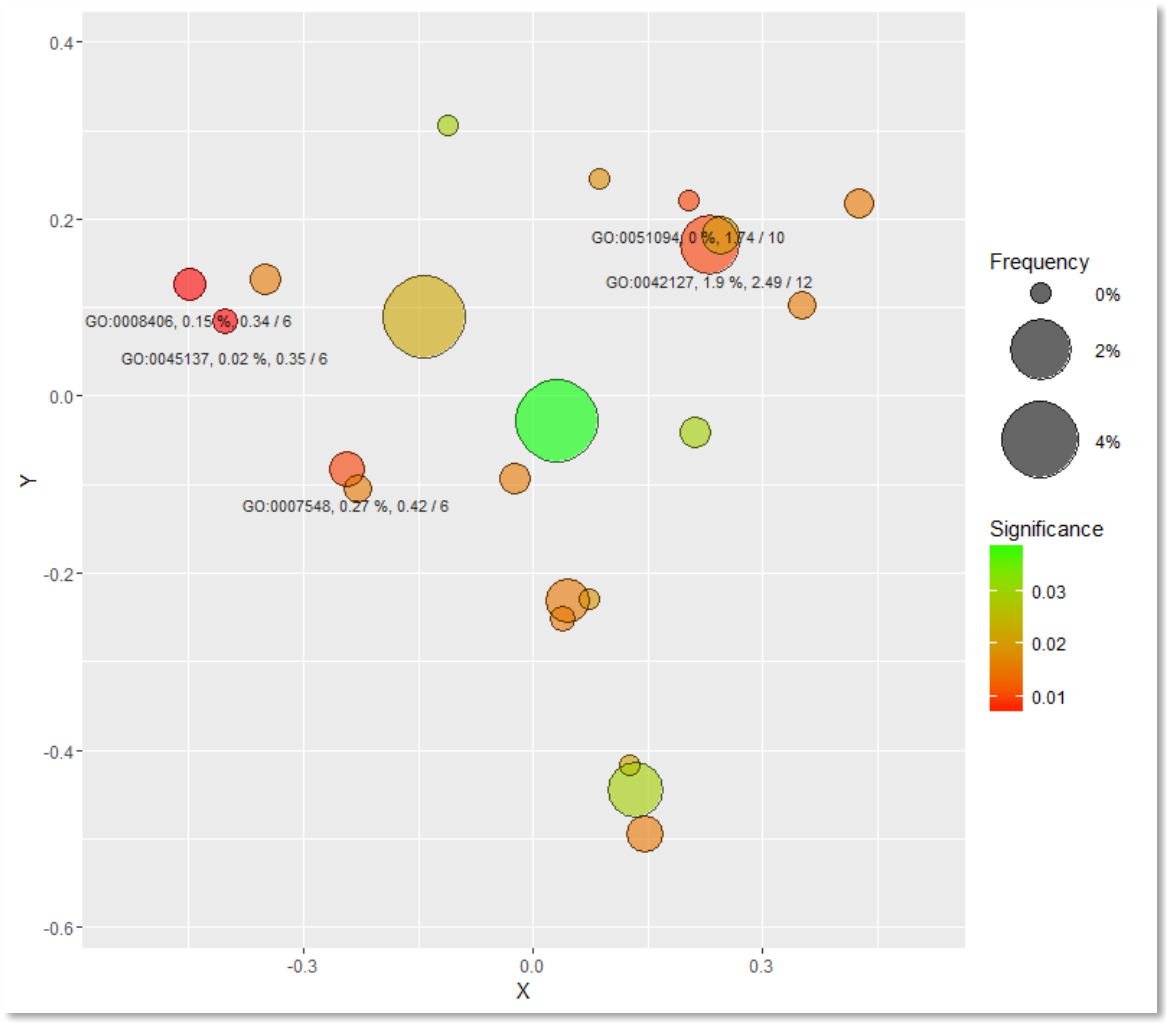


Figure 14: scatter plot showing GO terms clusters representatives on the 2D space obtained by MDS. Those terms more related according to the similarity metric employed are represented closer in the 2D space. The size of each GO term is related to its frequency in UniProt-GOA. The color of each GO term represents its significance by the Fisher exact test after FDR correction. This plot is based on Revigo output [11].

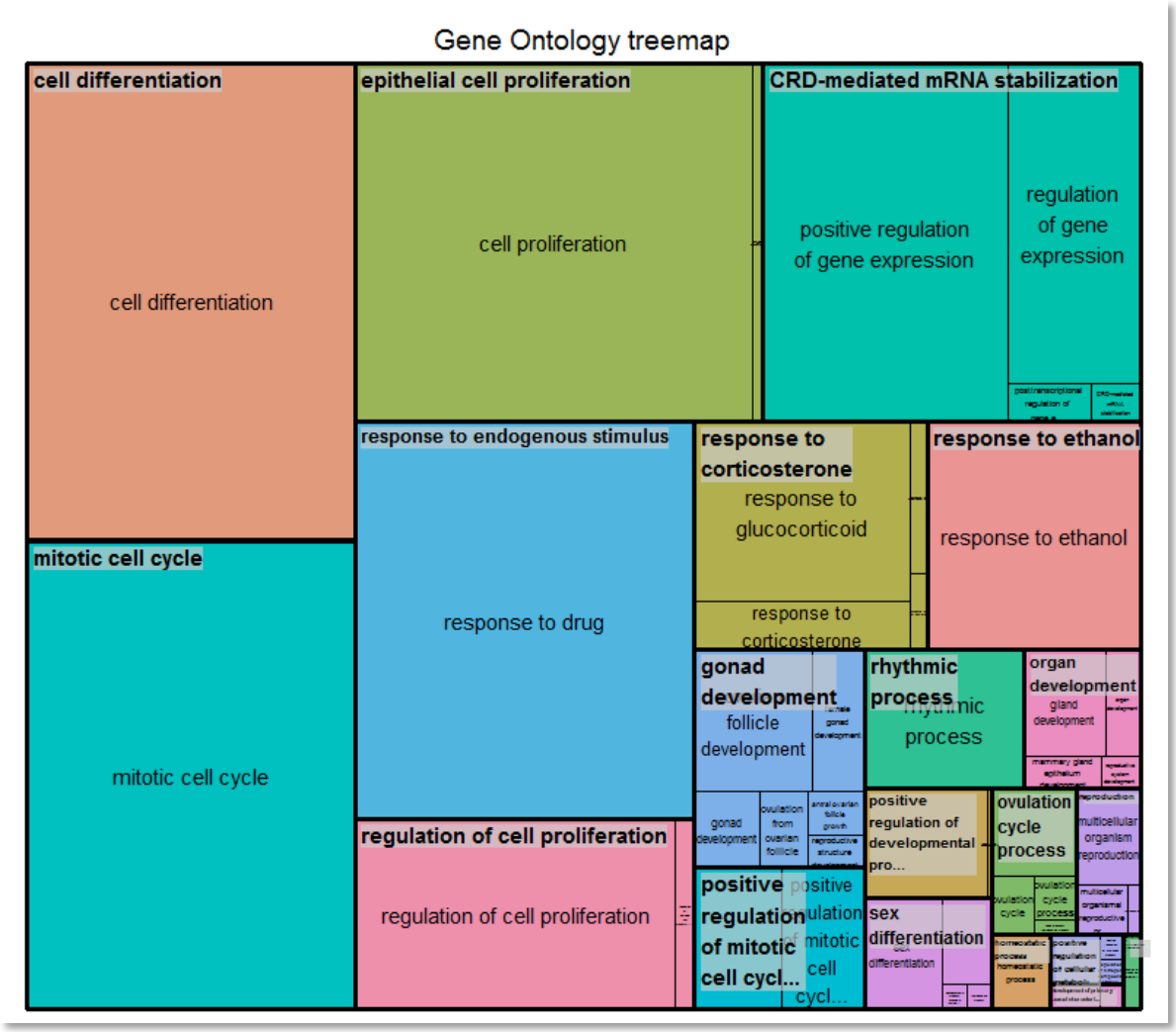


Figure 15: treemap showing the clusters of GO terms with different colors. The GO term representative on the top left corner of every cluster and those GO terms forming every cluster represented as cluster sectors. The size of clusters and of each GO term corresponds to the frequency in UniProt-GOA. This plot is based on Revigo output [11].

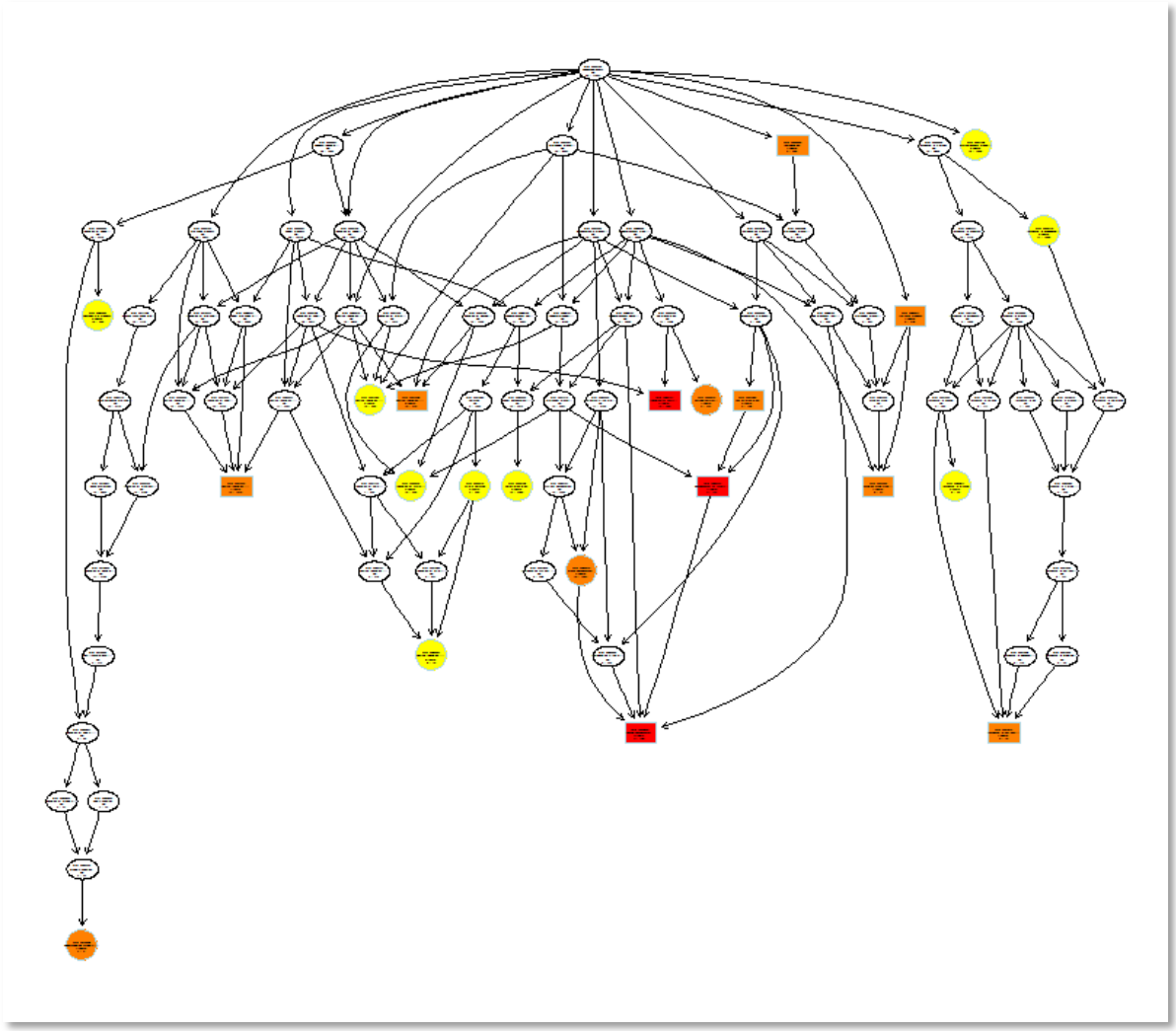


Figure 16: graph of the Gene Ontology hierarchy for the cluster representatives and their ancestors. The color indicates those GO terms more significant, being red the most significant and yellow the less significant. The ancestors up to the root of the ontology are shown without color. Other members of the cluster not being the representative are missing from this graph. This graph is obtained from the TopGO package [24].

Results

The pipeline

The pipeline was implemented as the TCGAome R package and published in GitHub https://github.com/priesgo/TCGAome with MIT license for public use. The results shown here analyze two tumor types: breast and ovarian cancer, nevertheless TCGAome is intended to support the analysis of any number of types of cancer available in TCGA having RNAseq and RPPA data, the list of supported tumor types is available on the package. The command executed for this analysis is `run_TCGAome( tumor_types = c("BRCA", "OV"), GO_similarity_measure = "all")` with default values for all other parameters.

In Figure 17 we represent the different stages in our TCGAome pipeline, the pipeline provides certain configurability, the values provided in this description correspond to the execution under analysis. The pipeline executes the following steps:

- **Data download** from the Broad's Institute FireHose using the R package RTCGAToolbox [16].
- **Normalization of identifiers** using the BiomaRt [17].
- **Preprocessing** as described previously in the methodology.
- Execution of multivariate methods on the data: **MCIA** and **sPLS.**
- **Variable selection** of the most relevant features (i.e.: genes) for each of the methods as described in the methodology. We selected the top 5 results on each extreme of the three first components and on each dataset, that is a maximum of 60 variables.
- **Enrichment analysis**. We run Fisher exact test that compares the counts of genes observed and the expected count of genes for each GO term using the R package TopGO [24]. False Discovery Rate (FDR) multiple test correction is finally not applied to these results as recommended by TopGO authors, but this is configurable. Results are filtered on the resulting p-value using a threshold of 0.01.
- Calculates the **frequency of each GO term** based on the number of genes associated to it. This will be employed to avoid the significance bias of more general terms.
- Calculates the **pairwise similarity measure** by any of the semantic similarity methods based on IC obtained with GOSemSim [25]: "Resnik", "Lin", "Rel", "Jiang"; shared ancestors: "Wang"; and functional similarity measures: "UI", "binary", "bray-curtis", "cosine". This is used to compute the dissimilarity matrix.
- Based on the dissimilarity matrix **clusters the GO terms**. First we evaluate the optimal number of clusters by silhouette analysis from 2 to 30 clusters using the Partitioning Around Medoids (PAM) clustering algorithm.
- Computes the **Multi-Dimensional Scaling (MDS)** on the clustering results to reduce cluster representation to two dimensions.
- For each cluster it selects the most significant term as **cluster representative** having preference to those with a frequency not higher than 5%.
- Finally, it plots results:
    - **Scatter plot** (Figure 14) showing cluster representatives in 2D.
    - **Treemap** (Figure 15) showing all those terms contained in each cluster.
    - **Ontology graph** (Figure 16) of cluster representatives and their ancestors.

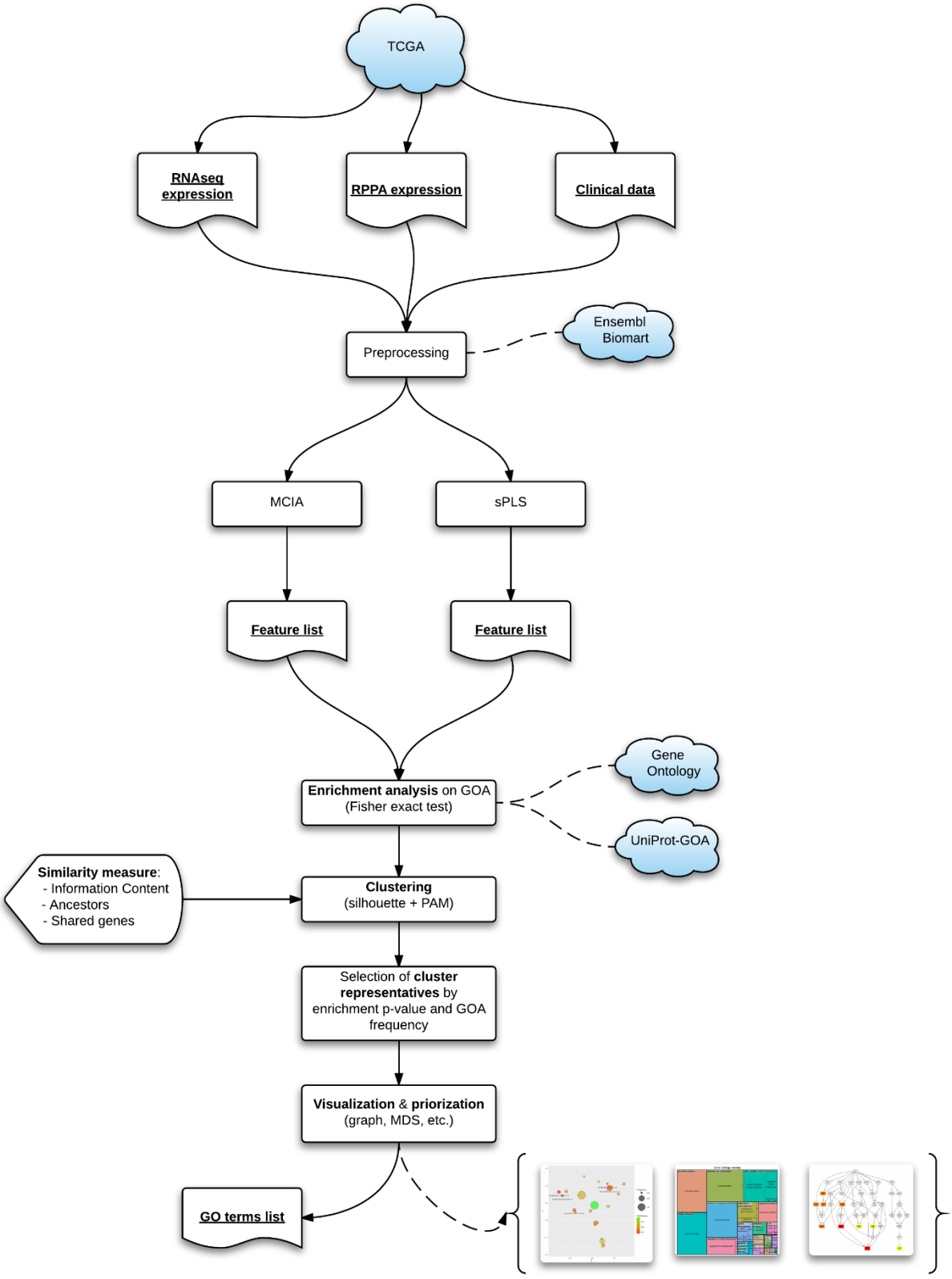


Figure 17: pipeline diagram representing from data retrieval to the final generation of visualizations.

MCIA vs sPLS

If we compare the results of both methods, we observe that only 21 genes – less than half – were selected by both methods (see Figure 18). This leaves 37 genes uniquely identified by each of the methods. The results were also compared with known tumor suppressor genes and oncogenes [14] – a list of 125 genes –finding that only 3 were selected by both methods and 1 and 5 more were selected by MCIA and sPLS respectively. We also compared it with the Tumor Suppressor Gene Database that identifies those genes being differentially expressed in specific tumor types; finding that only 4 were selected by both methods and 7 and 5 respectively for MCIA and sPLS. These results seem to show low concordance with knowledge in cancer and between methods.

There are two genes selected by both methods, AR and SMAD4, being known to be, respectively, an oncogene and a tumor suppressor gene (TSG) [14]. By doing a quick search in the literature for the gene name and tumor type both genes are known to be associated to breast and ovary cancer. It is interesting to analyze the coordinates of these genes in the MCIA results variable space (Figure 11), considering that the first component represents the inter-tumor variance, while the second component represents the intra-tumor variance. For AR we only have data on protein expression having the coordinates 1.11 and 1.50. While for SMAD4 we have 0.29 and -1.64 for protein expression and -0.37 and 0.03 for gene expression. These results are coherent with our literature search as both genes show a higher variance in the second component, indicating that they might be good candidates to separate tumor subtype, but not between our tumor types, breast and ovary cancer. Also, the MCIA method identifies two more genes, RB1 and EZH2, being known to be a TSG and an oncogene, respectively, and they are associated to both tumor types. The gene RB1 is a good candidate to separate both types of tumor according to its MCIA coordinates of 1.46 and 0.23 for protein expression. The gene EZH2 is strongly associated with the intra-tumor variance having coordinates for gene expression of -0.23 and -2.47. As a conclusion, we have identified 4 genes which are known to be associated with cancer and more concretely with breast and ovarian cancer. One of the genes was obtained from gene expression data while the other three genes were obtained from protein expression data. Even if we are making use of the two omics datasets to identify genes we have not demonstrated that the joint analysis is superior to the independent analysis. To complete the analysis, we can use sPLS correlation results to evaluate potential gene expression regulation candidates (see Figure 19). Analyzing the AR proximity, we can identify genes whose gene expression is positively correlated to AR's protein expression.

When we compute the GO term enrichment in each of the gene sets it is surprising that while the 58 genes selected by MCIA enrich significantly (p-value < 0.01, no multiple test correction) for 90 GO terms, the 57 genes selected by sPLS enrich significantly for 313 GO terms; having both sets of GO terms and overlap of only 38 (see Figure 20). Also, sPLS enrichment show higher concordance with biological knowledge. It would be difficult to drive conclusions from this results without a detailed analysis of the genes and GO terms identified with each method. But, definitely we can conclude that both methods are targeting different events. As both methods maximize variance using different techniques a deeper analysis is needed to fully understand the differences between both results.

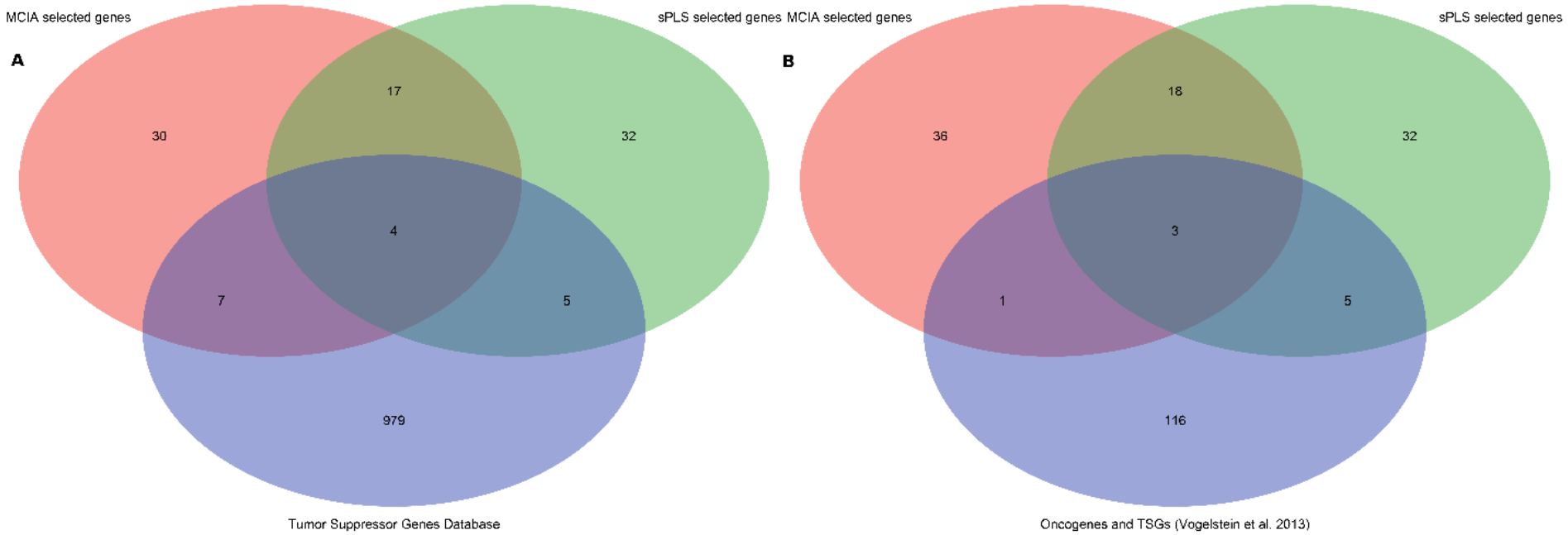


Figure 18: (A) Venn diagram comparing the selected variables obtained with the methods MCIA (red) and sPLS (green) and the tumor suppressor genes in the Tumor Suppressor Gene Database [26] (blue); (B) Venn diagram comparing the selected variables obtained with the methods MCIA (red) and sPLS (green) and the tumor suppressor genes and oncogenes published in [14](blue).

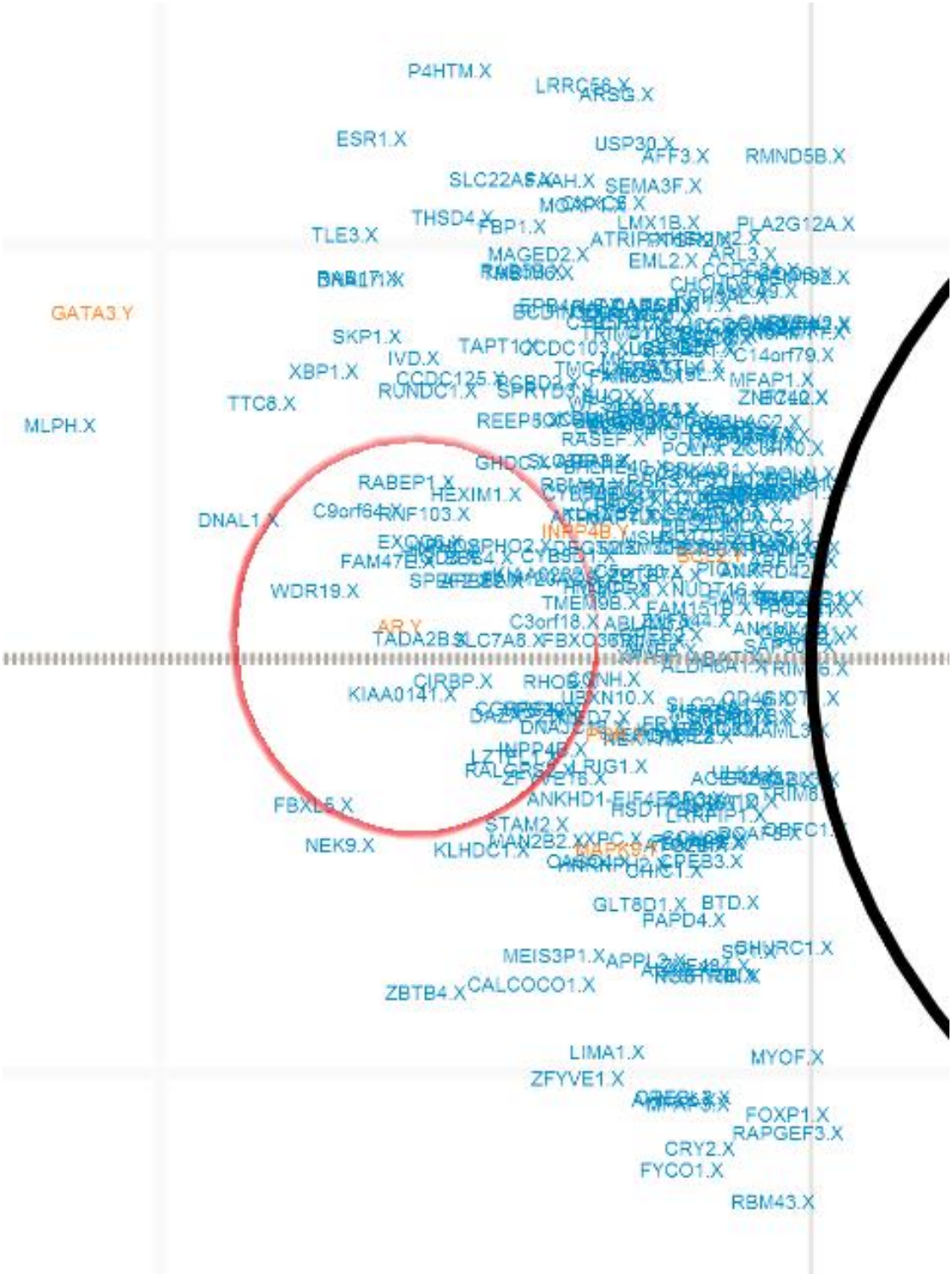


Figure 19: sPLS correlation plot detail for protein expression AR (in orange) and those gene expressions highly correlated (elements in blue in the proximity of AR)

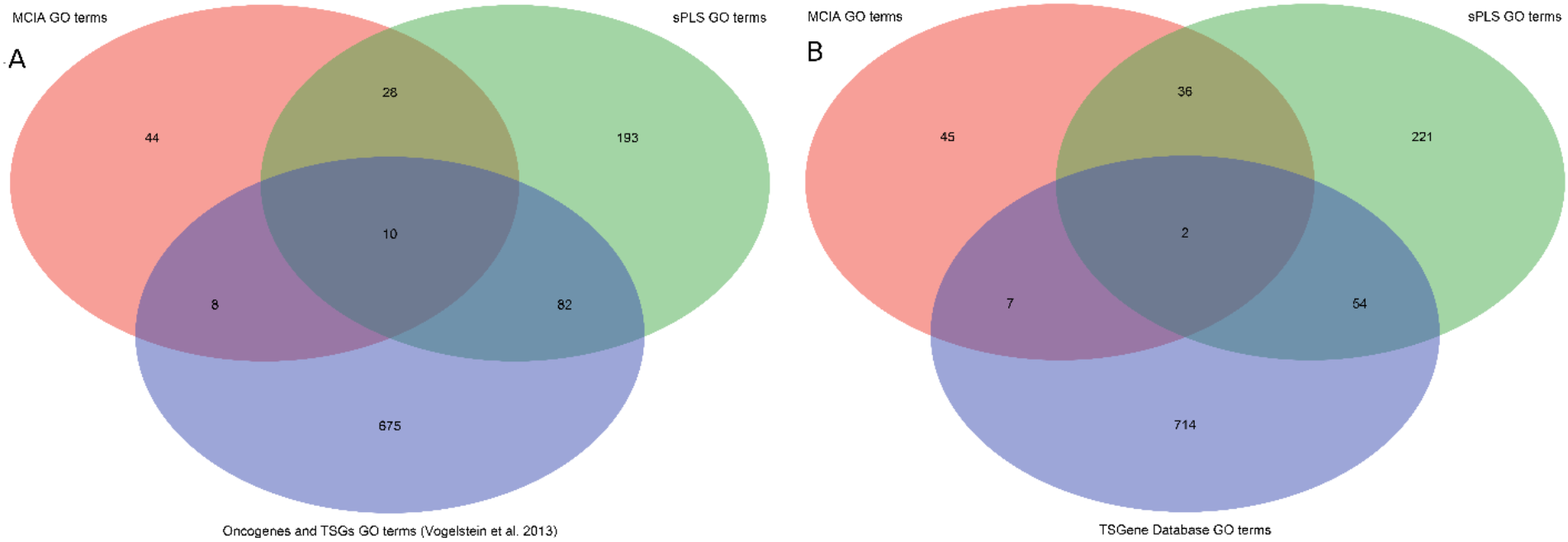


Figure 20: (A) Venn diagram comparing the enriched GO terms for the set of genes obtained with the methods MCIA (red) and sPLS (green) and the tumor suppressor genes in the Tumor Suppressor Gene Database [26] (blue); (B) Venn diagram comparing the enriched GO terms for the set of genes obtained with the methods MCIA (red) and sPLS (green) and the tumor suppressor genes and oncogenes published in [14](blue).

Description	MCIA	sPLS
Both MCIA and sPLS are time performant.	⊕	⊕
MCIA supports more than two matrices. Furthermore, sPLS assumes that one of the data matrices contains the explanatory variables, while the other contains the explained variables. This might a mathematical artifact but definitely this assumption of causality might be erroneous in many situations.	⊕	⊖
sPLS returns a result on covariance allowing to plot networks of relations between variables. MCIA does not return a covariance result for each pair of variables; instead it runs a joint analysis and returns an absolute result per variable.	⊖	⊕
MCIA plots samples and variables on the same space, allowing segregating samples by dominant variables driving their variance and vice versa.	⊕	⊖
Both MCIA and sPLS return no statistical significance results.	⊖	⊖
Both MCIA and sPLS variable selection uses top ranked criteria on the different datasets which leads to strictly symmetric variable selection which might be misleading.	⊖	⊖
MCIA and sPLS results are not very coherent (i.e.: more than half of variables selected are unique to each method) and further analysis needs to be done to understand the differences.	⊖	⊖
sPLS does not provide explained variance results on the latent variables.	⊕	⊖

Table 2: pros and cons of MCIA and sPLS

Comparison of similarity measures

In order to evaluate our results, we run some overall comparison across all similarity measures. The first thing that we want to understand is the dispersion of GO terms in the two-dimensional space after MDS and how informative the similarity measures are. To assess the dispersion, we compute for each measure the distances to the centroid of every GO term. We can observe in Figure 21 that the mean and median for measures based in shared genes are significantly smaller than the rest. Also, the distribution of distances to the centroid is more condensed and at the same time there exist more outliers. We can thus state that the 2D space is on average less informative for measures based on shared genes as GO terms tend to be closer. We will need to evaluate the adequacy of representing the distances between GO terms in just two dimensions.

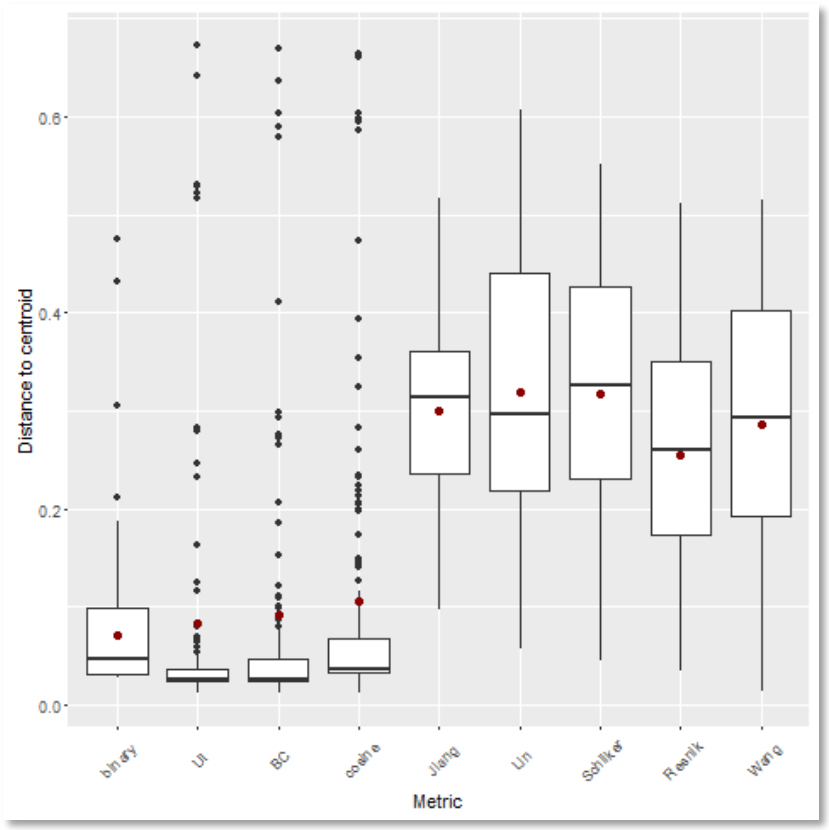


Figure 21: distribution of distances to centroid on the MDS 2 dimensions obtained for each similarity measure.

Next point to evaluate are the clustering results. On Figure 23 we can observe the number of clusters obtained with each measure. The number of clusters in this dataset is high for every similarity measure, being the maximum number of clusters of 30 every measure returned between 28 and 29 clusters, except the Bray-Curtis measure that returned 22 clusters. We have observed in other datasets that functional similarity measures tend to create more single-element clusters than semantic similarity measures. Also, functional similarity measure show the highest maximum number of elements per cluster (Figure 22). The binary measure is the most extreme case tending to create a big cluster and many single-element clusters. The other functional similarity measures show more reasonable results, even if they tend to create bigger clusters than semantic similarity measures.

Finally, for a detailed analysis we selected one similarity measure per approach: (1) Resnik for the IC-based measures, (2) Wang for the ancestors-based measures and (3) cosine for the shared genes measures, being the most coherent measure with the state of the art. When comparing the GO terms that were selected as cluster representatives we can observe in the Venn diagram in Figure 24 that the cosine measure is the most divergent measure. There are several terms selected by Wang and Resnik that are missed by the cosine measure. It is important to note also that the cosine measure has a high number of terms not shared with Wang or Resnik. It is clear that the cosine measure is missing the information in the ontology hierarchy, but the information contained in the shared annotated genes adds information that is being missed by ancestors and IC-based measures. A further evaluation of how informative clusters are will be required.

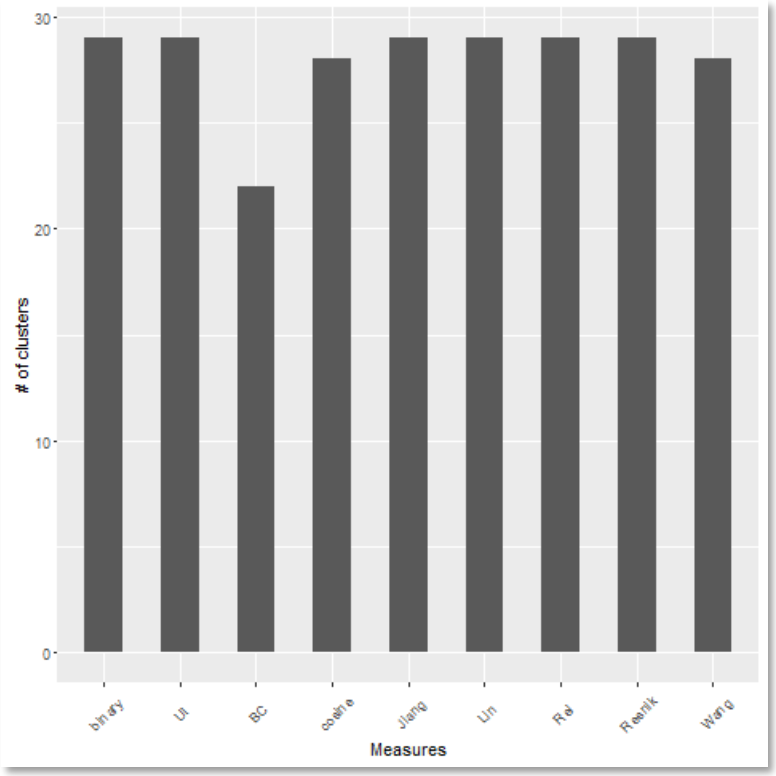


Figure 23: Number of clusters obtained with each similarity measure.

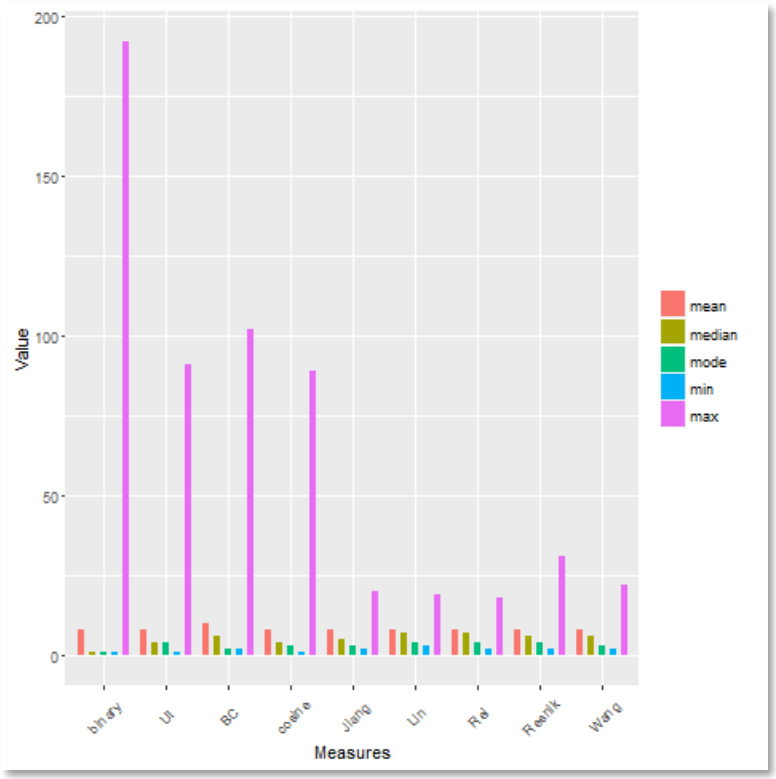


Figure 22: descriptive statistics measure for the number of elements per cluster for each similarity measure.

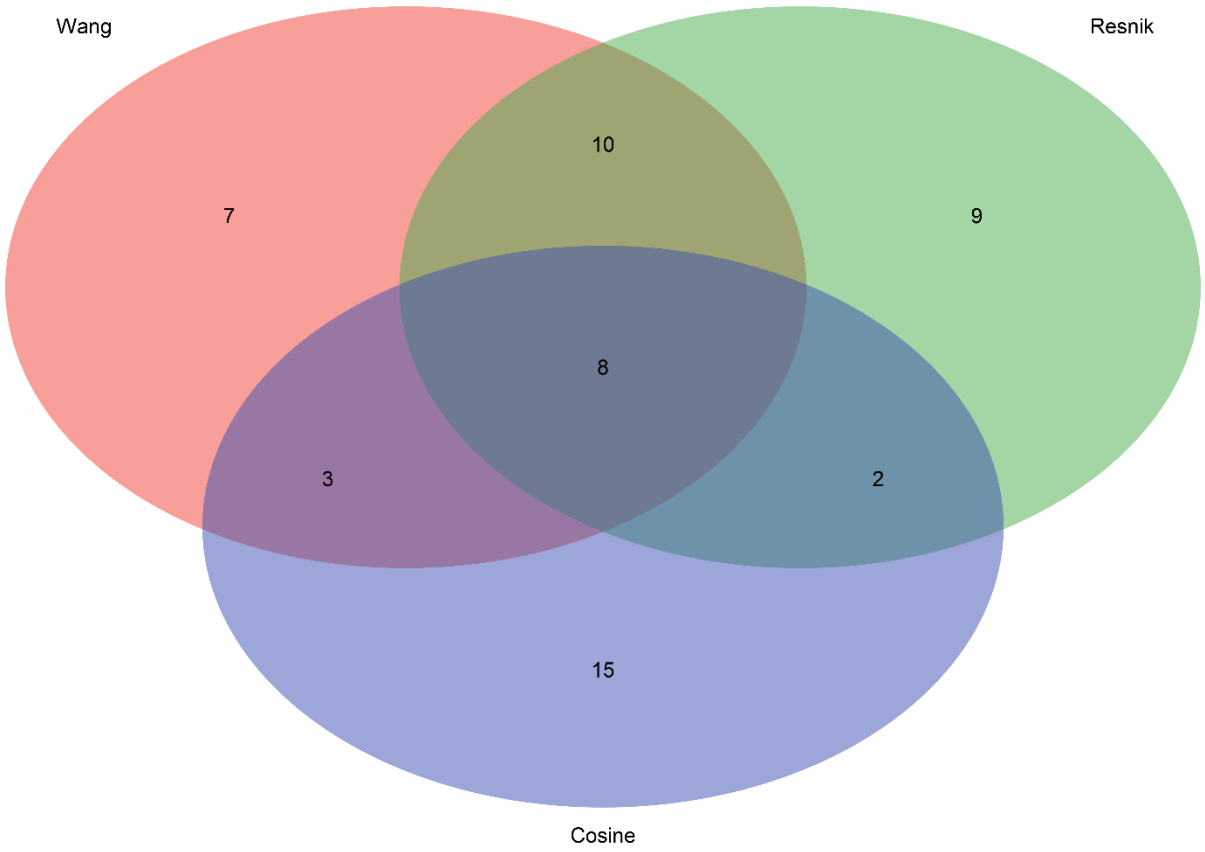
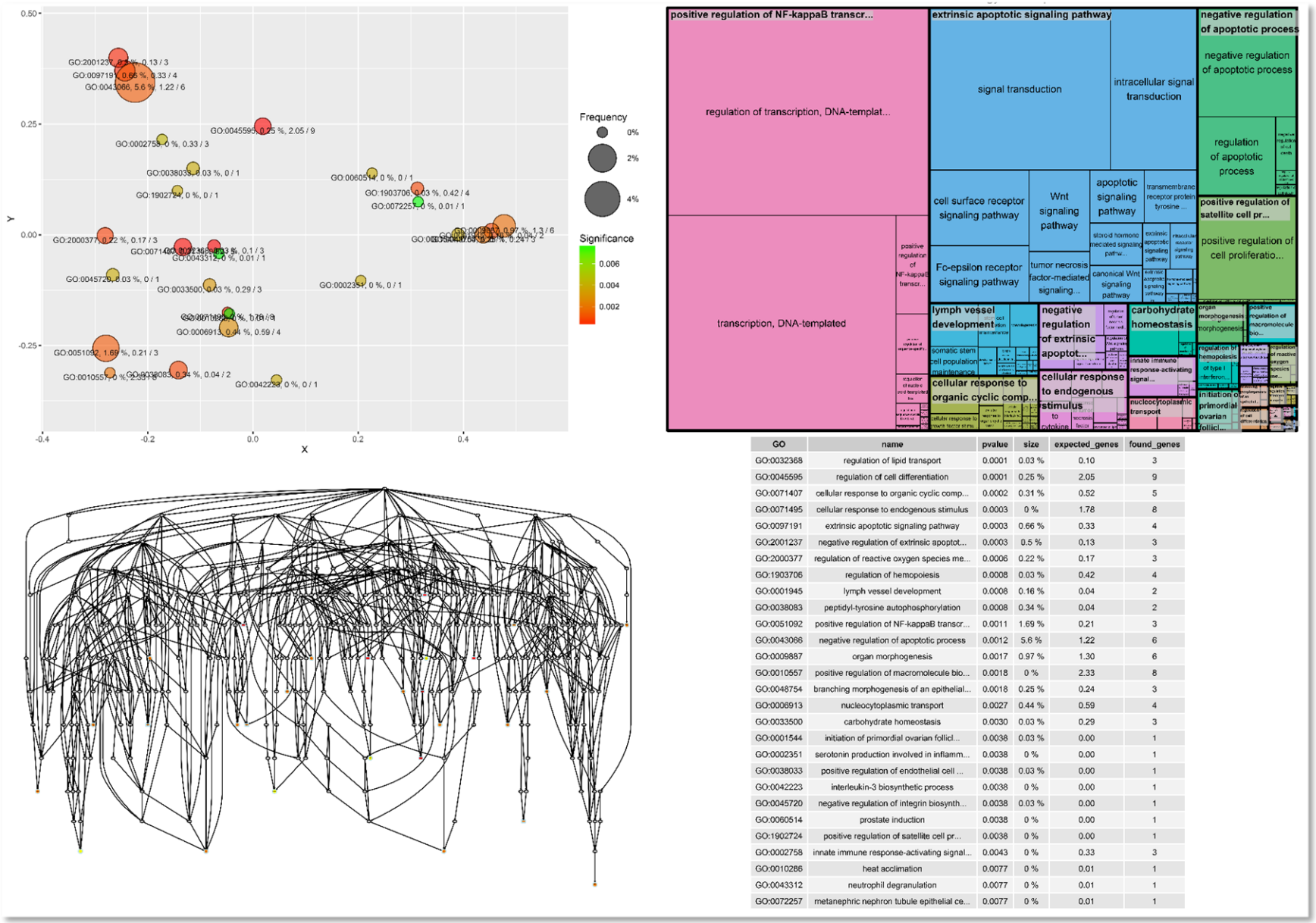


Figure 24: Venn diagram comparing the cluster representatives for the similarity measures of Wang, Resnik and cosine.

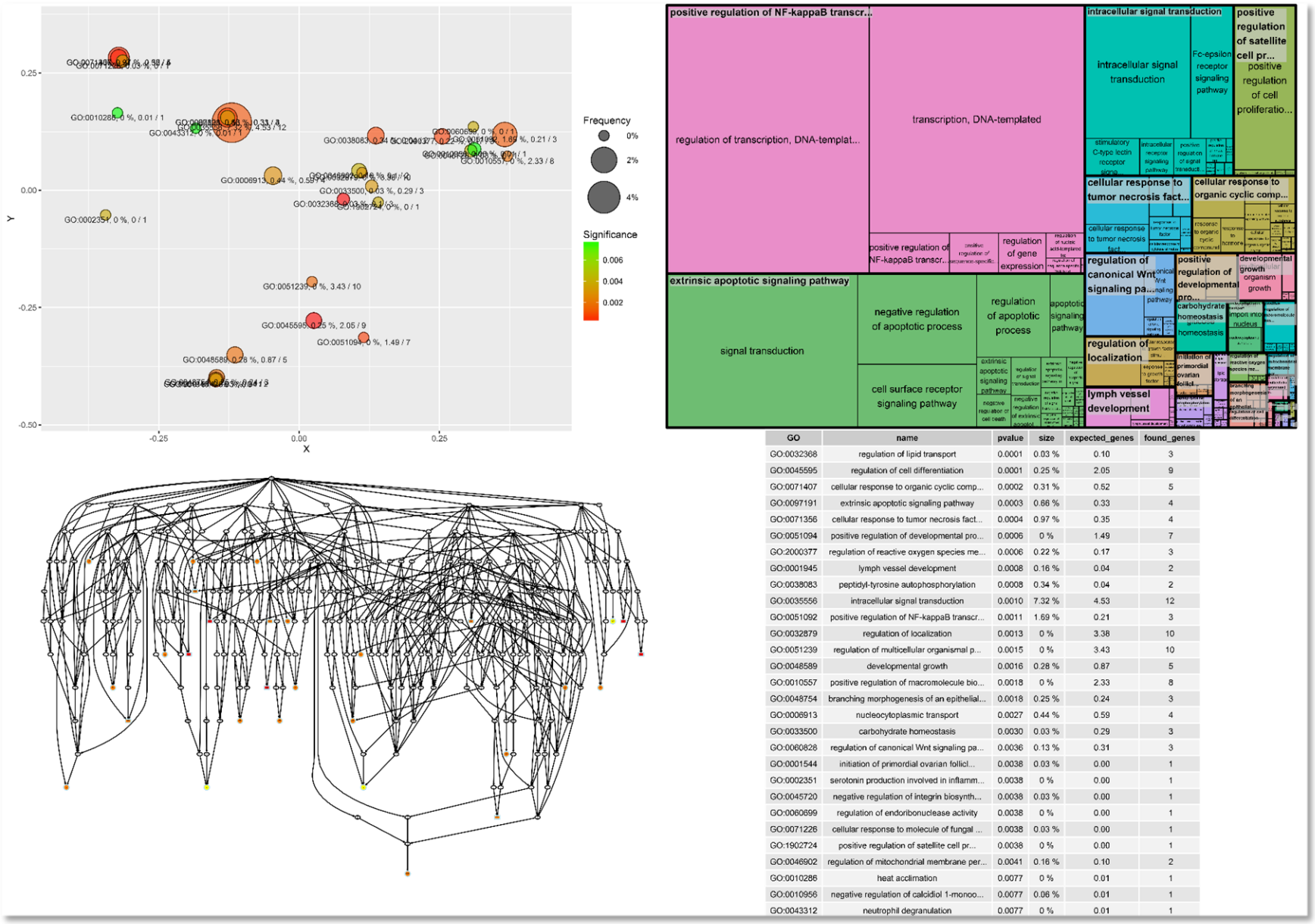
The graphical results for the three methods can be compared in Figure 25, Figure 26 and Figure 27. We can evaluate the composition of clusters between methods. The first difference is the biggest cluster in the treemap (being the size determined by its frequency of annotation); all three measures cluster together the terms GO:0006355 (regulation of transcription, DNA-templated) and GO:0006351 (transcription, DNA-templated) which seems correct as they are connected by a "regulates" relation in the ontology, they have annotation frequencies in UniProt-GOA of 28.16% and 25.47% and they are associated to 900 and 841 genes respectively sharing 503 of those genes. Now, the results for Resnik and Wang group these two terms under the representative GO:0051092 (positive regulation of NF-kappaB transcription factor activity) which as its name indicates is much more specific than the previous being associated to just 54 genes, a frequency of annotation of 1.6%. This is precisely one of the objectives of this methodology: given a set of significantly enriched terms group those more general and less informative under a more specific term. Unfortunately, the cosine measure does not make this grouping. It creates a big cluster with just the two initial terms GO:0006355 and GO:0006351; and a smaller cluster with GO:0051092. If we analyze in detail the cluster represented by GO:0051092 for all three measures the cluster size is of 10, 15 and 5 for Wang, Resnik and cosine respectively; Wang and Resnik seem quite coherent as they share 7 terms apart from the representative term, while cosine only shares 1 term with the other measures. The shared term between all measures is an ancestor of the representative term. The other 3 terms grouped uniquely by the cosine measure are related to the cluster representative with 1, 2 and 8 shared genes and they have no direct relation in the ontology with GO:0051092. See the details in Table 3.

To sum up, Wang and Resnik results tend to be similar, having more differences in the clustering of smaller terms. Cosine measure fails to correctly determine similarity between terms of very different size (i.e.: big terms will not be close to any small term). This issue could be improved by normalizing the cosine similarity measure with the term size. Nevertheless, the cosine measure is identifying relations between terms only backed by the shared genes between terms and these relations are missed



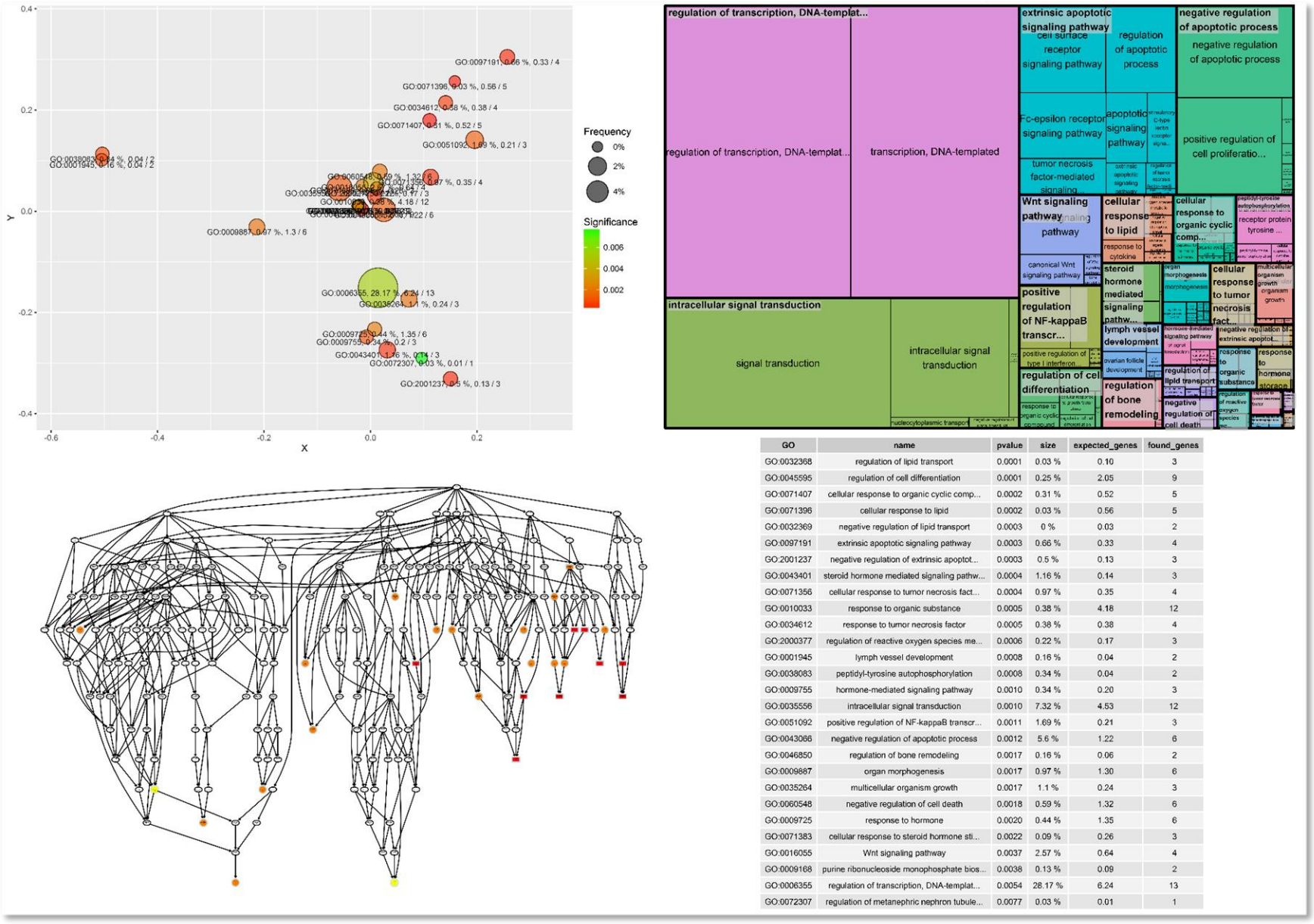
GO	name	pvalue	size	expected_genes	found_genes
GO:0032368	regulation of lipid transport	0.0001	0.03 %	0.10	3
GO:0045595	regulation of cell differentiation	0.0001	0.25 %	2.05	9
GO:0071407	cellular response to organic cyclic comp...	0.0002	0.31 %	0.52	5
GO:0071495	cellular response to endogenous stimulus	0.0003	0 %	1.78	8
GO:0097191	extrinsic apoptotic signaling pathway	0.0003	0.66 %	0.33	4
GO:2001237	negative regulation of extrinsic apoptot...	0.0003	0.5 %	0.13	3
GO:2000377	regulation of reactive oxygen species me...	0.0006	0.22 %	0.17	3
GO:1903706	regulation of hemopoiesis	0.0008	0.03 %	0.42	4
GO:0001945	lymph vessel development	0.0008	0.16 %	0.04	2
GO:0038083	peptidyl-tyrosine autophosphorylation	0.0008	0.34 %	0.04	2
GO:0051092	positive regulation of NF-kappaB transcr...	0.0011	1.69 %	0.21	3
GO:0043066	negative regulation of apoptotic process	0.0012	5.6 %	1.22	6
GO:0009887	organ morphogenesis	0.0017	0.97 %	1.30	6
GO:0010557	positive regulation of macromolecule bio...	0.0018	0 %	2.33	8
GO:0048754	branching morphogenesis of an epithelial...	0.0018	0.25 %	0.24	3
GO:0006913	nucleocytoplasmic transport	0.0027	0.44 %	0.59	4
GO:0033500	carbohydrate homeostasis	0.0030	0.03 %	0.29	3
GO:0001544	initiation of primordial ovarian follicl...	0.0038	0.03 %	0.00	1
GO:0002351	serotonin production involved in inflamm...	0.0038	0 %	0.00	1
GO:0038033	positive regulation of endothelial cell ...	0.0038	0.03 %	0.00	1
GO:0042223	interleukin-3 biosynthetic process	0.0038	0 %	0.00	1
GO:0045720	negative regulation of integrin biosynth...	0.0038	0.03 %	0.00	1
GO:0060514	prostate induction	0.0038	0 %	0.00	1
GO:1902724	positive regulation of satellite cell pr...	0.0038	0 %	0.00	1
GO:0002758	innate immune response-activating signal...	0.0043	0 %	0.33	3
GO:0010286	heat acclimation	0.0077	0 %	0.01	1
GO:0043312	neutrophil degranulation	0.0077	0 %	0.01	1
GO:0072257	metanephric nephron tubule epithelial ce...	0.0077	0 %	0.01	1

Figure 25: Results for those genes selected by MCIA after GO terms enrichment, clustering and MDS using the Wang semantic similarity measure.



GO	name	pvalue	size	expected_genes	found_genes
GO:0032368	regulation of lipid transport	0.0001	0.03 %	0.10	3
GO:0045595	regulation of cell differentiation	0.0001	0.25 %	2.05	9
GO:0071407	cellular response to organic cyclic comp...	0.0002	0.31 %	0.52	5
GO:0097191	extrinsic apoptotic signaling pathway	0.0003	0.66 %	0.33	4
GO:0071356	cellular response to tumor necrosis fact...	0.0004	0.97 %	0.35	4
GO:0051094	positive regulation of developmental pro...	0.0006	0 %	1.49	7
GO:2000377	regulation of reactive oxygen species me...	0.0006	0.22 %	0.17	3
GO:0001945	lymph vessel development	0.0008	0.16 %	0.04	2
GO:0038083	peptidyl-tyrosine autophosphorylation	0.0008	0.34 %	0.04	2
GO:0035556	intracellular signal transduction	0.0010	7.32 %	4.53	12
GO:0051092	positive regulation of NF-kappaB transcr...	0.0011	1.69 %	0.21	3
GO:0032879	regulation of localization	0.0013	0 %	3.38	10
GO:0051239	regulation of multicellular organismal p...	0.0015	0 %	3.43	10
GO:0048589	developmental growth	0.0016	0.28 %	0.87	5
GO:0010557	positive regulation of macromolecule bio...	0.0018	0 %	2.33	8
GO:0048754	branching morphogenesis of an epithelial...	0.0018	0.25 %	0.24	3
GO:0006913	nucleocytoplasmic transport	0.0027	0.44 %	0.59	4
GO:0033500	carbohydrate homeostasis	0.0030	0.03 %	0.29	3
GO:0060828	regulation of canonical Wnt signaling pa...	0.0036	0.13 %	0.31	3
GO:0001544	initiation of primordial ovarian follicl...	0.0038	0.03 %	0.00	1
GO:0002351	serotonin production involved in inflamm...	0.0038	0 %	0.00	1
GO:0045720	negative regulation of integrin biosynth...	0.0038	0.03 %	0.00	1
GO:0060699	regulation of endoribonuclease activity	0.0038	0 %	0.00	1
GO:0071226	cellular response to molecule of fungal ...	0.0038	0.03 %	0.00	1
GO:1902724	positive regulation of satellite cell pr...	0.0038	0 %	0.00	1
GO:0046902	regulation of mitochondrial membrane per...	0.0041	0.16 %	0.10	2
GO:0010286	heat acclimation	0.0077	0 %	0.01	1
GO:0010956	negative regulation of calcidiol 1-monoo...	0.0077	0.06 %	0.01	1
GO:0043312	neutrophil degranulation	0.0077	0 %	0.01	1

Figure 26: Results for those genes selected by MCIA after biological process GO terms enrichment, clustering and MDS using the Resnik semantic similarity measure.



GO	name	pvalue	size	expected_genes	found_genes
GO:0032368	regulation of lipid transport	0.0001	0.03 %	0.10	3
GO:0045595	regulation of cell differentiation	0.0001	0.25 %	2.05	9
GO:0071407	cellular response to organic cyclic comp...	0.0002	0.31 %	0.52	5
GO:0071396	cellular response to lipid	0.0002	0.03 %	0.56	5
GO:0032369	negative regulation of lipid transport	0.0003	0 %	0.03	2
GO:0097191	extrinsic apoptotic signaling pathway	0.0003	0.66 %	0.33	4
GO:2001237	negative regulation of extrinsic apoptot...	0.0003	0.5 %	0.13	3
GO:0043401	steroid hormone mediated signaling pathw...	0.0004	1.16 %	0.14	3
GO:0071356	cellular response to tumor necrosis fact...	0.0004	0.97 %	0.35	4
GO:0010033	response to organic substance	0.0005	0.38 %	4.18	12
GO:0034612	response to tumor necrosis factor	0.0005	0.38 %	0.38	4
GO:2000377	regulation of reactive oxygen species me...	0.0006	0.22 %	0.17	3
GO:0001945	lymph vessel development	0.0008	0.16 %	0.04	2
GO:0038083	peptidyl-tyrosine autophosphorylation	0.0008	0.34 %	0.04	2
GO:0009755	hormone-mediated signaling pathway	0.0010	0.34 %	0.20	3
GO:0035556	intracellular signal transduction	0.0010	7.32 %	4.53	12
GO:0051092	positive regulation of NF-kappaB transcr...	0.0011	1.69 %	0.21	3
GO:0043066	negative regulation of apoptotic process	0.0012	5.6 %	1.22	6
GO:0046850	regulation of bone remodeling	0.0017	0.16 %	0.06	2
GO:0009887	organ morphogenesis	0.0017	0.97 %	1.30	6
GO:0035264	multicellular organism growth	0.0017	1.1 %	0.24	3
GO:0060548	negative regulation of cell death	0.0018	0.59 %	1.32	6
GO:0009725	response to hormone	0.0020	0.44 %	1.35	6
GO:0071383	cellular response to steroid hormone sti...	0.0022	0.09 %	0.26	3
GO:0016055	Wnt signaling pathway	0.0037	2.57 %	0.64	4
GO:0009168	purine ribonucleoside monophosphate bios...	0.0038	0.13 %	0.09	2
GO:0006355	regulation of transcription, DNA-templat...	0.0054	28.17 %	6.24	13
GO:0072307	regulation of metanephric nephron tubule...	0.0077	0.03 %	0.01	1

Figure 27: Results for those genes selected by MCIA after biological process GO terms enrichment, clustering and MDS using the cosine functional similarity measure.

Discussion

TCGAome is a functional pipeline to analyze multiomics datasets downloaded from the TCGA public dataset with state of the art multivariate methods. It also combines these results with the existing biological knowledge. The pipeline has, nevertheless several limitations: (1) the analysis of results will only be possible to apply to biological entities univocally mapped to genes, for instance we won't be able to apply this approach to metabolites data and for variants we will need to aggregate the data into genes, (2) visualizations are static and many times don't allow the user to drill down into the details needing to go to data tables, (3) we are limited to TCGA datasets and (4) we only make use of the Gene Ontology while many other biomedical ontologies exist. Each of the previous limitations would need further development of tool functionality and methodology.

The differences observed between the results of MCIA and sPLS could not be explained without a deep analysis of the results. This understanding may lead to an interesting combination of both methods to target different biological events. Also, it would be interesting to evaluate both methods on other types of data, like metabolome vs expression. Regarding the test dataset it is an inconvenient having so few protein expression variables as compared to the number of gene expression variables.

We compared several semantic and functional similarity measures for the pairwise comparison of GO terms. Functional similarity measures making use only of the shared genes in Uniprot-GOA seem to separate GO terms with different criteria than other semantic similarity measures, which leaves an open question: could we improve the existing state of the art combining these two approaches? Also, the functional approaches analyzed here fail to correctly determine similarity between terms of very different sizes (i.e.: frequency of annotation). A normalization of the functional similarity with the size of terms might improve the results. Finally, we can conclude that the ontology structure is an important piece of information to determine similarity as is shown by Wang results. Furthermore, this approach is applicable to other biomedical ontologies that might not have such a complete annotation set as Uniprot-GOA.

The approach for the analysis of gene lists is "inherited" from the analysis of differential gene expression sets and it is thus missing the covariance information that both MCIA and sPLS provide. Using this information to map genes and their covariances into the Gene Ontology space remains a challenge. Integrating pathways enrichment might also add useful information.

Summing up, an end to end tool that generalizes the previous problem being able to analyze any number of datasets of any type, using any ontology to map the existing knowledge on a given domain would be a great achievement, but also a great interdisciplinary challenge.

References

1. Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! *Genome Biol.* **12,** 125 (2011).
2. Ivan Merelli, Horacio Pérez-Sánchez, Sandra Gesing, and D. D. Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives. at <http://www.hindawi.com/journals/bmri/2014/134023/>
3. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).
4. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **135,** 0–9 (2012).
5. Material, S. O., Web, S., Press, H., York, N. & Nw, A. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–40 (2004).
6. Gomez-Cabrero, D. *et al.* Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* **8 Suppl 2,** I1 (2014).
7. Culhane, A. C., Perrière, G. & Higgins, D. G. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics* **4,** 59 (2003).
8. Lê Cao, K.-A., Martin, P. G., Robert-Granié, C. & Besse, P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* **10,** 34 (2009).
9. Meng, C., Kuster, B., Culhane, A. C. & Gholami, A. M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **15,** 162 (2014).
10. Piwowar, M. & Jurkowski, W. ONION: Functional Approach for Integration of Lipidomics and Transcriptomics Data. *PLoS One* **10,** e0128854 (2015).
11. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6,** e21800 (2011).
12. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21,** 3448–9 (2005).
13. Isserlin, R., Merico, D., Voisin, V. & Bader, G. D. Enrichment Map - a Cytoscape app to visualize and explore OMICs pathway enrichment results. *F1000Research* **3,** 141 (2014).
14. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–58 (2013).
15. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6,** 813–23 (2006).
16. Samur, M. K. RTCGAToolbox: a new tool for exporting TCGA Firehose data. *PLoS One* **9,** e106397 (2014).
17. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4,** 1184–91 (2009).
18. González, I., Cao, K.-A. L., Davis, M. J. & Déjean, S. Visualising associations between paired 'omics' data sets. *BioData Min.* **5,** 19 (2012).
19. Alexander Budanitsky, G. H. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. (2001). at <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.545>
20. Slimani, T. Description and Evaluation of Semantic Similarity Measures Approaches. *Int. J. Comput. Appl.* **80,** 25–33 (2013).
21. Pedersen, T., Pakhomov, S. V. S., Patwardhan, S. & Chute, C. G. Measures of semantic similarity and relatedness in the biomedical domain. *J. Biomed. Inform.* **40,** 288–99 (2007).
22. Pesquita, C., Faria, D., Falcão, A. O., Lord, P. & Couto, F. M. Semantic Similarity in Biomedical Ontologies. *PLoS Comput. Biol.* **5,** e1000443 (2009).
23. Choi, S., Cha, S. & Tappert, C. A Survey of Binary Similarity and Distance Measures. *J. Syst. Cybern. Informatics* **8,** 43 – 48 (2010).
24. Alexa, A. & Rahnenfuhrer, J. topGO: Enrichment analysis for Gene Ontology. (2010).
25. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26,** 976–8 (2010).
26. Zhao, M., Sun, J. & Zhao, Z. TSGene: a web resource for tumor suppressor genes. *Nucleic Acids Res.* **41,** D970–6 (2013).

Annex A

Table 3: clusters for GO term GO:0051092 with the different measures; in yellow the cluster representatives; in gray the shared GO terms between all measures.

Measure	GO	Name	Found genes	Expected genes	p-value	size
cosine functional similarity	GO:0051092	positive regulation of NF-kappaB transcr...	3	0.21	0.00113	1.69%
	GO:0051091	positive regulation of sequence-specific...	3	0.33	0.00428	1.03%
	GO:0002218	activation of innate immune response	3	0.34	0.00471	0.09%
	GO:0032481	positive regulation of type I interferon...	2	0.13	0.00745	0.72%
	GO:0002238	response to molecule of fungal origin	1	0.01	0.00767	0.03%
	GO:0006355	regulation of transcription, DNA-templat...	13	6.24	0.00543	28.17%
	GO:0032386	regulation of intracellular transport	4	0.79	0.00757	0.06%
	GO:0006351	transcription, DNA-templated	13	6.58	0.00852	25.48%
	GO:0031328	positive regulation of cellular biosynth...	7	2.45	0.00973	0.03%
Wang semantic similarity	GO:0051092	positive regulation of NF-kappaB transcr...	3	0.21	0.00113	1.69%
	GO:0051090	regulation of sequence-specific DNA bind...	4	0.53	0.00189	0.28%
	GO:0051091	positive regulation of sequence-specific...	3	0.33	0.00428	1.03%
	GO:0006355	regulation of transcription, DNA-templat...	13	6.24	0.00543	28.17%
	GO:1903506	regulation of nucleic acid-templated tra...	13	6.29	0.00582	0.50%
	GO:2001141	regulation of RNA biosynthetic process	13	6.31	0.00597	0.03%
	GO:0010468	regulation of gene expression	14	7.21	0.00706	1.00%
	GO:0006351	transcription, DNA-templated	13	6.58	0.00852	25.48%
	GO:0097659	nucleic acid-templated transcription	13	6.63	0.00908	0.00%
	GO:0034654	nucleobase-containing compound biosynthe...	14	7.48	0.00983	0.00%
Resnik semantic similarity	GO:0051092	positive regulation of NF-kappaB transcr...	3	0.21	0.00113	1.69%
	GO:0051090	regulation of sequence-specific DNA bind...	4	0.53	0.00189	0.28%
	GO:0060699	regulation of endoribonuclease activity	1	0	0.00384	0.00%
	GO:1902380	positive regulation of endoribonuclease ...	1	0	0.00384	0.03%
	GO:0051091	positive regulation of sequence-specific...	3	0.33	0.00428	1.03%
	GO:0006355	regulation of transcription, DNA-templat...	13	6.24	0.00543	28.17%
	GO:1903506	regulation of nucleic acid-templated tra...	13	6.29	0.00582	0.50%
	GO:2001141	regulation of RNA biosynthetic process	13	6.31	0.00597	0.03%
	GO:0051252	regulation of RNA metabolic process	13	6.44	0.00717	0.00%
	GO:0010956	negative regulation of calcidiol 1-monoo...	1	0.01	0.00767	0.06%
	GO:0060558	regulation of calcidiol 1-monooxygenase ...	1	0.01	0.00767	0.00%
	GO:0060700	regulation of ribonuclease activity	1	0.01	0.00767	0.03%
	GO:2000630	positive regulation of miRNA metabolic p...	1	0.01	0.00767	0.03%
	GO:0006351	transcription, DNA-templated	13	6.58	0.00852	25.48%
	GO:0097659	nucleic acid-templated transcription	13	6.63	0.00908	0.00%