coGSEA : comparative Gene Set Enrichement Analysis

Maxime Borry

May 2017

Bioinformatics Master 1 Internship Report

Institut Pasteur-Bioinformatics and Biostatistics HUB

28 Rue du Dr Roux, 75015 Paris

Supervisor 1: Natalia Pietrosemoli, PhD

Supervisor 2: Vincent Guillemot, PhD

Acknowledgment

I would like to thank the following persons :

Natalia Pietrosemoli for her warm welcome and her consistent help throughout this internship.

Vincent Guillemot for his warm welcome and his help in statistics and document formatting.

Hugo Varet for his advices on the Shiny framework.

All the members of the Bioinformatics and Biostatistics HUB of Institut Pasteur for having me during these three months.

M.B.

Acronyms

DEG Differentially expressed gene

EGS Enriched gene-set in differentially expressed genes

FCS Functionnal class scoring

GSE Gene set enrichment

GUI Graphical user interface

ORA Over representation analysis

PT Pathway topology

SEGS Significantly enriched gene-set in differentially expressed genes

SS Single sample

Contents

coGSEA is a tool I developed to combine different gene-set enrichment (GSE) methods. In this report, I explain the context of GSE methods and their development, together with the theory behind coGSEA.

1. *Introduction*

When Francis Crick introduced the idea of the central dogma of molecular biology: “DNA makes RNA and RNA makes protein”, this revolutionized the way biologists thought of life. Nowadays, this central dogma is at the basis of different fields in the study of nucleic acids: genomics for DNA, transcriptomics for RNA, proteomics for proteins, among many other “-omics”. These fields have different focuses: while genomics looks for the presence of mutations, polymorphisms, and insertions, the field of transcriptomics is more focused on studying the number of transcript’s copies that a gene can give rise to. Using micro-array technologies and Next Generation Sequencing techniques, the main goal of transcriptomics studies is to measure the level of expression of the different transcripts for each gene, in different biological conditions, to reflect changes observed at the phenotypic level at the molecular level. As more experimental evidences about genes and interactions of their products, and their effect on the cell, became available, genes were assembled into biological pathways forming sets of genes. And with it, came the idea of analyzing lists of differentially expressed genes (DEGs) derived from different biological conditions, according to how many of such DEGs were present in these sets of genes. In other words, to assess the enrichment of these sets of genes in DEGs, and compare it across different biological conditions.

1.1 Gene Set Enrichment

Gene Set Enrichment (GSE) studies introduced the notion of defining sets of genes to group genes [21] using other criteria than the ones used in the classical "knowledge based driven pathways analysis” where only set of genes, or gene-sets derived from biological pathways were studied [18]. According to GSE, genes could also be grouped based on the gene ontology such as grouping by the cellular location of their products or by their biological functions. Therefore the definition of gene-sets became: **any** a priori classification of genes into biologically relevant groups.

However, several challenges arise while performing GSE studies such as defining gene-sets and the genes involved in it, calculating the enrichment of gene-sets in DEG, comparing across different conditions, and summarizing the results with relevant metrics.

1.2 GSE Databases

While most databases for functional annotation of genes were originally designed for pathway oriented analyses such as KEGG [10], Reactome [5], and IPA [12], the Molecular Signature Database (MSigDB) [21] allows for a more generic grouping of genes based on other criteria, including not only metabolic pathways, but also groups of genes that share a common biological function, a chromosomal location, or a common regulation. Together with the database, the authors of MSigDB proposed one of the first GSE method named Gene Set Enrichment Analysis (GSEA) for retrieving significantly enriched gene sets (SEGS) in DEG across different biological conditions [21].

1.3 GSE methods

Over the last twelve years, since GSEA, many tools and methods have been developed to perform GSE analyses. GSE methods can be broadly classified into 4 categories [11][22]. First, **Over-Representation Analysis (ORA)** methods evaluate the proportion of significant DEGs (at a given threshold α) belonging to different gene-sets. However, it has been shown that, because genes are classified in a binary way as differentially expressed or not according to α, the choice of this threshold strongly affects the outcome of the analysis[17]. Thus, to avoid user defined thresholds affecting the results, a second type of GSE method was introduced with GSEA: the **Functional Class Scoring (FCS)** methods. These methods generally follow the same procedure: a given statistic, different according to the method, is first calculated at the feature/gene level. Then, all the statistics for all the genes belonging to a same gene-set are aggregated into a gene-set level statistic. Different gene-set level statistics can be used by different methods. Finally, the statistical significance of the gene-set level statistic is assessed. To do so, two different null hypothesis can be evaluated, the **self-contained null hypothesis H0(self)** that states: *"no genes in the gene set of interest are differentially expressed"*, or the **competitive null hypothesis H0(comp)**: *"genes in the gene set of interest are at most as often differentially expressed as the genes not belonging to the gene set of interest"*[7]. In other words, in the competitive null hypothesis testing, genes in a chosen gene-set are compared to genes that do not belong to this gene-set, whereas in the self contained null hypothesis, genes that are not in the gene-set are not used. Self-contained null hypotheses are therefore rejected more often than competitive null hypothesis. The drawback of the self-contained null hypotheses is that in certain cases, too many gene-sets will be found significant, therefore diluting the relevance of the biological information. A different approach is taken by the **Single-Sample (SS)** methods: a gene-set score is calculated for each individual sample from the observed gene expression, which allows for easier analysis of complex experimental designs [22]. The last category of GSE methods belongs to the **Pathway Topology (PT)** based approaches. PT methods use the pathway topology informations to calculate gene level statistics.

In this work, 10 different GSE methods, belonging to the ORA, FCS, and SS class, were tested and compared to assess the underlying phenotypic differences on a human micro-array dataset using 292 gene-sets belonging to the Human KEGG signaling and diseases gene-set collection. These 10 methods, together with two additional ones were assembled in a common interface and distributed as an R package, as well as a Graphical User Interface (GUI). Our results show that different methods produce different results, and that a combination of all methods seems to perform better than individual methods in terms of sensitivity and in comparison to biologically verified results. These findings are in agreement with the results of Alhamdoosh et al. [1].

2. *Material and Methods*

2.1 Biological data

The chosen benchmark dataset consists of Affymetrix Human Genome U133 Plus 2.0 micro-array gene expression profiles of astrocytoma, the most common neuroepithelial cancer. It consists of 4 normal tissues samples and 17 samples of tumoral tissue at different clinical stages as defined by the World Health Organization tumor grading system. This dataset was first described by Liu et al. [14] and deposited on the Gene Expression Omnibus database [6] under the GSE19728 accession number. For the purpose of this analysis, the 4 normal/control samples and only the 5 stage IV samples, corresponding to the most advanced stage of the disease, were analyzed. The raw data was provided as CEL files, and loaded in the R environment using the *affy* package. The probesets were mapped onto genes using the *hgu133plus2.db* package. The expression data were $\log$ transformed, and normalized using the `rma()` function from the *affy* package. Finally the dispersion, that is how the variance deviates from the mean in the negative binomial distribution, of the samples was computed using the `estimateDisp()` function from the *edgeR* package [19].

2.2 Gene set Collection Database

The chosen gene set collection database for this work is the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways collection for Homo sapiens [10]. This gene set collection was retrieved using the `kegg.gsets()` function provided with the implementation of GAGE [15] GSE method. To this date, it includes 292 gene-sets containing a total of 25096 human genes identified by their `ENTREZ` accession number.

Method	**Package**	**Class**	**H0**
Camera	limma	FCS	competitive
GAGE	gage	FCS	competitive
GlobalTest	globaltest	FCS	self contained
GSVA	GSVA	FCS	competitive
ssGSEA	GSVA	SS	competitive
zscore	GSVA	FCS	not reported
PLAGE	GSVA	FCS	self contained
ORA	eGSEA	ORA	competitive
PADOG	PADOG	$\sim$ FCS	self contained
Roast	limma	FCS	self contained
SAFE	safe	FCS	competitive
SetRank	SetRank	$\sim$ FCS	competitive

Table 1: GSE methods classified according to their null hypothesis, category, and their R package implementation

2.3 GSE Methods

There are 12 GSE methods implemented in the R package I developed (table 1): *Camera* [26], *GAGE* [15], *GlobalTest* [8], *GSVA* [9], *ssGSEA* [2], *zscore* [13], *PLAGE* [24], *ORA*, *PADOG* [23], *Roast* [25], *SAFE* [3], and *SetRank* [20]. Each method calculates the *p* value of each gene-set in the gene-set collection, and ranks them accordingly. For a more detailed review of different methods, see [16]. However, due to the fact that some of these methods cannot be implemented for our dataset (PLAGE), or because of run-time optimization considerations (SetRank), only 10 methods are compared on the astrocytoma dataset. The implementations of 11 of the methods (all but SetRank) were extracted from the *eGSEA* R package. [1].

2.4 Statistical processing

For each of the methods' results, and for each gene-set, the rank r of the gene-set in the gene-set collection was extracted, as well as the associated *p* value p. Let M be the number of methods m, G the number of gene sets g. A $G \times M$ matrix of raw *p* values p was retrieved. *p* values were adjusted across G for multiple hypothesis testing using the Benjamini-Hochberg method, and finally, p values were combined across M using the Fisher's method (equation 2.1) giving a G sized vector with one adjusted and combined p value X_{p_f} per gene-set.

$$X_{p_f} = -2 \sum_{m=1}^{M} \log p_m \tag{2.1}$$

$$\overline{R} = \frac{1}{M} \sum_{m=1}^{M} r_m \tag{2.2}$$

$$logFC = \log Expr_{tumor} - \log Expr_{normal} \tag{2.3}$$

$$\overline{logFC}_{gene-set} = \frac{1}{G} \sum_{g=1}^{G} \log FC_g \tag{2.4}$$

$$S = -\log_{10}(pvalue\ _G) \times \frac{1}{G} \sum_{g=1}^{G} \log FC_g \tag{2.5}$$

It must be noted that some of Fisher method's assumptions are violated:

- The independence of tests, because the results of the methods are coming from the same samples.
- Homogeneity of null hypotheses, because the null hypotheses can be different depending on the method.

Regarding the ranks, a $G \times M$ matrix is calculated as well, with the rank r being defined as the position of a gene-set in the sorted gene-set collection on raw p values, after analysis. Ranks are then combined using equation 2.2 in a G sized vector of one rank $\overline{R}$ per gene-set. The log Fold Change ($\overline{logFC}_{gene-set}$) of each gene-set is computed for each gene in the gene-set by averaging the $\log FC$ of each gene present in the gene-set using equation 2.4 . The gene-level $\log FC$s are calculated as the difference between the expression level of a given gene in two different biological conditions (equation 2.3). $\log FC$ is calculated as performed by *edgeR*, producing a G sized vector of one $\overline{logFC}_{gene-set}$ per gene-set. Another metric is defined as the significance S (equation 2.5) and computed for each gene-set and takes into account both the $\log FC$ and the p value of a gene-set. Finally, for combining the results of the different GSE methods and identifying SEGS in DEGs, a threshold $\alpha = 0.05$ was chosen. All these statistical analyses were performed according to the procedures described in Alhamdoosh et al. [1]. The result is a table exported as a `.csv` file where each row corresponds to a gene-set, and in the columns, all the metrics previously described.

2.5 Comparison of GSE methods and combination graphs

To investigate the relation between the different GSE methods, several statistical methods were applied on the ranks of the gene-sets for each GSE method including hierarchical clustering (euclidean distance, and Ward.D2 clustering) (fig 3), principal component analysis (fig S1, S2), and correlation plots (fig S3). Additionally a `multi-set` plot using the *SuperExactTest* package was produced to visualize the intersection between different methods, with intersections being defined as the SEGS found in common by different methods at threshold α (fig 1). A heatmap was also produced to identify the SEGS at these intersections (fig S4). Finally, a custom scatter-plot (fig 2) was produced using *ggplot2* package to examine the relationship between the average $\log FC$ and the significancy of the resulting gene-sets according to the combination of different GSE methods.

2.6 Package and GUI

The implementation of the methods described above is available in the coGSEA R package `https://github.com/maxibor/coGSEA`. An interactive GUI of coGSEA was designed using the Shiny Framework [4] and offers a simplified interface while improving the results exploration possibilities `https://github.com/maxibor/coGSEA_shiny`.

3. *Results*

In order to validate the results of the individual GSE methods and their combination, I selected a target gene-set that was biologically relevant for the astrocytoma dataset: `hsa:05214 Glioma` [27], comprising many genes related to brain cancers. It should therefore be expected to be ranked as the most SEGS.

Our results show that, regarding the adjusted *p* value, and taking into account all the SEGS found by each individual method, one can find a total of 272 different SEGS, which is the equivalent of 93.2% of the gene-set collection.

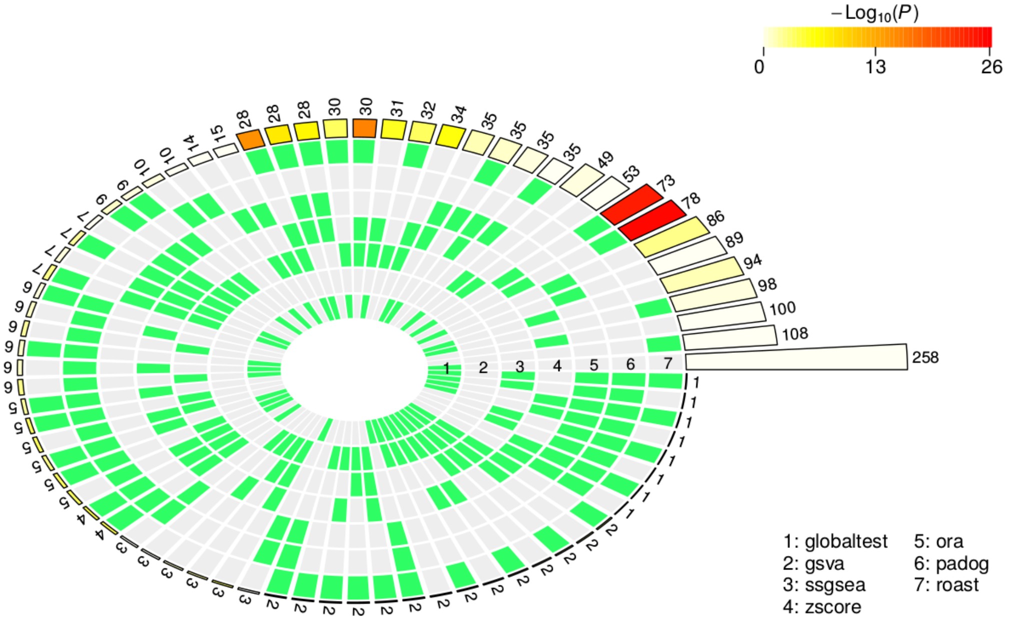

Figure 1: `Multi-set` intersection plot showing the results of the human astrocytoma dataset GSE19728. Each concentric ring represent a GSE method, and each green box in a wheel spoke, an intersection. An intersection is defined as significantly enriched gene sets (SEGS) under $\alpha < 0.05$, found in common by different methods involved in it. Each outer bar represents the size of the intersection (number on top) while the color represents the probability of the intersection happening by chance. Methods not retrieving any significant SEGS are not shown.

Detailed analysis of the SEGS found by each method (fig 1) shows that there are a lot of differences in the results of each method, and therefore a procedure for coming to an agreement on the true SEGS needed to be devised.

The combination of p values was one solution: using the Fisher's method described to combine adjusted p values for each method, with $\alpha = 0.05$, resulted in a total number of 88 different SEGS which is equivalent to 30% of the gene-set collection. However, still having 30% of a gene-set collection significantly enriched is excessive. Because yielding too many SEGS will not be very biologically informative, after the combined p values, the rank of the different enriched gene sets (EGS) was examined. The target gene-set `hsa:05214` was ranked as first by coGSEA, whereas in the ranking by each individual method, it was found in the top 5 only by two methods (gage and roast) and by none of the methods in the top 3 (table S1). Another way to investigate which SEGS are the most significant is to look at the combination of combined adjusted p value, $\log FC$ and S as defined in equation 2.5 (fig 2). Looking at the highest significance scores on the top-right of the plot, `hsa:05214` is also found as the most differentially expressed gene-set.

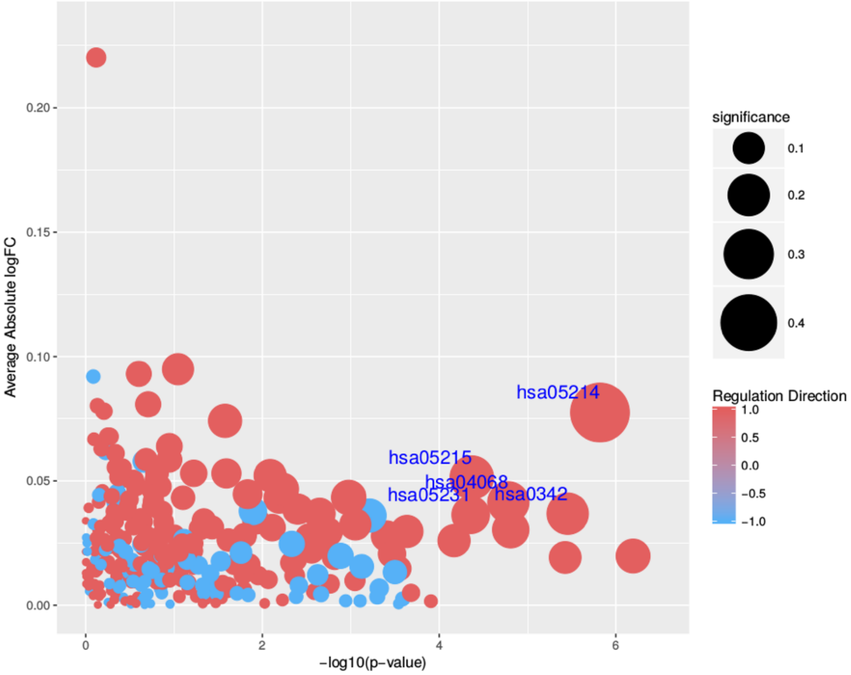

Figure 2: Gene-sets retrieved for the human astrocytoma dataset GSE19728. On the x-axis the -log10(combined adjusted p value) for each EGS, on the y-axis, the average $\log FC$ for each EGS. The bubble size is the significance S defined in equation 2.5 . The bubble color displays the direction of the $\log FC$. The top 5 most significant enriched gene-sets have their KEGG accession number diplayed in blue.

Finally, comparing the different methods results on the ranks of the gene-sets does not show a clustering pattern that reflects a grouping by the method's class nor by the null hypothesis (fig 3).

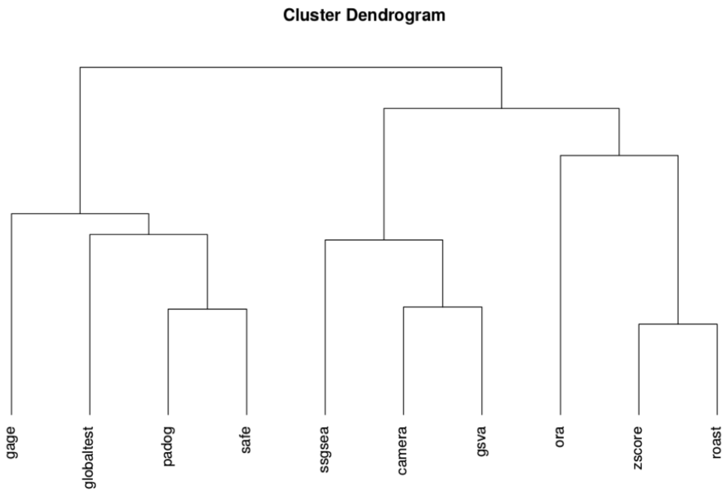

Figure 3: Clustering of the GSE methods according to the gene-sets ranks with an euclidean distance, and the ward.D2 method

4. *Discussion and Conclusion*

I have shown that overall, coGSEA performs better on the studied dataset: only the combination of 10 methods allowed to retrieve the glioma target gene-set, whereas the individual GSE methods failed to retrieve it in the top 3 significant gene-sets (table S1) and individually produced very different results. Despite the violation of some of Fisher's assumptions for combining *p* values, coGSEA's conclusions are in agreement with the biologically validated results. Another way to investigate the validity of the *p* value combination is to look at the correlation between gene-set ranks and combined *p* values: there seems to be a strong correlation between average ranks and combined adjusted *p* values, with a Spearman correlation ρ of 0.76, which validates the *p* value combination method. Nevertheless, there exist other *p* values combination methods that could be used with assumptions more compatible with this problem. They are however out of the scope of this work. The implementation of coGSEA as a documented R package and as a GUI is available to the scientific community in a reproducible and flexible way, and is ready to use. It has been tested with RNA-seq data as well as with micro-array data, and could potentially be used for proteomics data, however, this hasn't been tested.

Bibliography

[1] Alhamdoosh, M., Ng, M., Wilson, N. J., Sheridan, J. M., Huynh, H., Wilson, M. J., and Ritchie, M. E. (2017). Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics*, 33(3):414–424.

[2] Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., Gupta, P. B., Wadlow, R. C., Le, H., Hoersch, S., Wittner, B. S., Ramaswamy, S., Livingston, D. M., Sabatini, D. M., Meyerson, M., Thomas, R. K., Lander, E. S., Mesirov, J. P., Root, D. E., Gilliland, D. G., Jacks, T., and Hahn, W. C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108–112.

[3] Barry, W. T., Nobel, A. B., and Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949.

[4] Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2017). *shiny: Web Application Framework for R*.

[5] Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., and Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Database issue):D691–D697.

[6] Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207.

[7] Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.

[8] Goeman, J. J., Geer, S. A. v. d., Kort, F. d., and Houwelingen, H. C. v. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, (20/1).

[9] Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics*, 14:7.

[10] Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

[11] Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Computational Biology*, 8(2):e1002375.

[12] Krämer, A., Green, J., Pollard, J., and Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*, 30(4):523–530.

[13] Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., and Lee, D. (2008). Inferring Pathway Activity toward Precise Disease Classification. *PLOS Computational Biology*, 4(11):e1000217.

[14] Liu, Z., Yao, Z., Li, C., Lu, Y., and Gao, C. (2011). Gene Expression Profiling in Human High-Grade Astrocytomas. *Comparative and Functional Genomics*, 2011:1–10.

[15] Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10:161.

[16] Nam, D. and Kim, S.-Y. (2008). Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, 9(3):189–197.

[17] Pan, K.-H., Lih, C.-J., and Cohen, S. N. (2005). Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 102(25):8961–8965.

[18] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47.

[19] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140.

[20] Simillion, C., Liechti, R., Lischer, H. E., Ioannidis, V., and Bruggmann, R. (2017). Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinformatics*, 18:151.

[21] 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., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

[22] Tarca, A. L., Bhatti, G., and Romero, R. (2013). A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PLOS ONE*, 8(11):e79217.

[23] Tarca, A. L., Draghici, S., Bhatti, G., and Romero, R. (2012). Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13:136.

[24] Tomfohr, J., Lu, J., and Kepler, T. B. (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6:225.

[25] Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M.-L., Visvader, J. E., and Smyth, G. K. (2010). ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182.

[26] Wu, D. and Smyth, G. K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133–e133.

[27] Zyla, J., Marczyk, M., Weiner, J., and Polanska, J. (2017). Ranking metrics in gene set enrichment analysis: do they matter? *BMC Bioinformatics*, 18:256.

Supplementary material

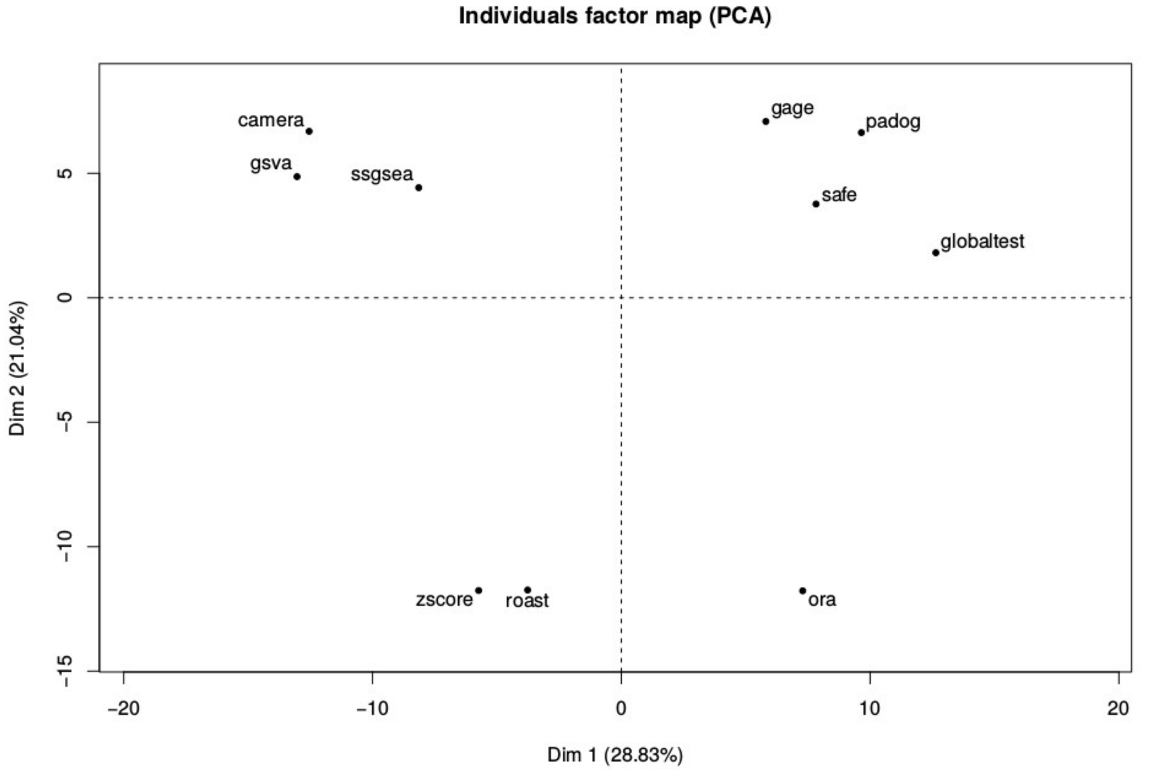

Figure S1 : PCA on the first 2 components of the methods on gene-sets ranks. The total variance explained by the two first component is 49.87%.

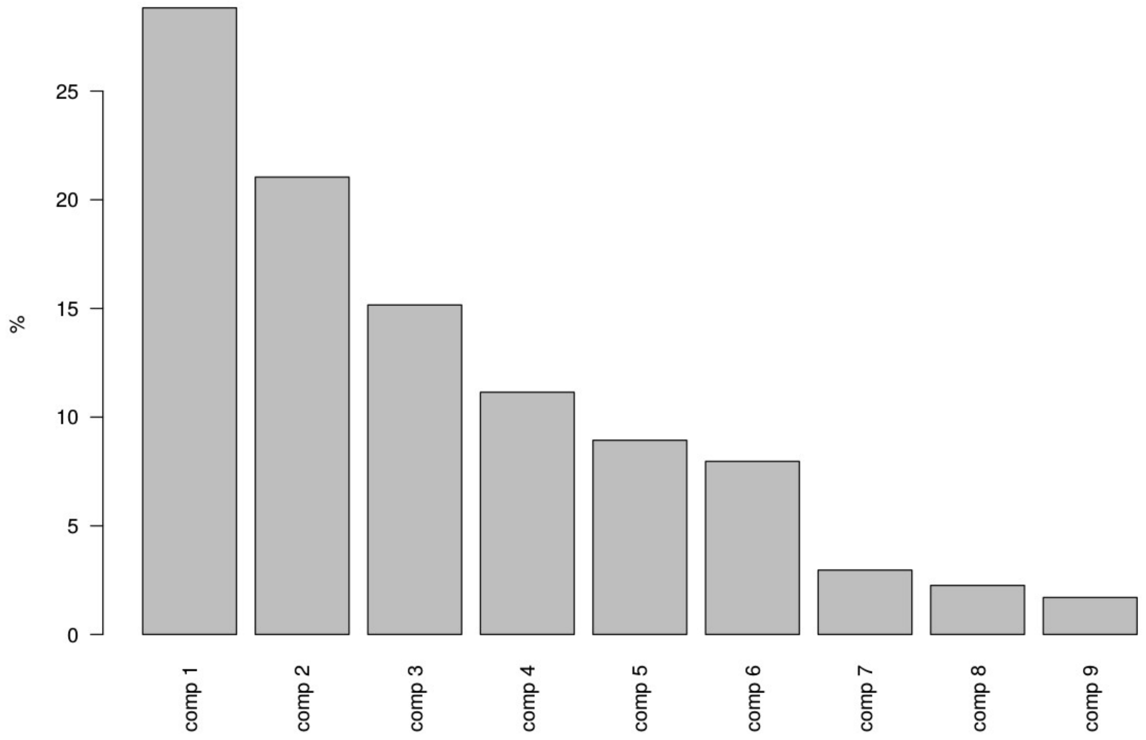

Figure S2 : Fall of the eigen values (in percentage of the explained variance) of the PCA on the first 9 components

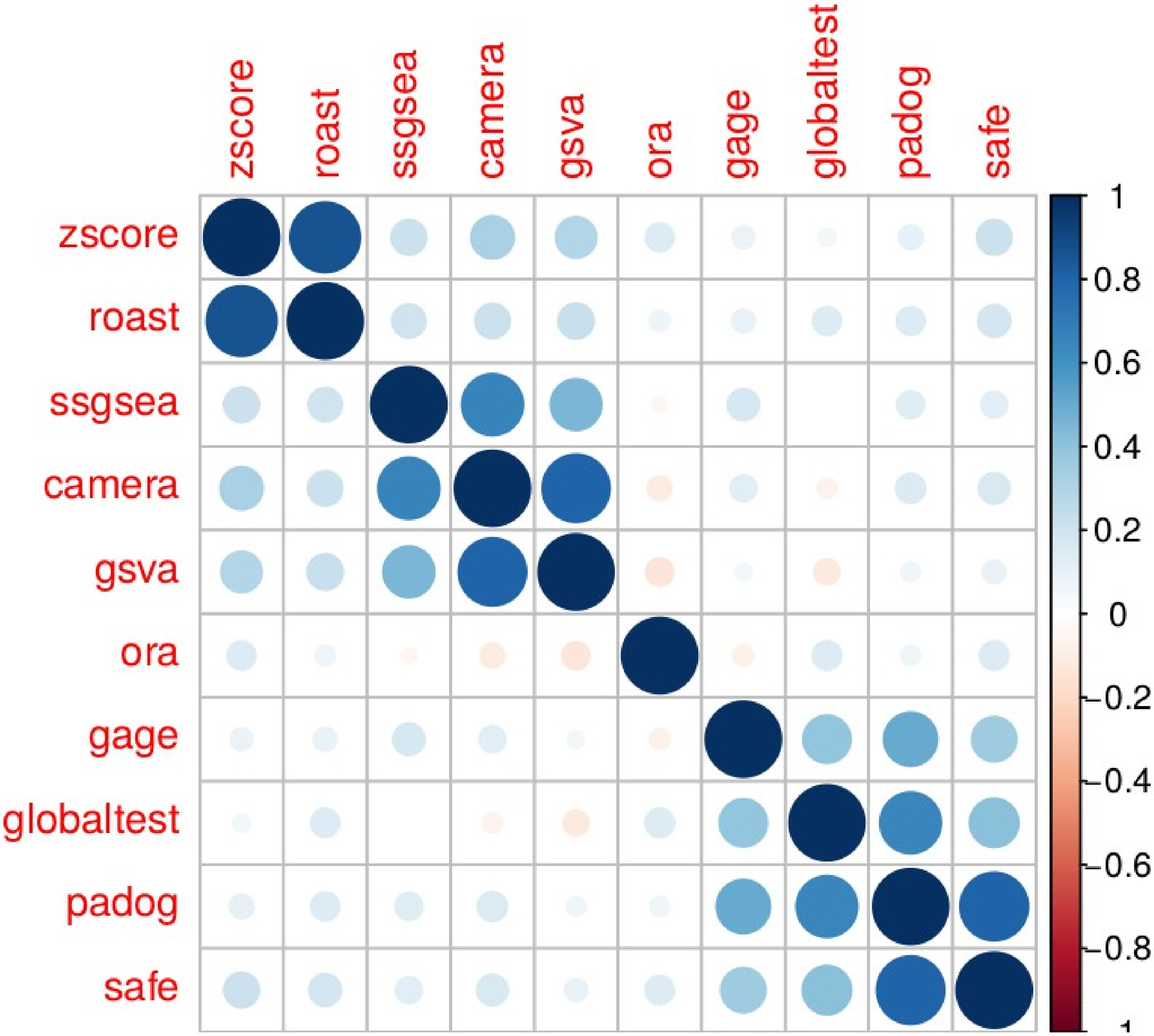

Figure S3 : Correlation plot of the methods on the gene-sets ranks. The color of the dots shows the sign of correlation coefficient between two methods, while the size of the dot shows the correlation coefficient itself. Methods correlating well together are clustered using hierarchical clustering (euclidean distance, ward.D2 method).

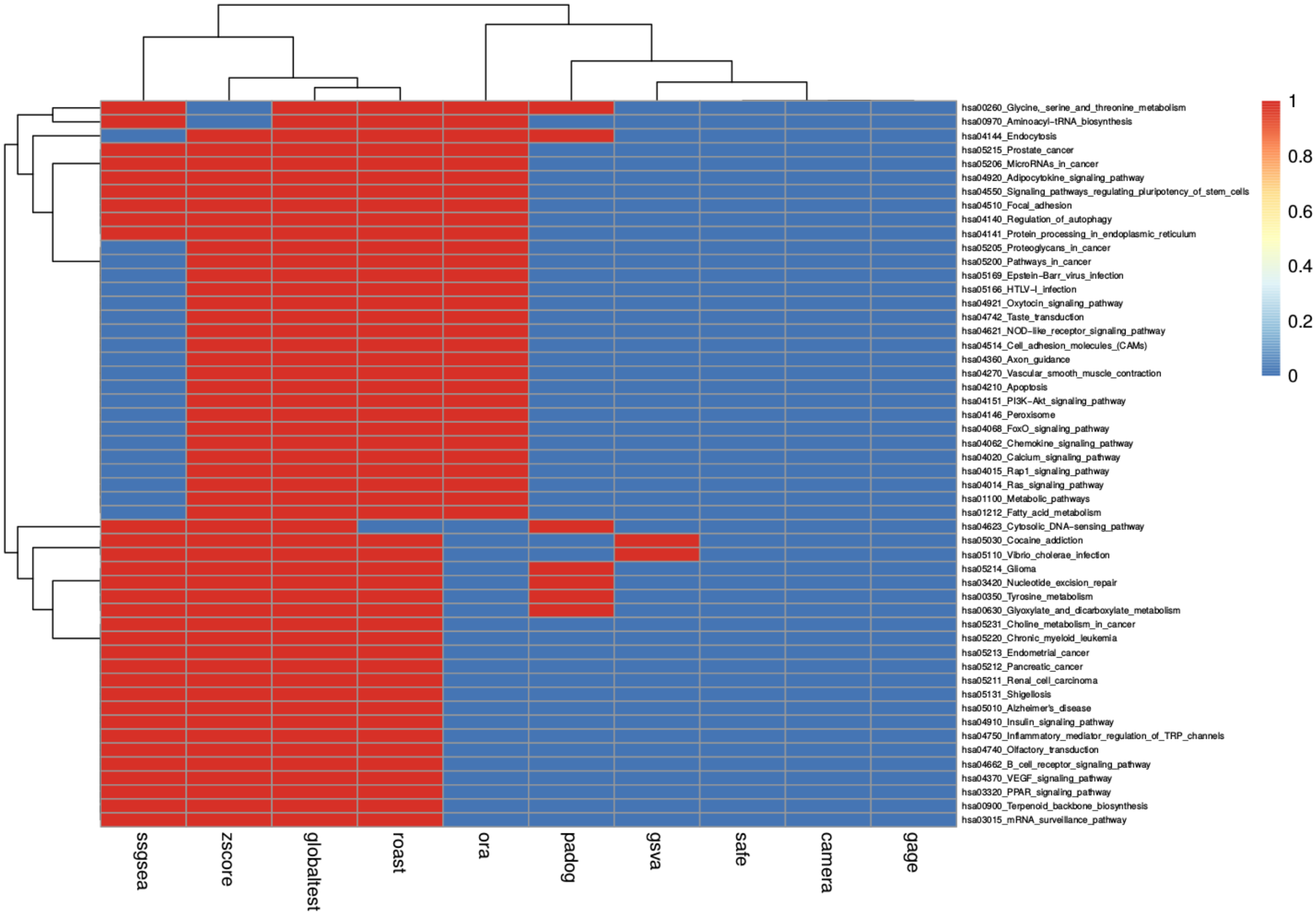

Figure S4 : Heatmap of the most commonly found gene-sets for each method at α = 0.05. SEGS are shown in red, in blue if not significant. Hierarchical clustering performed using a euclidean distance, and a ward.D2 method.

Method	camera	gage	globaltest	gsva	ssgsea	zscore	ora	padog	roast	safe	coGSEA
Rank of hsa05214 Glioma gene-set	7	5	28	16	6	9	97	10	4	44	1

Table S1 : Ranking of the target gene-set hsa05214 glioma by the different GSE methods.

The script for the analysis of the glioma dataset GSE19728 can be found at this address: https://github.com/maxibor/GSE19728_Analysis

Abstract

English

There exist many Gene Set Enrichment (GSE) methods for assessing the enrichment of gene-sets in differentially expressed genes, in the context of transcriptomics. However, applied on the same data-set and biological conditions, the results of those methods can be very different one from the other, and from the true biological state. Combining GSE methods allows for a result which is closer to the biological truth. Here, I present a tool I developped to combine GSE methods: coGSEA, now available as an R package and as an interactive Graphical User Interface.

Français

Il existe de nombreuses méthodes de vérification d'enrichissement de gene set (GSE) en gènes significativement différentiellement exprimés. Cependant, ces méthodes, utilisées sur un même jeu de données, et avec les mêmes conditions biologiques, peuvent produirent différents résultats l'une de l'autre, mais également différents de la réalité biologique expérimentalement validée. En combinant ces différents méthodes de GSE, le résultat obtenu permets de mieux retrouver la réalité biologique. Dans ce rapport, je présente une méthode qui combine différentes méthodes de GSE: coGSEA, maintenant disponible en tant que package R et en tant qu'interface graphique interactive.