Computers in Biology and Medicine 79 (2016) 10–20

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Improving information retrieval in functional analysis

Juan C. Rodriguez[a,b], Germán A. González[a,c], Cristóbal Fresno[a], Andrea S. Llera[d], Elmer A. Fernández[a,e,*]

[a] UA AREA CS. AGR. ING. BIO. Y S, Universidad Católica de Córdoba, CONICET, Córdoba, Argentina
[b] Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba, Córdoba, Argentina
[c] Instituto Nacional de Cáncer, MinSal, Córdoba, Agentina
[d] IIBBA, Fund. Instituto Leloir, CONICET, Buenos Aires, Argentina
[e] Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de Córdoba, Córdoba, Argentina

ARTICLE INFO

Keywords:
Big omics data
Gene set enrichment analysis
Functional class scoring
Over representation analysis
Singular enrichment analysis
Biological insight
Knowledge discovery
Breast cancer
R framework

ABSTRACT

Transcriptome analysis is essential to understand the mechanisms regulating key biological processes and functions. The first step usually consists of identifying candidate genes; to find out which pathways are affected by those genes, however, functional analysis (FA) is mandatory. The most frequently used strategies for this purpose are Gene Set and Singular Enrichment Analysis (GSEA and SEA) over Gene Ontology. Several statistical methods have been developed and compared in terms of computational efficiency and/or statistical appropriateness. However, whether their results are similar or complementary, the sensitivity to parameter settings, or possible bias in the analyzed terms has not been addressed so far. Here, two GSEA and four SEA methods and their parameter combinations were evaluated in six datasets by comparing two breast cancer subtypes with well-known differences in genetic background and patient outcomes. We show that GSEA and SEA lead to different results depending on the chosen statistic, model and/or parameters. Both approaches provide complementary results from a biological perspective. Hence, an Integrative Functional Analysis (IFA) tool is proposed to improve information retrieval in FA. It provides a common gene expression analytic framework that grants a comprehensive and coherent analysis. Only a minimal user parameter setting is required, since the best SEA/GSEA alternatives are integrated. IFA utility was demonstrated by evaluating four prostate cancer and the TCGA breast cancer microarray datasets, which showed its biological generalization capabilities.

1. Introduction

Cancer is so heterogeneous that single the analysis of differentially expressed (DE) genes is not enough to gain biological insight of this complex disease [1]. On the contrary, it is the starting point for an interpretation process in which biologists search for patterns using different information sources [2]. The process to uncover those functionalities is known as Functional Analysis (FA), which is based on the assessment not of individual genes but of genes grouped due to their association with a biological mechanism (gene sets), under the assumption that their coordinated action impacts the same biological process [2,3]. There are two main approaches to perform this task: Over Representation Analysis and Functional Class Scoring [4,5]. According to Huang et al., the most commonly used methods in those categories are Singular and Gene Set Enrichment Analysis (SEA and GSEA), respectively [6]. The former uses an interest gene list as input, which is usually the DE gene list. Then, given a statistical test based on a contingency table, each term is evaluated and considered enriched if the observed proportion of DE genes in the term differs from the expected distribution when compared against a background reference (BR). One of the main criticisms towards SEA is that it requires a user-defined DE gene list (usually by setting a threshold) [2,4,5,7–9]. GSEA methods have overcome this limitation by using all gene expression levels available in the experiment. These genes are sorted according to some metric related to the analyzed phenotype.

Several SEA and GSEA algorithms have been proposed [9] with their own assumptions and input parameters, which could potentially lead to different results. Indeed, some gene sets such as the ones provided by the Gene Ontology (GO) Consortium [10] are organized in some particular structure that yields additional penalization strategies to consider. Therefore, selecting the appropriate algorithm and its parameter settings is not trivial decision to make for researchers that face a biological problem and has not been comprehensively addressed. In addition, what each method returns from an information retrieval point of view is not clear; moreover, whether these results are independent of the method and parameters, complementary or are

* Corresponding author at: UA AREA CS. AGR. ING. BIO. Y S, Universidad Católica de Córdoba, CONICET, Córdoba, Argentina.
E-mail address: efernandez@bdmg.com.ar (E.A. Fernández).

http://dx.doi.org/10.1016/j.compbiomed.2016.09.017
Received 20 April 2016; Received in revised form 21 September 2016; Accepted 22 September 2016

equally useful is also unclear. Manoli et al. [4] compared both approaches taking into account only the top-20 highly raked pathways from 227 evaluated, using three datasets involving a total of 160 subjects. Pavlidis et al. [5] also compared SEA and GSEA approaches, but using 41 paired brain sample experiment and considering only the top-10 pathways from 965 evaluated. However, both Manoli and Pavlidis obtained unexpected GSEA results, since they tested only one parameterization that was not the recommended one. Manoli used an un-weighted Kolmogorov-Smirnov statistic, whereas Pavlidis used a pre-ranked alternative. Thus, a comprehensive analysis over a wider range of gene sets (GSs) as well as over larger cohorts is crucial in order to design a comprehensive and unified FA analysis approach.

One of the main drawbacks in comparing methods is the lack of "gold standards" or a benchmark dataset, as stated by Khatri et al.; in this case, the use of real biological datasets is preferable than simulated data since the latter lack biological factors [9]. To overcome this issue, here we propose the use of several experiments to evaluate the same (and very contrasting) cancer phenotypes, assuming that they should exhibit similar functional profiles across experiments. Our hypothesis is that in a horizontal cohort meta-analysis that contrasts two phenotypes with well-known differences, functional enrichment consensus patterns should be shared across all datasets, independently of the method used. Cohort differences could be regarded as biological particularities that can be further explored. For instance, although little or no overlap was found between several molecular signatures (outcome or phenotype-related gene sets) in different patient cohorts with the same phenotype [11], common functionalities in terms of biological functions have been reported [1]. Thus, FA results should be quite similar, showing high consensus between datasets despite their differentially expressed genes or their ranking over each cancer dataset. In addition, to determine if the enrichment results are truly related to the breast cancer concept, a literature validation will be required.

Here, we analyzed and comprehensively compared four and two methods for SEA and GSEA, respectively, as well as the combinations of their parameter effects, with the aim of finding the best strategy to improve biological information gain in FA. The methods were evaluated in six breast cancer datasets contrasting Basal-Like versus Luminal A subtypes [12] over GO. The relationship of the results with breast cancer was validated in the literature using the Comparative Toxicogenomics Database (CTD) [13]. Finally, using the best method/parameter combinations, we presented an Integrative Functional Analysis (IFA) framework. IFA simultaneously performs SEA and GSEA analyses, providing a unified and simplified FA approach and minimizing user-defined parameters. The proposed IFA was further evaluated using breast cancer samples from The Cancer Genome Atlas and in four prostate cancer datasets from Bioconductor.

2. Materials and methods

2.1. Input data

The analyzed datasets correspond to breast cancer and were obtained from the Bioconductor repository (Table 1). For each of the six datasets, subjects were classified into breast cancer intrinsic subtypes (Basal-Like, Her2-Enriched, Luminal B, Luminal A and Normal-Like) using the PAM50 algorithm [14] by means of the genefu R library [15] and processed as suggested by Sorlie et al. [16]. Only those subjects classified as Basal-Like or Luminal A were included (741 subjects in total). The comparison of survival outcomes of both subtypes showed that Luminal A has a better prognosis than Basal-Like [14,17,18]. Hence, we expect to identify many deregulated genes impacting on several (enriched) terms that should be shared across datasets. The expression matrices of these subjects were obtained for each dataset. Those genes with a valid Entrez Gene ID and expression values reliably detected for at least 50% of samples per condition were considered for further analyses.

2.2. Gene sets

Both SEA and GSEA methods require to be provided with the gene sets. Except for DAVID and dEnricher, in which the GSs are held in their knowledge bases, the *org.Hs.eg.db* (*v3.0.0*) R library [19] was used to retrieve GO terms.

2.3. Functional analysis algorithms

The tested algorithms described in the sections below were evaluated using their default cutoff options (see Supplementary Material).

2.3.1. Singular enrichment analysis

2.3.1.1. Methods and parameter combinations. In SEA, one of the most widely used tools is the Database for Annotation, Visualization and Integrated Discovery (DAVID) web platform [20]. The DE gene list of each dataset was submitted to DAVID through the RDAVIDWebService R package [21] (hereafter WD). One of the main drawbacks of the DAVID platform is that, in the 6.7 version, its knowledge base has not been updated since 2010 [6,20], and the 6.8 version is still in beta and cannot be accessed programmatically through the R environment. In order to overcome this issue, RD, an R version of DAVID's EASE score (see Supplementary Material) was developed, which allows us to analyze any knowledge base. Moreover, unlike DAVID and RDAVIDWebService, RD is not web based and therefore does not require an internet connection. The third SEA method evaluated was GOstats [22], which was designed specifically to perform GO enrichment analysis through a Hypergeometric test and takes into account the GO structure to penalize term enrichment using the elim algorithm. A competing method is dEnricher [23], which accounts for the GO structure hierarchy using its own definition of GO gene sets. It provides three different statistical tests: Hypergeometric, Binomial and Fisher, as well as four different enrichment penalizing algorithms based on the GO structure: none, lea, elim and pc. Except for pc algorithm, which is still under development, every statistic/algorithm was tested, which resulted in nine dEnricher combinations.

2.3.1.2. SEA inputs. SEA methods require as input the GSs, the gene background reference list and the DE gene list. Since different lengths of the background reference list could lead to different results [24,25], the BR strategies proposed by Fresno et al. [25] were evaluated and compared. These BRs are the genome (BRI) and those genes reliably detected in the experiment (BRIII, Table 1). The dEnricher method only allows the use of the genes in its knowledge base, and not any other BR.

The second input is the list of DE genes, obtained from genes with an absolute fold change greater than a given threshold (treatLfc) using *treat* [26] function from limma R library [27]. A false discovery rate, FDR, adjusted p-value ≤ 0.01 was used to define DE genes. To provide a comparable gene list length between datasets, the treatLfc was chosen to yield a gene list of about 5% of the BRIII length.

2.3.2. Gene set enrichment analysis

2.3.2.1. Methods and parameter combinations. The Subramanian Method (SM) [7] Java implementation *(v2-2.2.1)*, available at the BROAD Institute website, was used because the R version is deprecated. The SM can be fed with both a pre-ranked gene list (SMpr), obtained through a suitable metric or with the expression matrix. For the latter, the significance of the enrichment score statistic is estimated through a permutation strategy over the gene labels (SMgp) or the sample phenotype (SMpp). To determine gene set enrichment, SM calculates an enrichment score to which it applies a

weighted Kolmogorov-Smirnov like statistic [7], where this weight "w" is set to 1 by default but may be also set to 0 or 2. However, "w" effect is not clear, since different parameterizations are suggested according to the input data (expression matrix or pre-ranked gene list). Here, pre-ranked as well as both permutation strategies and the set of "w" values were evaluated.

The other GSEA method evaluated was the mGSZ R library, which is based on gene set Z-scoring function and asymptotic p-value estimation using sample and (implicit) gene permutation [36]. The mGSZ ranks the genes using limma's eBayes function, and orders them according to the moderated t-statistic [27].

2.3.2.2. GSEA inputs. For SM the default GSs size limits filtering function was used, which analyzes GSs including between 15 and 500 genes. Since a one column pre-ranked gene list can also be used as input, in order to pre-rank these genes, statistical p-values were obtained for each gene applying the limma's *eBayes* function to each dataset to make it comparable to mGSZ. Then, the order was assigned according to the t-statistic, 1-p-value and -log(*p*-value) metrics. In mGSZ, the gene expression matrix and the sample phenotype labels were used because it does not accept a pre-ranked list alternative. Since mGSZ limits the GSs sizes by default, using a minimum of 5 genes, it was also set between 15 and 500 as in the SM method in order to make a robust comparison.

2.4. Tested methods

$$
\begin{aligned}
&Subra-manian \begin{cases}
Genepermutation \begin{cases} w=0 & \textbf{SMgp0} \\ w=1 & \textbf{SMgp1} \\ w=2 & \textbf{SMgp2} \end{cases} \\
Phenotypepermutation \begin{cases} w=0 & \textbf{SMpp0} \\ w=1 & \textbf{SMpp1} \\ w=2 & \textbf{SMpp2} \end{cases} \\
Pre-rank \begin{cases}
t-score \begin{cases} w=0 & \textbf{SMprtScore0} \\ w=1 & \textbf{SMprtScore1} \\ w=2 & \textbf{SMprtScore2} \end{cases} \\
-logp-value \begin{cases} w=0 & \textbf{SMpr}-\textbf{log(p)0} \\ w=1 & \textbf{SMpr}-\textbf{log(p)1} \\ w=2 & \textbf{SMpr}-\textbf{log(p)2} \end{cases} \\
1-p-value \begin{cases} w=0 & \textbf{SMpr1}-\textbf{p0} \\ w=1 & \textbf{SMpr1}-\textbf{p1} \\ w=2 & \textbf{SMpr1}-\textbf{p2} \end{cases}
\end{cases}
\end{cases} \\
&mGSZ \begin{cases} GSfilter\,[15,\,500) & \textbf{mGSZ[15, 500)} \\ GSfilter\,[5,\,\infty) & \textbf{mGSZ[5, }\infty\textbf{)} \end{cases} \\
&RD \begin{cases} BRI & \textbf{RDBRI} \\ BRIII & \textbf{RDBRIII} \end{cases} \\
&DAVID \begin{cases} BRI & \textbf{WDBRI} \\ BRIII & \textbf{WDBRIII} \end{cases} \\
&GOstats \begin{cases} BRI & \textbf{GOstatsBRI} \\ BRIII & \textbf{GOstatsBRIII} \end{cases} \\
&dEnricher \begin{cases}
Hypergeo-metric \begin{cases} none & \textbf{dEHnone} \\ lea & \textbf{dEHlea} \\ elim & \textbf{dEHelim} \end{cases} \\
Binomial \begin{cases} none & \textbf{dEbnone} \\ lea & \textbf{dEblea} \\ elim & \textbf{dEbelim} \end{cases} \\
Fisher \begin{cases} none & \textbf{dEFnone} \\ lea & \textbf{dEFlea} \\ elim & \textbf{dEFelim} \end{cases}
\end{cases}
\end{aligned}
$$

Table 1
PAM50 algorithm assignment.

Breast cancer datasets	PAM50 subjects		Number of BRIII genes
	Basal	Luminal A	
Vdx [28,29]	91	108	13091
Nki [30,31]	70	105	13108
Transbig [32]	46	66	13091
Upp [33]	34	69	18528
Unt [34]	22	40	18528
Mainz [35]	33	57	13091
Total	296	445	

BRIII: Background reference, i.e. genes reliably detected in the experiment. Dataset references are included between brackets. Every analyzed dataset was manufactured by Affymetrix®, except for Nki which is Agilent®.

2.5. Information retrieval analysis

The results of the different methods were evaluated in terms of their stability, GO depth, and consensus analyses, as described below.

2.5.1. Stability analysis

Enrichment stability was evaluated using boxplots to assess the distribution of the number of enrichment yielded for each method/parameter combination.

2.5.2. Gene Ontology depth analysis

The GO is structured as three hierarchical directed acyclic graphs (trees; "molecular functions", "biological processes" and "cellular components"), where a child node represents a more biologically specific term than its parent. To explore the biological specificity over the GO structure, the minimum number of arcs between the node and the root (depth) of each enriched term was calculated. Then, a frequency table of the number of enriched terms by depth was calculated (grouping results for every dataset).

2.5.3. Consensus analysis

To evaluate the hypothesis that the compared phenotypes should present similar enrichment profiles across datasets, an enrichment matrix $E = e_{mdt}$ was built. In this matrix, each row holds a GO term and each column holds a method/parameter per dataset combination. Where each e_{mdt} cell of the matrix was defined as follows:

$$
e_{mdt} = \begin{cases} 1 & if\ m\ \textbf{enriched}\ t\ on\ dataset\ d \\ 0 & if\ m\ \textbf{did not enrich}\ t\ on\ dataset\ d \\ NA & if\ m\ \textbf{did not analyze}\ t\ on\ dataset\ d \end{cases}
$$

where $m = 1 \ldots M$ method/parameter combination; $t = 1 \ldots T$ a gene set; $d = 1 \ldots D$ a dataset.

For consensus analysis, those GO terms that were not enriched in any dataset were filtered. Using the vegan R library [37] a hierarchical clustering was applied to E via Jaccard distance and average linkage to automatically group similar enrichment profiles; both the matrix and the clustering were displayed using a heatmap. Since all experiments compare the same breast cancer phenotypes, we expected to find concordant results across datasets for every method/parameter combination. Thus, the number of enriched terms across almost all datasets (enrichment consensus; EC), as well as those terms that were not enriched across almost all datasets. (non-enrichment consensus; NEC), were used as an indicator of the stability of the method. Based on this assumption, we define the comparison metrics listed in Table 2.

2.5.4. Complementary enrichment and relevance

In order to determine whether the methods could be considered complementary or not from an information retrieval perspective, the

Table 2
Comparison metrics.

Name	Definition
Inter-method term enrich frequency	$F_{mt} = \frac{1}{D}\sum_{d=1}^{D} e_{mdt}$
Enrichment consensus	$EC_m = \sum_{t=1}^{T} I(F_{mt} > t_e)$
Non-enrichment consensus	$NEC_m = \sum_{t=1}^{T} I(F_{mt} < t_n)$
Non-consensus	$NC_m = \sum_{t=1}^{T} I(t_n \leq F_{mt} \leq t_e)$
EC frequency	$FEC_m = \frac{EC_m}{EC_m + NC_m}$
NEC frequency	$FNEC_m = \frac{NEC_m}{NEC_m + NC_m}$

$d = 1 \ldots D$: selected dataset; $t = 1 \ldots T$: selected gene set; m: any method/parameter combination used; t_e=0.8 and t_n=0.2: thresholds for enrichment and non-enrichment, respectively; I: indicator function.

exclusively enriched terms were analyzed for each method, i.e., terms present, for m, in 80% of the datasets but less than 20% in all the others $m' \neq m$, see Eq. (1).

$$EET_m = \{t | F_{mt} > t_e \wedge \forall\, m' \neq m\colon F_{m't} < t_n\} \quad (1)$$

where t is any gene set; m is any method/parameter combination; t_e=0.8 and t_n=0.2 are the thresholds for enrichment and non-enrichment, respectively.

To assess the annotated phenotype association of exclusively enriched terms, each EET_m was queried through the CTD [13] in order to ascertain its pathological condition in relation to the "breast cancer" concept.

3. Results

3.1. Differentially expressed genes per dataset

The number of DE genes for each dataset is presented in Table 3. Although we compare the same phenotypes of breast cancer, very little overlap between pairs of datasets is observed. For instance, Unt and Nki datasets only share 383 DE genes, with Unt having 1059 DE genes (only 36% overlap). Overall, only 12% (195 genes) of the DE genes in at least one dataset (their intersection) were found to overlap between the union (1678 genes) of all the evaluated datasets.

3.2. Stability analysis

Integration of results from the different datasets allows us to provide inter-study validation as stated by Edelman et al. [38]. The boxplot of the number of enriched terms for each SEA/GSEA method and parameterization is shown in Fig. 1 and as Supplementary Table S3.

Table 3
Differentially expressed genes between datasets.

Datasets	Vdx	Nki	Transbig	Upp	Unt	Mainz
Vdx (0.75)	611 (4.7)	292	465	430	425	412
Nki (0.2)		568 (4.3)	310	374	383	286
Transbig (0.6)			628 (4.8)	448	461	433
Upp (0.3)				932 (5)	632	428
Unt (0.25)					1059 (5.7)	437
Mainz (0.45)						605 (4.6)
	Intersection=195			Union=1678		

First column: Dataset names with treat log fold change threshold shown between parenthesis. Principal diagonal: Number of differentially expressed (DE) genes for each dataset (FDR ≤ 0.01) and percentage of the total genes in the experiment between parenthesis. Superior triangular: Number of intersected DE genes for every pair of datasets. Notice that the global intercept of DE genes is only 195 of a total union of 1678 genes.

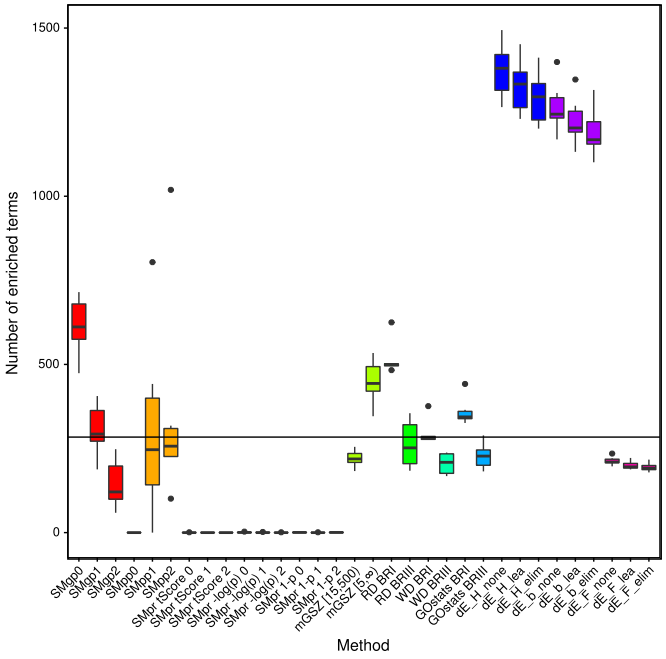

Fig. 1. Boxplot of the number of enriched terms for each method over the different datasets. Method acronyms are described in Section 2.4. Notice that SM pre-ranked methods enrich almost no gene sets, whereas the median number of enriched terms is 284 for the remaining methods (horizontal black line). The SMpp1 method obtained the most variable number of enriched terms. Except for dEnricher with Fisher test, all the other dEnricher method/parameter combinations returned extreme values.

A high variability between methods as well as within their parameterizations is observed. A median of 284 enriched terms was found including only datasets with at least one enriched term. The SM method seems to be very sensitive to different parameterizations as well as to the way in which the gene information is fed into the algorithm, i.e., through the gene expression matrix or by a pre-ranked gene list. Interestingly, for SMpr every value of the weighting factor "w" returned almost zero enriched terms (see Supplementary Material). In addition, for SMgp and SMpp, the selection of "w" could yield very different enrichments, ranging from zero terms in SMpp0 for Nki or 59 terms in SMgp2 for Vdx to extreme values such as 1019 in SMpp2 for Nki or 474 terms in SMgp0 for Vdx. In particular, the SMpp method presents very different behaviors, depending on the "w" value. For instance, no enrichment was found for w=0, great variability resulted with w=1 with an interquartile range (IQR) of 257.57, and concordant results were achieved with w=2, i.e., small dispersion over datasets with an IQR of 83.25. However, SMpp2 showed an extreme number of enrichments for one dataset (1019 for Nki) and a very low number for another (101 for Mainz), resulting in two outliers. This could pose a problem when analyzing only one dataset. For SMgp, the enrichment is quite stable across datasets, with IQRs of 105, 91.25 and 98.5 for SMgp0, SMgp1 and SMgp2, respectively, but a decreasing number of enrichments was obtained as "w" increased from 0 to 2, yielding a median of 611.5, 293 and 121, respectively. However, for w=1 the number of enriched terms is similar to overall median (284 enriched terms) of the different methods, suggesting SMgp1 appropriateness.

In the case of mGSZ, the results showed an adequate overall stability across datasets (IQRs of 26.5 for mGSZ[15,500) and 73 for mGSZ[5, ∞)), yielding a very similar number of enriched terms between datasets. This method is sensitive to the size of the gene sets presented in the analysis. When was set between [15,500), a quite conservative number of enriched terms was achieved, i.e., a lower number of enrichment values. These results were more stable when

compared with the value obtained with gene set limits between [5,∞). In addition, the enriched terms obtained with the [15,500) gene set size limit were mostly contained (93% on average) by the one accomplished with the default limits. These additional enriched terms usually contained a lower number of genes, i.e. more specific terms that can be much more useful to elicit the phenomenon under study. Based on this concept hereafter we followed mGSZ author's recommendation for gene sets filtering.

Except for dEnricher methods, the results of SEA alternatives were quite similar between them, yielding fairly stable enrichment across datasets: IQRs of 6.25 for RD BRI; 115.75 RD BRIII; 8.75 WD BRI; 58 WD BRIII; 21.75 GOstats BRI; and 45.5 GOstats BRIII. The RD with the BRIII showed greater variability than the WD counterpart, probably because a higher number of GSs were analyzed (76% more GSs on average). For every SEA alternative, enrichments obtained with the BRIII were in general contained in those obtained using BRI (99% of terms on average for RD and WD; and 86% for GOstats) in concordance with the observation of Fresno et al. [25].

In the case of dEnricher, Hypergeometric and Binomial tests returned an extreme number of enriched terms compared to overall median of the methods, returning above 1168, regardless of the penalizing algorithm applied. On the other hand, Fisher test returned quite stable results, as the other SEA methods: median values of 210, 195.5 and 191.5 for dE_F_none, dE_F_lea and dE_F_elim, respectively, with IQR values below 13.75. Enriched terms obtained with the dE_F_lea and dE_F_elim algorithms were 100% contained in those obtained with dE_F_none; moreover, 86% of these additional enriched terms found by dE_F_none were related to breast cancer when queried at the CTD. For the following analyses dEnricher combinations, except dE_F_none, were discarded.

Only those methods and configurations which returned a concordant number of enriched terms between datasets and enriched around the overall median of the methods (284 enriched terms) were considered for the following analyses, i.e., SMgp1, SMpp2, mGSZ [5, ∞), RD BRI, RD BRIII, WD BRI, WD BRIII, GOstats BRI, GOstats BRIII and dE_F_none.

3.3. Gene Ontology depth analysis of enriched terms

The percentage of enriched terms grouped by depth for each method is shown in Fig. 2 (see Supplementary Table S4). All the methods tend to explore depths mostly between three and six. The mGSZ, dE_F_none, SMpp2 and GOstats BRIII enriched the highest number of specific terms, i.e., lower nodes or leaves in the GO structure with depth >6: 13%, 12.8%, 10.7% and 9.8% of the total enrichment, respectively. Furthermore, WD and RD provided enrichment of general terms, i.e., depth <3, nodes closer to the GO tree root node: 13.7% for RD BRI, 11.9% for WD BRI, 9.7% for WD BRIII and 8.8% for RD BRIII. Within SEA methods, except dEnricher, proportionally more terms are enriched near the root when using BRI.

For every SEA method, except dEnricher, BRI enriched a higher number of terms than BRIII. However, as stated above, in both WD, RD - as in GOstats - most of the terms enriched by BRIII were also enriched by BRI. As discussed by Fresno et al., using BRIII statistically makes more sense than using BRI; thus, its use is suggested and used hereafter.

3.4. Consensus analysis

A biclustering analysis was conducted on the consensus E matrix and is displayed as a heatmap in Fig. 3. The top dendrogram shows that the analyzed datasets tend to cluster together according to the applied method, except for SMpp2, GOstats and RD, which present one spread dataset. A noticeable difference can also be observed between the results achieved using SEA vs. GSEA, i.e., GSEA methods form one separate cluster, whereas SEA methods are divided into two major

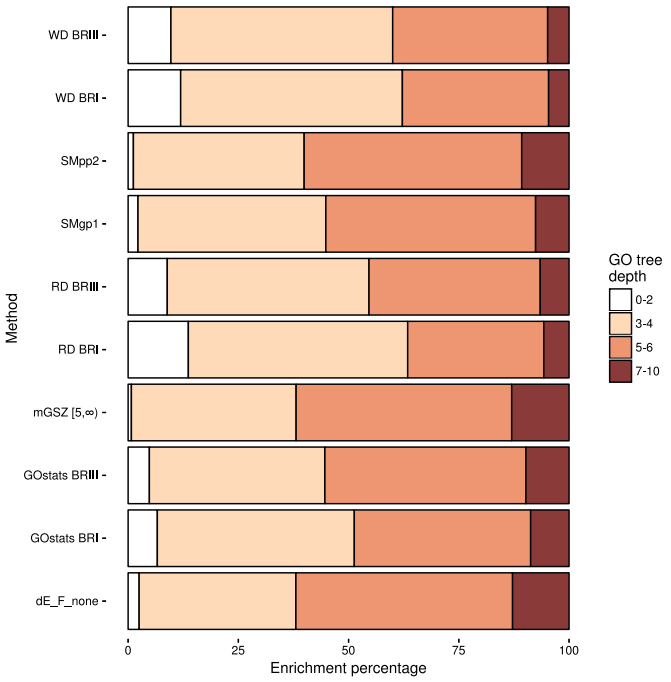

Fig. 2. Gene Ontology (GO) enrichment depth for each method. Darker colors represent deeper GO tree hierarchy terms. Notice that all methods tend to enrich depths mostly between three and six. The WD and RD methods enriched shallower terms of the GO tree structure. On the other hand, mGSZ, dE_F_none, SMpp2 and GOstats BRIII enriched deeper terms.

clusters, the RD/WD and the GOstats/dE_F_none. A subset of terms enriched by almost all methods can be seen across datasets, as expected (rows tagged as **A**). For instance, an average of 64% of those terms enriched in at least 80% of the datasets for each method were also enriched by mGSZ in the same proportion. This suggests that, to some extent, all the methods tend to provide the same information. However, each method also provides exclusively enriched terms (EET_m; rows tagged as **E**). The comparison of GSEA and SEA approaches shows that RD and WD tend to enrich some terms that are not enriched by any other method; the same happens with dE_F_none and with mGSZ, suggesting that GSEA and SEA complement each other. The dE_F_none, GOstats and RD, as well as mGSZ, analyze more terms than any other method (a lower number of $e_{mdt} = NA$). Moreover, 63% of the terms enriched in at least 80% of the datasets by SM methods were also enriched by mGSZ in the same proportion, whereas 47% of the terms enriched by GOstats, RD and WD were enriched by dE_F_none, suggesting that mGSZ and dE_F_none can be used as reference methods for GSEA and SEA, respectively.

The concordance across datasets for each method showed that mGSZ outperforms with 45% of concordant enriched terms (*FEC*), followed by dE_F_none, RD and WD with 39%, SMgp1 with 36%, SMpp2 with 30%, and GOstats with 29%. The concordance across terms that were not enriched (*FNEC*) yielded 91% for dE_F_none, RD and GOstats, 89% for WD, 85% for mGSZ, 83% for SMgp1, and 82% for SMpp2. Therefore, all methods seem to have high consensus for non-enriched terms and a low consensus for enriched ones. Both concepts are important when facing FA, since no biologically significant terms should be lost, nor should there be incorrectly enriched terms. Accordingly, GSEA analysis using mGSZ showed to be the most consensual method, whereas dE_F_none and RD for the SEA counterpart.

Since RD was able to analyze more GSs than WD, it can be used with an up-to-date GO knowledge base and does not depend on an internet connection, it was preferred over WD. Thus, the latter was left out from the analysis hereafter.

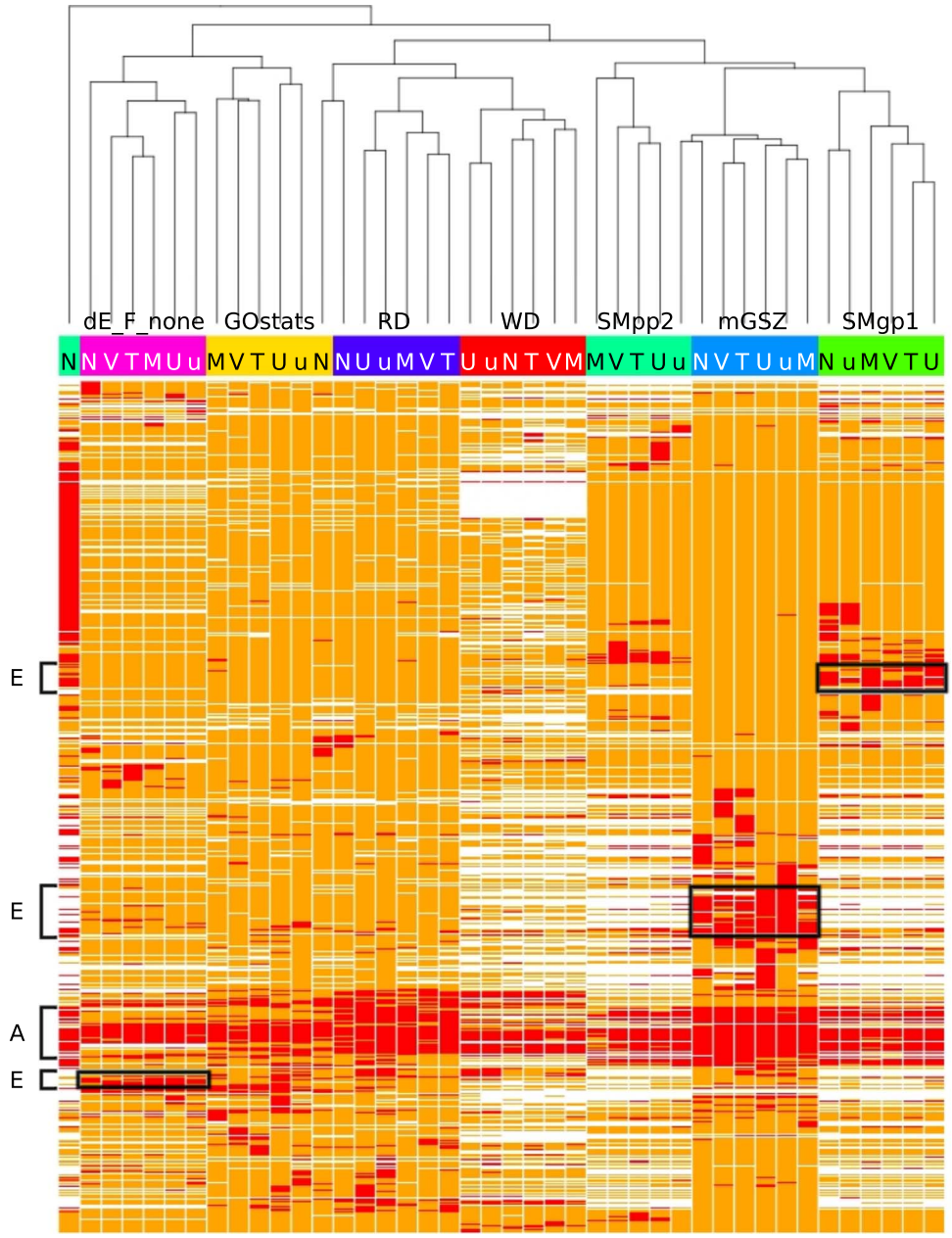

Fig. 3. Enrichment heatmap. In columns, each method/parameter combination per dataset; in rows, those gene sets (terms) enriched in at least one dataset. Notice that GSEA and SEA methods are separately clustered in the dendrogram. Red cells indicate enrichment, orange cells indicate no enrichment, and white cells show terms that were not analyzed. There are subsets of terms that resulted enriched across almost every analyzed method (**A**) and subsets of terms enriched (in every dataset) exclusively by only one method/parameter combination (**E**). The label color of each column represents the used algorithm, and the letter stands for the initial letter of the dataset. V: Vdx. N: Nki. T: Transbig. U: Upp. u: Unt. M: Mainz.

Table 4
Number of exclusive enriched terms.

	GO tree depths				
Method	0-2	3-4	5-6	7-10	Total
SMpp2			2(0)	1(1)	3(1)
SMgp1	2(1)	15(9)	7(4)		24(14)
mGSZ		26(13)	27(12)	8(3)	61(28)
dE_F_none	1(1)	4(3)	5(3)	3(1)	13(8)
RD	4(0)	2(1)			6(1)
GOstats		1(0)			1(0)

The number of enriched terms related to breast cancer according to the Comparative Toxigenomics Database [13] is presented in parenthesis. Notice that mGSZ and dE_F_none enriched the highest number of terms for GSEA and SEA, respectively.

3.5. Exclusive enrichment and term relevance

When considering those terms exclusively enriched by each method (EET_m), we found that mGSZ and dE_F_none yielded more EET_m for GSEA and SEA, respectively (See Table 4). In addition, mGSZ also provides more terms related to breast cancer according to the CTD as well as much more specific terms (depth >6). In the case of SEA, dE_F_none provided more EET_m, also related to breast cancer. For a detailed description about which are the EET_m for each method, and if they are related to the disease, see Table S5.

3.6. Integrative functional analysis

As a consequence of this comprehensive study, the Integrative

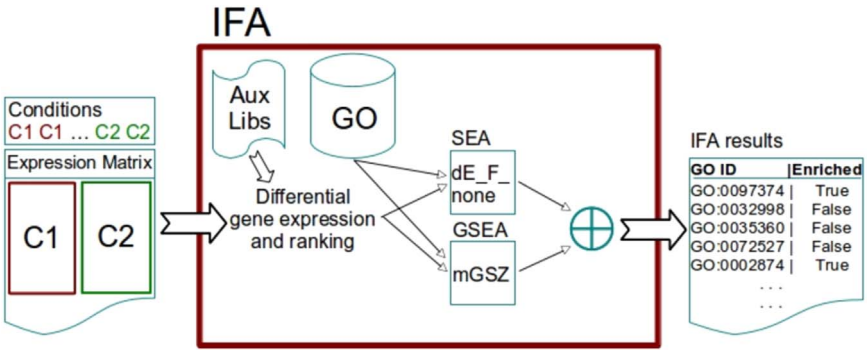

Fig. 4. Integrative Functional Analysis (IFA) workflow. The user provides the expression matrix and corresponding subject phenotype labels as input. IFA uses auxiliary R libraries to obtain differentially expressed genes, ranks them and performs SEA and GSEA analyses. Finally, enrichment results obtained by IFA integrate both SEA and GSEA results.

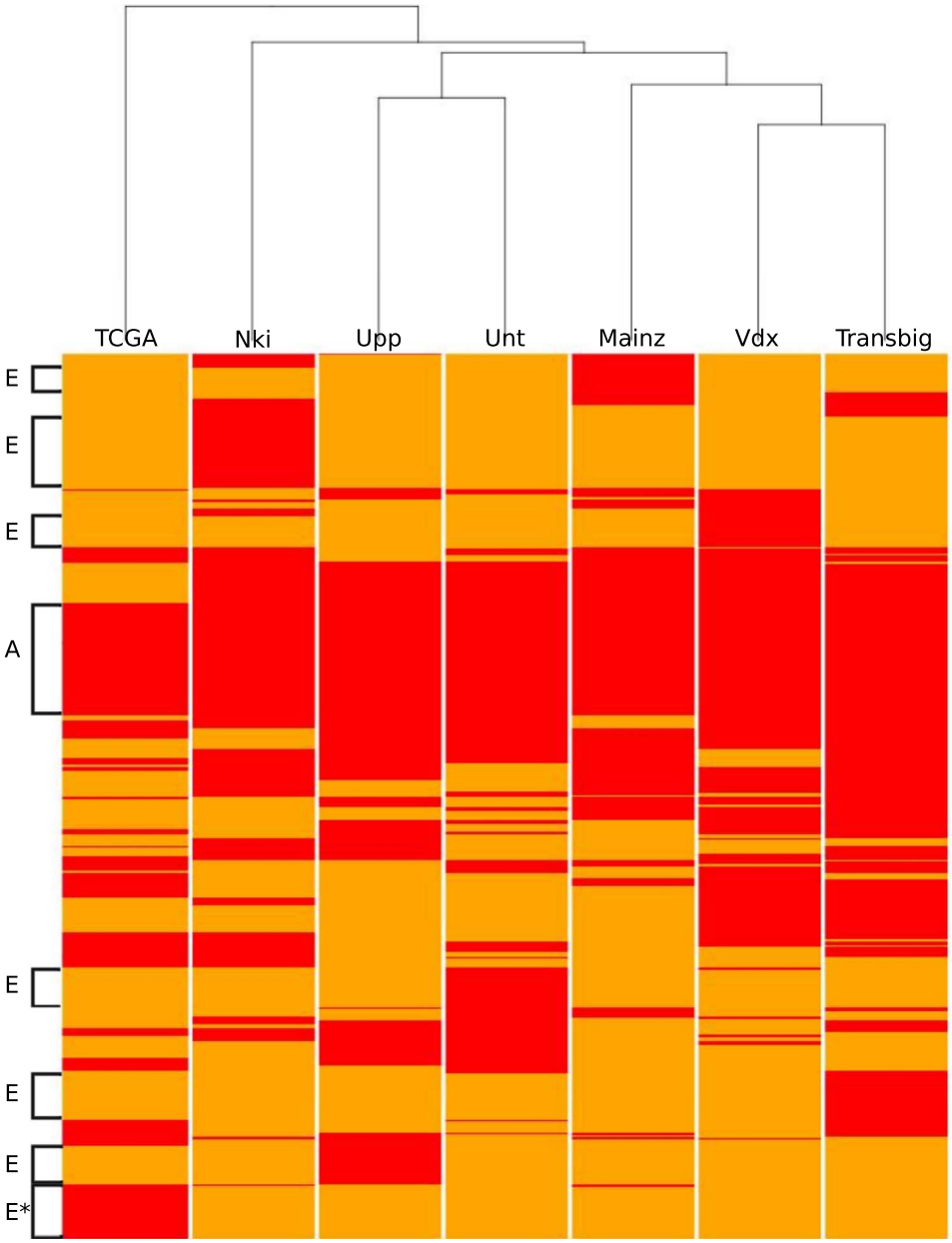

Fig. 5. Integrative Functional Analysis enrichment heatmap. Breast cancer datasets are presented in columns and those gene sets (terms) enriched in at least one dataset are presented in rows. Red cells indicate enrichment and orange cells indicate no enrichment. Notice concordant subsets of terms enriched between every dataset (**A**) and subsets of terms enriched exclusively only in one dataset (**E**).

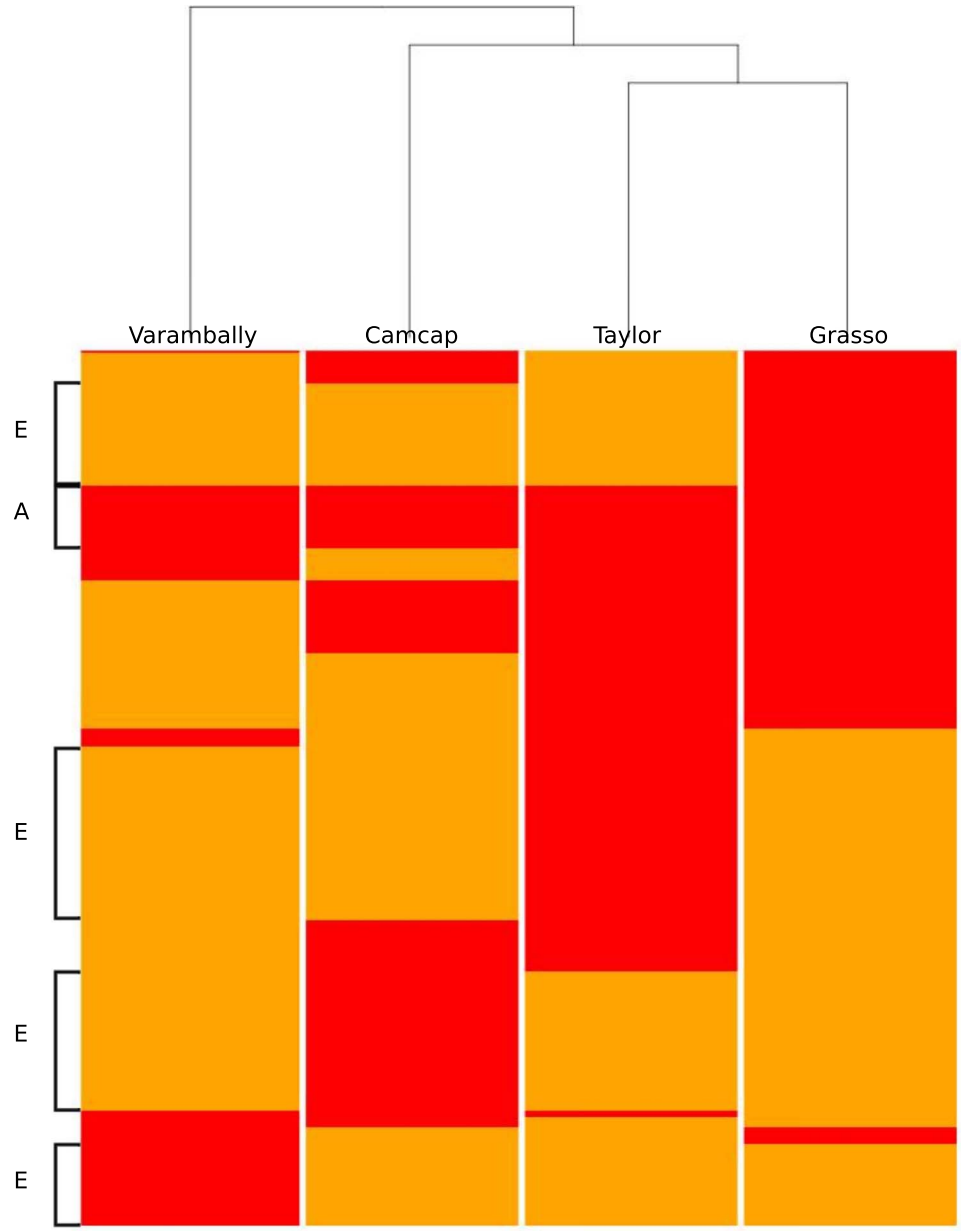

Fig. 6. Integrative Functional Analysis enrichment heatmap. Prostate cancer datasets are presented in columns and those gene sets (terms) enriched in at least one dataset are presented in rows. Red cells indicate enrichment and orange cells indicate no enrichment. Notice concordant subsets of terms enriched between every dataset (**A**) and subsets of terms enriched exclusively only in one dataset (**E**).

Functional Analysis framework is presented in Fig. 4 and provided as an R source code (at Github: http://github.com/jcrodriguez1989/IFA). The user should provide the expression matrix and the phenotypes specified as condition labels. Then, if no GSs are provided, IFA will use the up-to-date GO gene sets from the *org.Hs.eg.db* R library, and obtain the DE gene list and gene ranking by a linear model through limma R library in order to perform both mGSZ and dE_F_none analyses. Thus, it provides a simple, unified and comprehensive FA approach.

3.6.1. IFA over TCGA

To test the IFA framework, the mRNA breast invasive carcinoma dataset from TCGA was obtained, which consists of 86 Basal-Like and 198 Luminal A subjects. A treatLfc=1 cutoff was used to obtain about 5% of DE genes. The IFA results for this dataset were used as test case and, therefore, were evaluated and compared with the previously analyzed datasets (Table 1). As in Section 2.5.3, the IFA consensus matrix results were presented as a heatmap in Fig. 5, where 812 terms resulted enriched in TCGA dataset. Thirty-three percent of the terms (270) were also enriched in all the other datasets (tag **A** in Fig. 5), 43% (352 terms) were also enriched in more than 80% of the other datasets (CONCORDANT terms in Table S6), and 15% (123 terms) were exclusively enriched by TCGA (tag **E*** in Fig. 5, TCGA-EXCLUSIVE in Table S6). In this consensus matrix, 445 terms resulted enriched in at least 80% of all the datasets, of which 232 (52%) were related to breast cancer according to the CTD. Particularly, the mean proportion of exclusively enriched terms present in CTD in each dataset was 42%, with Nki leading with 170 exclusive terms (84 in CTD), and with Mainz dropping to only 58 terms (30 in CTD, see Table S7)..

Terms related to hormone and estrogen receptor, G1/S transition of mitotic cell cycle, DNA replication, mitotic spindle organization, DNA duplex unwinding, histone kinase activity, annealing helicase activity, among others, were commonly found across all datasets (tag **A** in Fig. 5, CONCORDANT terms in Table S6), supporting the proliferation differences between Basal-Like and Luminal A breast cancer subtypes. Furthermore, terms such as receptor signaling protein tyrosine kinase

activity, stem cell differentiation and others related to cell differentiation, were found only in the TCGA datasets (tag **E*** in Fig. 5, TCGA-EXCLUSIVE in Table S6).

3.6.2. IFA over prostate cancer datasets

With the purpose of checking that IFA results were not dependent on the analyzed pathology but may be extended to other scenarios, IFA was tested over four prostate cancer datasets from Bioconductor. In total, 519 subjects were retrieved: Camcap [39] 74 subjects with benign prostate cancer versus 125 with tumor; Taylor [40] with 29 benign versus 150 tumor; Varambally [41] 6 versus 13; and Grasso [42] 28 versus 94 (see Supplementary Material). In order to obtain about 5% DE genes over the total, treatLfc values of 0.2, 0.15, 0.2, and 0.45 were used for Camcap, Taylor, Varambally and Grasso, respectively. For Varambally analysis, p-values of the genes were not adjusted, since zero DE genes were reached at any treatLfc value with a fixed FDR p-adjusted cutoff of 0.01.

The resulting consensus matrix is shown in Fig. 6, in which 163 terms were enriched in at least 80% of the datasets (see Table S8), with 99 of them (61%) being related to prostate cancer according to the CTD. Particularly, the mean proportion of exclusively enriched terms present in CTD in each dataset was 44%, with Taylor leading with 448 exclusive terms (194 in CTD) and Varambally dropping to only 212 terms (98 in CTD, see Table S7)..

4. Discussion

Here it is shown that FA results may strongly vary depending on the method and parameters used, as previously noted in [43]. This could negatively influence the biological interpretation if not addressed appropriately. For instance, Subramanian's GSEA method showed high sensitivity to different parameter configurations and input data. The SMpp0 and SMpr methods returned almost no enriched terms for the analyzed datasets using the recommended statistical cutoff. These results were quite unexpected because the nature of the considered breast cancer subtypes has quite contrasting underlying biological mechanisms and highly opposing survival outcomes have been reported [14]. However, when the user has only an ordered list of genes, there is no other alternative than the pre-ranked version (SMpr) to perform GSEA. In this case, we suggest using the 1-p-value ranking and w=1 with different enrichment cutoffs, in order to obtain comparable results in terms of information retrieval between datasets (data not shown). When SM was fed with the expression matrix and its phenotype labels, it was found that both phenotype and gene permutations were very sensitive to the chosen weighting value, yielding different numbers of enriched terms as well as different levels of enrichment variability between datasets. The phenotype permutation with w=2 (SMpp2) seems to provide stable results, but the appearance of datasets with an extreme number of enriched terms discourages its use. In addition, when w=1 results were very unstable (high IQR). The gene permutation strategy with both w=1 and w=0 (SMgp1 and SMgp0) provided very stable results but the latter enriched almost twice as many terms as all the other used methods. When w=2 was used (SMgp2), it returned a low number of enriched terms. In disagreement with SM's authors, we recommend SMgp over SMpp. However, we agree with their recommended weighting value w=1, i.e., we recommend SMgp1 over any other SM configuration.

The mGSZ method was the most stable across datasets, yielding high consensus between datasets (high *FEC*) and providing the highest number of informative and exclusively enriched terms (EET_m). It was observed that, when no GS size upper limit was used, additional specific and informative terms were enriched. Thus, we encourage its use. Moreover, another advantage of mGSZ over SM is that the former has an up-to-date R implementation, whereas the latter requires a Java environment.

For SEA methodologies, although using BRI yields stable results and contains those terms enriched by the use of BRIII, the latter is more appropriate from a statistical point of view [25]. In addition, it was shown that BRIII does not present an outlying number of enriched terms, unlike BRI, but it has a more variable range of enriched terms over the datasets used. The implemented R version of DAVID's EASE score was developed to perform as similarly as DAVID web platform, with the advantage of using an up-to-date GO annotation database. Even more, any desired GS of interest can potentially be tested. Moreover, RD does not require either an internet connection or a DAVID registered account. In the case of GOstats, it was shown that it is too variable when testing enrichment across datasets (low *FEC*). This could turn problematic when only one dataset is analyzed, i.e., it would give a very limited biological insight of the experiment under analysis. For dEnricher algorithms, extreme values of enrichment were obtained when using Hypergeometric or Binomial; however, when using Fisher test, compliant results were retrieved. Moreover, when not applying any penalizing algorithm (dE_F_none) additional terms were obtained related to the disease under study. The dE_F_none resulted to be the most stable method across datasets for SEA alternatives (highest *FEC*) and outperformed its competitors in terms of the number of informative and exclusively enriched terms (EET_m).

In summary, we conclude that if parameters are properly set, FA retrieves consensual and meaningful biological information despite the low level of DE genes overlapped between datasets. It was demonstrated that both SEA and GSEA provided complementary results that could be integrated to gain biological insight. Moreover, their integration allows us to span the complete GO structure depth, a desirable feature when contrasting experimental conditions [25]. Accordingly, we propose to use the IFA framework which performs simultaneous SEA and GSEA analyses through dE_F_none and mGSZ, which resulted to be the most representative methods respectively. Both approaches are based upon the same linear model through the well-known limma library [27]. Thus, it provides a comprehensive and unified framework using only the expression matrix, the experimental design and (if GSs not provided) the GO database from *org.Hs.eg.db* [19]. Although this work was based upon the GO gene sets, any other knowledge base could be applied.

Applying IFA to study the terms differentially regulated between Luminal A and Basal-Like breast cancer subtypes resulted in several terms deeply related to breast cancer. For instance, those related to hormone and estrogen receptor signaling pathways are strongly related to the analyzed breast cancer subtypes, since Luminal A subjects are estrogen-dependent, whereas Basal-Like subjects are not [17,44,45]. From IFA results, concordant results revealed those terms associated with DNA unwinding process (Table S6), an event associated with the initiation of DNA synthesis and related to the facilitation of helicases activity. The helicase BACH1/FANCJ has been reported to be mutated in early onset breast cancer, especially linked to the hereditary breast cancer gene BRCA1 [46]. As most BRCA1-related breast cancers are both triple negative and Basal-Like [47], differences in the genes regulating DNA unwinding process between Luminal A and Basal-Like are highly expected. Furthermore, IFA revealed a group of terms that were only differentially regulated in TCGA dataset. Of them, the histone methyltransferase activity at H3-K9 was one of the deepest terms found in the GO tree structure. Histone H3-K9 methylation has been correlated with heterochromatin formation and transcriptional repression, which can regulate estrogen receptor (ER) expression [48]. In addition, the tyrosine kinase activity is involved in therapy circumvent in triple negative (Basal-Like) breast cancers [49]. The evaluation of the generalization capacity of IFA under prostate cancer datasets also returned concordant and informative results. Moreover, as expected and seen in the breast cancer case, it got enriched terms present in consensus between datasets, as well as specific enriched terms for each one.

These findings support the usefulness of our proposal from a biological data mining perspective. Furthermore, the proposed IFA

framework overcomes the limitations presented by the knowledge bases of the methods, minimizes the user defined parameters, facilitates the FA, allows the comparison of different patient cohorts, as shown with TCGA and prostate cancer results, and is freely provided as an R source code (at Github).

5. Funding

This work was supported by grants from the following Argentine institutions: Universidad Católica de Córdoba (BOD/2016 to EAF), Ministerio de Ciencia, Tecnología e Innovación Productiva (PPL 6/2011 to EAF), Secretaría de Ciencia y Tecnología - Universidad Nacional de Córdoba (30720150101719CB to EAF) and the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

6. Conflict of interest

None declared.

Acknowledgments

The authors would like to thank the contribution of three anonymous reviewers who helped us to improve the quality of this work. We would also like to thank Roxana Schillaci, PhD and Patricia Elizalde, PhD for assistance with the biological interpretation that greatly improved the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.compbiomed.2016.09.017.

References

[1] J.S. Reis-Filho, L. Pusztai, Gene expression profiling in breast cancer: classification, prognostication, and prediction, Lancet 378 (9805) (2011) 1812–1823.
[2] J.J. Goeman, P. Bühlmann, Analyzing gene expression data interms of gene sets: methodological issues, Bioinformatics 23 (8) (2007) 980–987.
[3] H. Maciejewski, Gene set analysis methods: statistical models and methodological differences, Brief. Bioinforma. (2013) [bbt002].
[4] T. Manoli, N. Gretz, H.-J. Gröne, M. Kenzelmann, R. Eils, B. Brors, Group testing for pathway analysis improves comparability of di erent microarray datasets, Bioinformatics 22 (20) (2006) 2500–2506.
[5] P. Pavlidis, J. Qin, V. Arango, J.J. Mann, E. Sibille, Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex, Neurochem. Res. 29 (6) (2004) 1213–1222.
[6] D.W. Huang, B.T. Sherman, R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucleic Acids Res. 37 (1) (2009) 1–13.
[7] A. Subramanian, et al., P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, Gene set enrichment analysis: a knowledge based approach for interpreting genome-wide expression pro-les, Proc. Natl. acad. Sci. 102 (43) (2005) 15545–15550.
[8] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, P. J. Park, Discovering statistically significant pathways in expression profiling studies, Proceedings of the National Academy of Sciences of the United States of America 102 (38) (2005) 13544–13549.
[9] P. Khatri, M. Sirota, A.J. Butte, Ten years of pathway analysis: current approaches and outstanding challenges, PLoS Comput. Biol. 8 (2) (2012) e1002375.
[10] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., Gene ontology: tool for the unification of biology, Nat. Genet. 25 (1) (2000) 25–29.
[11] L. Ein-Dor, O. Zuk, E. Domany, Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer, Proceedings of the National Academy of Sciences 103 (15) (2006) 5923–5928.
[12] C.M. Perou, T. Sørlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, et al., Molecular portraits of human breast tumours, Nature 406 (6797) (2000) 747–752.
[13] A.P. Davis, C.J. Grondin, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B.L. King, T.C. Wiegers, C.J. Mattingly, The comparative toxicogenomics database's 10th year anniversary: update 2015, Nucleic Acids Res. (2014) [gku935].
[14] J.S. Parker, M. Mullins, M.C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, et al., Supervised risk predictor of breast cancer based on intrinsic subtypes, J. Clin. Oncol. 27 (8) (2009) 1160–1167.
[15] D.M. Gendoo, N. Ratanasirigulchai, M.S. Schröder, L. Paré, J.S. Parker, A. Prat, B. Haibe-Kains, Genefu: an r/bioconductor package for computation of gene expression-based signatures in breast cancer, Bioinformatics (2015) [btv693].
[16] T. Sørlie, E. Borgan, S. Myhre, H.K. Vollan, H. Russnes, X. Zhao, G. Nilsen, O.C. Lingjærde, A.-L. Børresen-Dale, E. Rødland, The importance of gene-centring microarray data, Lancet Oncol. 11 (8) (2010) 719–720.
[17] X. Dai, T. Li, Z. Bai, Y. Yang, X. Liu, J. Zhan, B. Shi, Breast cancer intrinsic subtype classification, clinical use and future trends, Am. J. Cancer Res. 5 (10) (2015) 2929.
[18] L. Ein-Dor, I. Kela, G. Getz, D. Givol, E. Domany, Outcome signature genes in breast cancer: is there a unique set?, Bioinformatics 21 (2) (2005) 171–178.
[19] M. Carlson, S. Falcon, H. Pages, N. Li, org. hs. eg. db: Genome wide annotation for human, 2013.
[20] D.W. Huang, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using david bioinformatics resources, Nat. Protoc. 4 (1) (2009) 44–57.
[21] C. Fresno, E.A. Fernández, Rdavidwebservice: a versatile r interface to david, Bioinformatics (2013) [btt487].
[22] S. Falcon, R. Gentleman, Using gostats to test gene lists for go term association, Bioinformatics 23 (2) (2007) 257–258.
[23] H. Fang, J. Gough, Thednetapproach promotes emerging research on cancer patient survival, Genome Med. 6 (8) (2014) 1.
[24] I. Rivals, L. Personnaz, L. Taing, M.-C. Potier, Enrichment or depletion of a go category within a class of genes: which test?, Bioinformatics 23 (4) (2007) 401–407.
[25] C. Fresno, A.S. Llera, M.R. Girotti, M.P. Valacco, J.A. López, O.L. Podhajcer, M.G. Balzarini, F. Prada, E.A. Fernández, The multi-reference contrast method: facilitating set enrichment analysis, Comput. Biol. Med. 42 (2) (2012) 188–194.
[26] D.J. McCarthy, G.K. Smyth, Testing significance relative to a fold-change threshold is a treat, Bioinformatics 25 (6) (2009) 765–771.
[27] C. Berkeley, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, E-book available at ⟨http://www.bepress.com/sagmb/vol3/iss1/art3.[PubMed]⟩.
[28] A.J. Minn, G.P. Gupta, D. Padua, P. Bos, D.X. Nguyen, D. Nuyten, B. Kreike, Y. Zhang, Y. Wang, H. Ishwaran, et al., Lung metastasis genes couple breast tumor size and metastatic spread, Proceedings of the National Academy of Sciences 104 (16) (2007) 6740–6745.
[29] Y. Wang, J.G. Klijn, Y. Zhang, A.M. Sieuwerts, M.P. Look, F. Yang, D. Talantov, M. Timmermans, M.E. Meijer-van Gelder, J. Yu, et al., Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, Lancet 365 (9460) (2005) 671–679.
[30] M.J. Van De Vijver, Y.D. He, L.J. van't Veer, H. Dai, A.A. Hart, D.W. Voskuil, G.J. Schreiber, J.L. Peterse, C. Roberts, M.J. Marton, et al., A gene-expression signature as a predictor of survival in breast cancer, New Engl. J. Med. 347 (25) (2002) 1999–2009.
[31] L.J. Van't Veer, H. Dai, M.J. Van De Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (6871) (2002) 530–536.
[32] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M.S. d'Assignies, et al., Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series, Clin. Cancer Res. 13 (11) (2007) 3207–3214.
[33] K.D. Miller, L.I. Chap, F.A. Holmes, M.A. Cobleigh, P.K. Marcom, L. Fehrenbacher, M. Dickler, B.A. Overmoyer, J.D. Reimann, A.P. Sing, et al., Randomized phase iii trial of capecitabine compared with bevacizumab plus capecitabine in patients with previously treated metastatic breast cancer, J. Clin. Oncol. 23 (4) (2005) 792–799.
[34] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, et al., Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis, J. Natl. Cancer Inst. 98 (4) (2006) 262–272.
[35] M. Schmidt, D. Böhm, C. von Törne, E. Steiner, A. Puhl, H. Pilch, H.-A. Lehr, J.G. Hengstler, H. Kölbl, M. Gehrmann, The humoral immune system has a key prognostic impact in node-negative breast cancer, Cancer Res. 68 (13) (2008) 5405–5413.
[36] P. Mishra, P. Törönen, Y. Leino, L. Holm, Gene set analysis: limitations in popular existing methods and proposed improvements, Bioinformatics 30 (19) (2014) 2747–2756.
[37] J. Oksanen, R. Kindt, P. Legendre, B. OHara, M.H.H. Stevens, M. J. Oksanen, M. Suggests, The vegan package, Community ecology package 10.
[38] E. Edelman, A. Porrello, J. Guinney, B. Balakumaran, A. Bild, P.G. Febbo, S. Mukherjee, Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles, Bioinformatics 22 (14) (2006) e108–e116.
[39] H. Ross-Adams, A. Lamb, M. Dunning, S. Halim, J. Lindberg, C. Massie, L. Egevad, R. Russell, A. Ramos-Montoya, S. Vowler, et al., Integration of copy number and transcriptomics provides risk stratification in prostate cancer: a discovery and validation cohort study, EBioMedicine 2 (9) (2015) 1133–1144.
[40] B.S. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, B.S. Carver, V.K. Arora, P. Kaushik, E. Cerami, B. Reva, et al., Integrative genomic profiling of human prostate cancer, Cancer Cell 18 (1) (2010) 11–22.
[41] S. Varambally, J. Yu, B. Laxman, D.R. Rhodes, R. Mehra, S.A. Tomlins, R.B. Shah, U. Chandran, F.A. Monzon, M.J. Becich, et al., Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression, Cancer Cell 8 (5) (2005) 393–406.
[42] C.S. Grasso, Y.-M. Wu, D.R. Robinson, X. Cao, S.M. Dhanasekaran, A.P. Khan,

M.J. Quist, X. Jing, R.J. Lonigro, J.C. Brenner, et al., The mutational landscape of lethal castration-resistant prostate cancer, Nature 487 (7406) (2012) 239–243.
[43] J.C. Rodriguez, G. González, C. Fresno, E.A. Fernández, Integrative functional analysis improves information retrieval in breast cancer, in: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer, 2015, pp. 43–50.
[44] A. Goldhirsch, E.P. Winer, A. Coates, R. Gelber, M. Piccart-Gebhart, B. Thürlimann, H.-J. Senn, K.S. Albain, F. André, J. Bergh, et al., Personalizing the treatment of women with early breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2013, Ann. Oncol. 24 (9) (2013) 2206–2223.
[45] R.R. Bastien, Á Rodríguez-Lescure, M.T. Ebbert, A. Prat, B. Munárriz, L. Rowe, P. Miller, M. Ruiz-Borrego, D. Anderson, B. Lyons, et al., Pam50 breast cancer subtyping by rt-qpcr and concordance with standard clinical molecular markers, BMC Med. Genom. 5 (1) (2012) 1.
[46] S.B. Cantor, D.W. Bell, S. Ganesan, E.M. Kass, R. Drapkin, S. Grossman, D.C. Wahrer, D.C. Sgroi, W.S. Lane, D.A. Haber, et al., Bach1, a novel helicase-like protein, interacts directly with brca1 and contributes to its dna repair function, Cell 105 (1) (2001) 149–160.
[47] D.P. Atchley, C.T. Albarracin, A. Lopez, V. Valero, C.I. Amos, A.M. Gonzalez-Angulo, G.N. Hortobagyi, B.K. Arun, Clinical and pathologic characteristics of patients with brca-positive and brca-negative breast cancer, J. Clin. Oncol. 26 (26) (2008) 4282–4288.
[48] D. Sharma, J. Blum, X. Yang, N. Beaulieu, A.R. Macleod, N.E. Davidson, Release of methyl cpg binding proteins and histone deacetylase 1 from the estrogen receptor α (er) promoter upon reactivation in er-negative human breast cancer cells, Mol. Endocrinol. 19 (7) (2005) 1740–1751.
[49] M. Scaltriti, M. Elkabets, J. Baselga, Molecular pathways: Axl, a membrane receptor mediator of resistance to therapy, Clinical Cancer Research (2016) clincanres–1458.

Elmer A. Fernández: He is a Biomedical Engineer with a PhD in Advanced Computing with more than 10 years in research experiencie in the Data Mining in Biomedicine field. He leads the Bioscience Data Mining Group (BDMG) at Catholic University of Córdoba. He is a staff researcher at CONICET and a director of the Biodata Mining Node of the National Bioinformatic Platform (Bioinformatics Argentina). His main research interests are Data Mining and Big-Data Analytics in Biological Sciences.

Elmer A. Fernández (EAF) conceived the original idea and conceived the experiment. Juan Cruz Rodriguez (JCR) introduced original concepts to the proposal, developed and performed the computational analyses. German A. González (GAG) and Cristóbal Fresno (CF) contributed to the formal methodology and in the algorithms development. Andrea S. Llera (ASLL) performed the biological interpretation and provided significant feedback to the development of the proposal. EAF, JCR, GAG, CF, and ASLL wrote the paper and reviewed the manuscript. All authors agreed with the final version of the manuscript.

Juan Cruz Rodriguez: He has a degree in Computer Science. He is currently doing his doctorate in the BDMG lab. His main interest is computer development of big-data analytics systems in bioscience.

Germán A. Gonzalez: He has a degree in Bioinformatics from the National Unirversity of Entre Rios. He is currently working as a technical assistant at BDMG in gene expresison analysis and Data Base administration.

Cristobal Fresno: He got his PhD at BDMG. He is currently finishing his Postdoc in collaboration with the BDMG and the Proteomic Branch of the Molecular and Cellular Therapy Group at Leloir Institute Foundation, lead by PhD. Andrea S. Llera. His main interests are bioinformatic applications for Breast Cancer research.

Andrea S. Llera: She has a PhD in Inmunology from National University of Buenos Aires with more than 25 years in research experience and more than 5 years of experience in genomics. She is an Independent Researcher at CONICET and the director of the Institute Leloir's node of the Argentine Consortium of Genomics Technology and of the proteomic laboratory in the molecular and cellular therapy group at Leloir Institute Foundation.