Journal of Genetics and Genomics xxx (2018) 1–16

Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Original research

CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways

Chen Ai, Lei Kong*

Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 26 February 2018
Received in revised form
11 August 2018
Accepted 13 August 2018
Available online xxx

Keywords:
Gene expression
Differential expression
Gene set enrichment
Support vector machine

ABSTRACT

Gene set enrichment (GSE) analyses play an important role in the interpretation of large-scale transcriptome datasets. Multiple GSE tools can be integrated into a single method as obtaining optimal results is challenging due to the plethora of GSE tools and their discrepant performances. Several existing ensemble methods lead to different scores in sorting pathways as integrated results; furthermore, it is difficult for users to choose a single ensemble score to obtain optimal final results. Here, we develop an ensemble method using a machine learning approach called Combined Gene set analysis incorporating Prioritization and Sensitivity (CGPS) that integrates the results provided by nine prominent GSE tools into a single ensemble score (R score) to sort pathways as integrated results. Moreover, to the best of our knowledge, CGPS is the first GSE ensemble method built based on a priori knowledge of pathways and phenotypes. Compared with 10 widely used individual methods and five types of ensemble scores from two ensemble methods, we demonstrate that sorting pathways based on the R score can better prioritize relevant pathways, as established by an evaluation of 120 simulated datasets and 45 real datasets. Additionally, CGPS is applied to expression data involving the drug panobinostat, which is an anticancer treatment against multiple myeloma. The results identify cell processes associated with cancer, such as the p53 signaling pathway (*hsa04115*); by contrast, according to two ensemble methods (EnrichmentBrowser and EGSEA), this pathway has a rank higher than 20, which may cause users to miss the pathway in their analyses. We show that this method, which is based on a priori knowledge, can capture valuable biological information from numerous types of gene set collections, such as KEGG pathways, GO terms, Reactome, and BioCarta. CGPS is publicly available as a standalone source code at ftp://ftp.cbi.pku.edu.cn/pub/CGPS_download/cgps-1.0.0.tar.gz.

1. Introduction

High throughput technologies, such as RNA-Seq and microarray, provide efficient methods for acquiring whole transcriptional profiles. To extract the biological meaning from profiles of thousands of genes on a functional level, gene set enrichment (GSE) analyses have been developed. The GSE analyses are typically used to identify groups of genes that are related to each other (i.e., genes belonging to the same pathway or gene set) based on expression data. Each pathway or gene set is ultimately assigned a pathway-level statistic by the GSE method to assess its relevance to the biological conditions under study. After sorting all pathways based on the statistics, the relevant pathways are expected to be ranked at the top position, allowing researchers to easily identify noteworthy pathways. These statistics can be significance (p-value), rank, or enrichment score depending on the specific GSE method. For example, GSEA (Subramanian et al., 2005) assigns an enrichment score and p-value to each pathway, and the pathways are ranked by either the value of the enrichment score or p-value. The most commonly used statistics among the GSE algorithms are the p-value and rank.

A number of GSE methods have been published, and these methods are divided into three generations according to their features as reviewed (Khatri et al., 2012). The first-generation GSE

* Corresponding author.
E-mail address: kongl@mail.cbi.pku.edu.cn (L. Kong).

https://doi.org/10.1016/j.jgg.2018.08.002

method named over-representation analysis (ORA) is most commonly used for GSE analyses. It utilizes some naïve statistics such as the hypergeometric test, and works by assessing differentially expressed (DE) genes with top ranks at a particular significance threshold. However, this approach uses predefined thresholds to identify the most significant genes and ignores the association of their expression values (Goeman and Bühlmann, 2007). The second-generation GSE methods, named functional class scoring or set-based methods, apply a gene-level statistic, such as the correlation between the expression value and the experimental conditions. Subsequently, a pathway-level statistic is applied, and the statistical significance is assessed. Set-based methods can avoid the artificial threshold of DE genes but do not incorporate gene-gene interactions in the analysis (Barry et al., 2005; Subramanian et al., 2005). The third-generation GSE methods, which are called network-based or pathway topology methods, incorporate the pathway topology or gene regulatory network and treat the genes unequally according to their roles in the network (Fang et al., 2012b; Gu and Wang, 2013). Some investigations have demonstrated the benefit of these network-based methods (Fang et al., 2012a, 2012b).

Importantly, the performance of these methods is quite distinct. In 2013, Tarca et al. (2013) compiled 42 real disease gene expression datasets, and for each disease, a known disease-specific pathway served as a "target pathway" for comparing the performance of the GSE methods assessed. These authors used the *p*-values of the target pathway to evaluate the sensitivity of each GSE method since a sensitive method could produce small *p*-values for pathways that are indeed relevant to a given condition. These authors also used the rank of the target pathway to evaluate prioritization, and methods with high prioritization ability could rank relevant gene sets close to the top. They showed that the performance of the 16 GSE methods was quite distinct. Some methods, including globaltest (Goeman et al., 2004) and plage (Tomfohr et al., 2005), had high sensitivity but a low prioritization ability. In contrast, the method padog (Tarca et al., 2012) performed well in prioritization but had low sensitivity. These observations demonstrated that different GSE methods could produce substantially different results using the same dataset. These discrepancies highlight two questions that must be addressed to extract meaningful pathways from the results. First, which GSE method should be used for a specific dataset to obtain optimal results? Second, should we use the *p*-value or rank as the threshold to obtain the enriched pathway? Due to the plethora of GSE methods and their distinct performances, selecting the proper tool and/or criteria for analysis is challenging.

Recently, two ensemble GSE methods were proposed to solve the first problem by integrating the results of multiple individual GSE methods. For instance, EnrichmentBrowser (Geistlinger et al., 2016) utilizes naïve statistics, such as the mean, sum, or median, to integrate gene set rankings from both set-based methods and net-based methods into an ensemble score (e.g., average rank). Then, the gene sets are sorted based on this score to generate the final integrated results. Similarly, EGSEA (Alhamdoosh et al., 2017) utilizes the results from multiple set-based methods and integrates these results into an ensemble score, such as median rank, average rank, vote rank, combined *p*-value, minimum *p*-value, and/or significance score. An advantage of EGSEA over EnrichmentBrowser is that EGSEA provides not only ensemble scores integrating the gene set rankings but also the gene set *p*-values and gene fold change. For instance, the combined *p*-value in EGSEA is a combination of *p*-values from multiple methods, and the significance score in EGSEA assigns high scores to gene sets with strong fold-changes and high statistical significance. However, although these two ensemble methods integrate results obtained using multiple methods, the various ensemble scores provided make it difficult to choose the most reliable score and identify the best and most comprehensive result. Thus, a method that can integrate comprehensive information into a single reliable score needs to be explored.

To overcome the shortcomings mentioned above, we propose a new ensemble GSE method called Combined Gene set analysis incorporating Prioritization and Sensitivity (CGPS). CGPS simultaneously integrates the *p*-value and rank obtained from nine prominent GSE methods into a reliable ensemble score (R score) and then generates a new gene set ranking based on the R score, representing a powerful approach for prioritizing biologically relevant gene sets. The comparative analyses of the R score using multiple datasets demonstrated that the CGPS is better than EnrichmentBrowser, EGSEA, and 10 individual methods in prioritizing relevant gene sets. We also showed that the R score can reflect the differential expression levels of a gene set. Furthermore, we showed that a priori knowledge of pathways and phenotypes, which is fundamental for CGPS, plays an important role in its performance. Moreover, we applied CGPS to a gene expression dataset related to the anticancer drug panobinostat. The results indicated that CGPS prioritized important biological pathways relevant to this drug, such as the p53 and TGF-beta signaling pathways, even though these pathways were not ranked near the top using EnrichmentBrowser or EGSEA. Finally, we also applied CGPS to acute lymphoblastic leukemia, which is not drug-related. Through this application, we demonstrated that CGPS can be applied to various gene set collections in addition to the KEGG pathway database.

2. Results

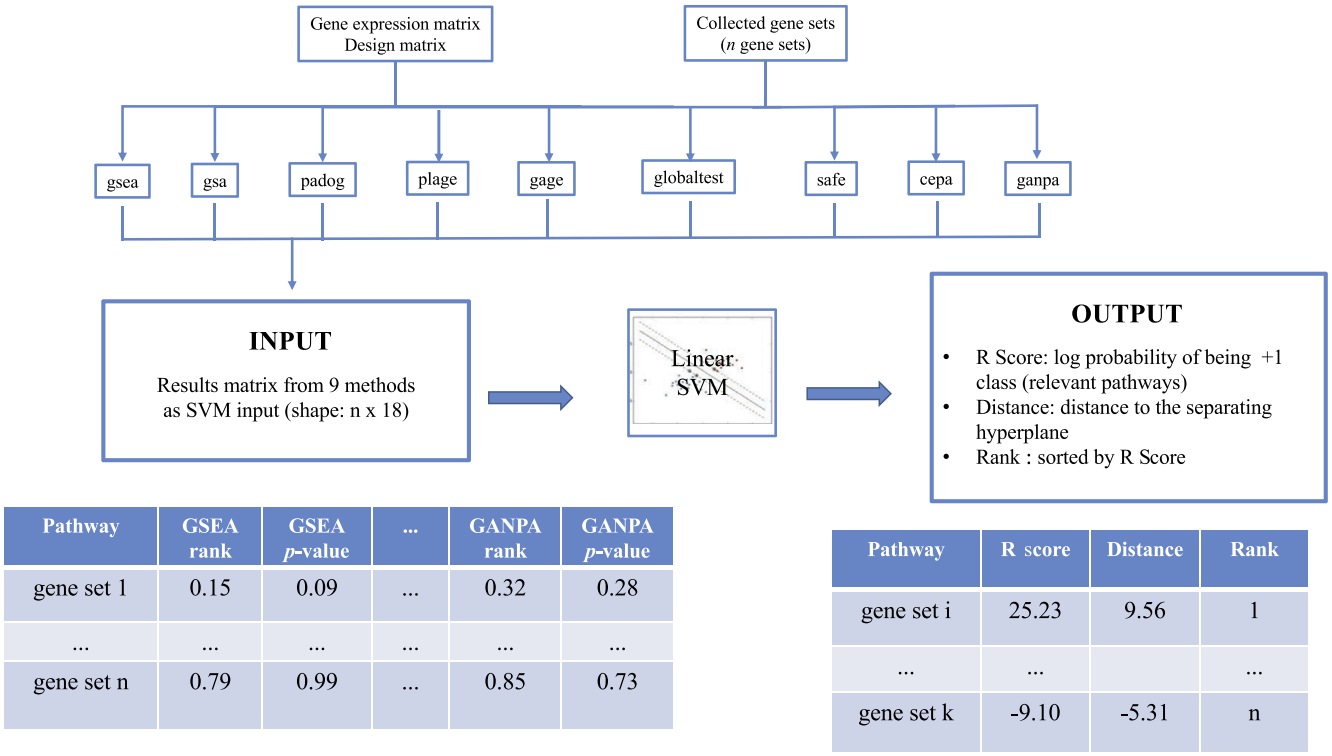
2.1. Overview of CGPS

To reliably and comprehensively integrate results obtained from multiple methods, we propose an ensemble method named CGPS. The workflow of CGPS is shown in Fig. 1. Given the *p*-values and rankings of a gene set obtained from multiple GSE methods, the core of CGPS, which is a linear support vector machine (SVM), can predict the relevance of a gene set to an experimental condition and produce an R score to measure its relevance. The relevant gene sets are usually denoted by a high positive R score value. CGPS sorts all gene sets in a gene set collection based on their R score to prioritize the relevant gene sets that users usually hope to find. The SVM combines nine GSE methods, including seven set-based methods, i.e., gsea (Subramanian et al., 2005), gsa (Efron and Tibshirani, 2007), padog (Tarca et al., 2012), plage (Tomfohr et al., 2005), gage (Luo et al., 2009), globaltest (Goeman et al., 2004), and safe (Barry et al., 2005), and two network-based methods, i.e., cepa (Gu and Wang, 2013) and ganpa (Fang et al., 2012b).

The SVM was constructed based on real gene expression datasets. We collected datasets related to drug treatment experiments performed in human and mouse cell lines, and each dataset involved a drug with known target pathways. These target pathways received a positive class (+1) label to train the SVM, and the features of each pathway included the *p*-value and ranking obtained using the nine individual methods. In contrast, the pathways that had a negative class (−1) label in training the SVM showed little gene expression difference between the case and the control sample groups (see details in Materials and methods). In total, 84 datasets, involving 255 target pathways and 1038 negative class pathways, were used to train the SVM. Additionally, 21 datasets involving 51 target pathways were employed to test and evaluate the SVM. The procedures used for the training and testing of the SVM are presented in greater details in the section of Materials and methods.

In the following sections, we investigate whether sorting by the R score properly prioritizes the relevant pathways and compare

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002



Pathway	GSEA rank	GSEA *p*-value	...	GANPA rank	GANPA *p*-value
gene set 1	0.15	0.09	...	0.32	0.28
...	...	...	...	...	...
gene set n	0.79	0.99	...	0.85	0.73

Pathway	R score	Distance	Rank
gene set i	25.23	9.56	1
...	...		...
gene set k	-9.10	-5.31	n

Fig. 1. Workflow of CGPS. Given gene expression data and the sample design information of a dataset, the gene set enrichment analysis is conducted using the nine GSE methods. Receiving these results, CGPS generates a comprehensive pathway table of the rank (prioritization) and *p*-value (sensitivity) produced by each method. This 18-dimensional pathway table is used as input for the support vector machine (SVM) algorithm. In the output results, a pathway's R score is computed as a log transformation of the probability of being a positive class pathway. Finally, the pathways are sorted based on the R score in descending order in the results list.

CGPS's results with the results obtained by using EnrichmentBrowser, EGSEA, ORA and the nine GSE methods. The approach used to evaluate the prioritization ability of the GSE methods uses real datasets followed the methods described by Tarca et al. (2012, 2013) based on known target pathways. These researchers assumed that a successful method could rank the relevant pathways close to the top and compared the median rank of the target pathways obtained from 16 GSE methods based on their benchmark datasets, i.e., *KEGGdzPathwaysGEO*. This approach has been used in many published works (Dong et al., 2016; Geistlinger et al., 2016). Notably, the drug-treated datasets in the following results refer to the 21 datasets used for testing and consisted of 11 RNA-Seq datasets and 10 microarray datasets. Additionally, KEGG pathway datasets (Kanehisa and Goto, 2000) were used to analyze all real datasets, including 305 pathways in human and 301 pathways in mouse.

2.2. Comparison of CGPS, EnrichmentBrowser and EGSEA using RNA-Seq drug-treated datasets and simulated datasets

As CGPS is an ensemble method, we first compared CGPS with two other ensemble GSE methods, EnrichmentBrowser and EGSEA. For CGPS, we used only the ensemble score (R score) to rank the gene sets. For EGSEA, we used four representative scores, including the average rank, the vote rank (the majority rank), the significance score (the log fold change in gene expression), and the combined *p*-value (a transformation and combination of *p*-values obtained from the individual methods), to generate the gene set rankings. For EnrichmentBrowser, we used the average rank as the representative ensemble score because the other statistics, such as the min rank, max rank, and sum rank, were all naïve statistics in terms of rank. Then, we compared the prioritization ability of six different ensemble scores obtained from three different ensemble GSE methods. We used real datasets and simulated datasets to evaluate the prioritization ability in two aspects. First, we used the RNA-Seq datasets in the collected drug-treated datasets to evaluate whether ranking the pathways by the ensemble score could prioritize the relevant pathways. Second, we used the simulated datasets to evaluate the false discovery rate of the gene set rankings.

2.2.1. RNA-Seq drug-treated datasets

In this study, we only used 11 RNA-Seq datasets among the 21 drug-treated datasets because EGSEA cannot be applied to microarray data according to a publication by Alhamdoosh et al. (2017). Each dataset contains expression data from the following two groups of samples: one group treated with a drug and the other untreated. The drug has several known target pathways according to the KEGG DRUG (Kanehisa et al., 2010) database (http://www.genome.jp/kegg/drug/). In total, we identified 29 target pathways within the 11 datasets.

For a given method, we first sorted the pathways by their ensemble score within each dataset, and then, we divided the absolute rank of the target pathway by the total number of pathways to generate a decimal rank ranging from 0 to 1. A low value demonstrated that the target pathway was ranked close to the top, suggesting that the method successfully prioritized the relevant pathways.

Among the six types of ensemble scores compared (Fig. 2), the R score provided by CGPS achieved the best median rank (0.223) for the 29 target pathways, while the average rank provided by EGSEA, average rank provided by EnrichmentBrowser, significance score provided by EGSEA, combined *p*-value provided by EGSEA, and vote rank provided by EGSEA were 0.272, 0.315, 0.334, 0.338, and 0.446, respectively. Furthermore, CGPS ranked more than half of the 29

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002

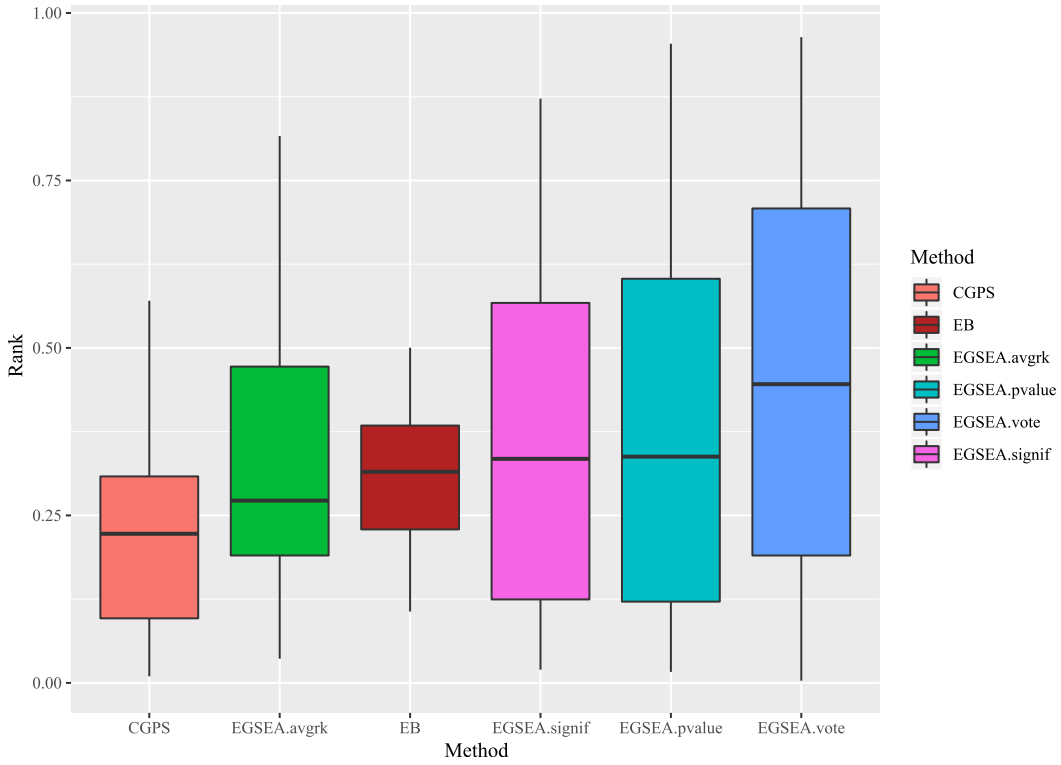


Fig. 2. Comparison of the ensemble GSE methods using the RNA-Seq drug-treated datasets. The boxes show the distributions of the ranks of the 29 target pathways received from CGPS, EnrichmentBrowser (EB), and the four different ranking methods of EGSEA. The four ranking methods of EGSEA are the average rank, vote rank, significant score, and combined *p*-value. EnrichmentBrowser uses the average rank method to sort the pathways. The lower values reflect the method that performs better in prioritizing the relevant pathways. The methods are ordered from best to worst according to the median rank value (represented by the bold line in the box) of the 29 target pathways. The median rank given by the six methods from left to right is 0.223, 0.272, 0.315, 0.334, 0.338, and 0.446.

target pathways in the first quarter among all pathways, while the other five scores did not achieve this ranking in the analysis. Compared with the other five scores, the 29 ranks obtained from the 11 different datasets provided by CGPS (R score) seemed to be concentrated at approximately 0.25 and below 0.40, showing that the R score could consistently prioritize target pathways from different datasets. Altogether, these observations indicate that the R score provided by CGPS performed the best among these six ensemble scores in ranking the target pathways near the top.

2.2.2. Simulated datasets

To compare the performance of CGPS with that of other methods in different settings, a cut-off threshold of 60 (top-ranked affected gene sets) was used to evaluate each method's retrieval power. The false discovery rate (FDR) of each method was calculated in each configuration of simulated datasets to evaluate the retrieval power (Alhamdoosh et al., 2017). Through the simulated datasets, the effect of the differential expression level on performance was investigated by using fold change (FC) and detection call (DC, the percentage of DE genes in affected gene sets) as variables. The FC varied over four levels between 1.3 and 2.0, and the DC varied over three levels (20%, 30%, and 50%). The different configurations of FC and DC provided a comprehensive and sensitive way to display their effects on a method's FDR. The FDRs of CGPS, EnrichmentBrowser, and EGSEA and the standard deviation of 10 datasets in each configuration were calculated (Table S1). As either the FC or DC increased, the values of the FDR of all methods decreased, indicating that the performance of all methods improved as the level of differential expression increased. Of the 12 configurations, CGPS's R score achieved the second best FDR average (0.134) following the FDR of EnrichmentBrowser's average rank (0.127), but the R score was better than the FDR of EGSEA's average rank, vote rank, significance, and combined *p*-value (0.142, 0.216, 0.254, and 0.259, respectively). Additionally, as the value of the FC increased, the absolute difference in the FDRs among the six methods decreased. At the lowest FC of 1.3 and DC of 0.2, CGPS (FDR = 0.385) greatly outperformed the four EGSEA methods (none of the four FDRs were below 0.4), with a slightly higher score (0.017) than EnrichmentBrowser (FDR is 0.368). CGPS was among the top two best performing methods of the six configurations (Table S1), indicating that CGPS has good retrieval power.

In summary, the evaluation of the RNA-Seq drug-treated datasets and simulated datasets (real datasets) demonstrated that CGPS's R score prioritized the relevant pathways better than EnrichmentBrowser or EGSEA's four representative methods, and the evaluation of the simulated datasets demonstrated that CGPS's R score was among the top two methods due to its low FDR value.

2.3. Comparison of CGPS and nine individual methods using the drug-treated datasets

To compare the nine individual methods integrated into CGPS, we applied 21 drug-treated datasets, including the 11 RNA-Seq datasets described above and 10 microarray datasets. Similarly, each drug within the datasets has several known KEGG target pathways, and these datasets were associated with 51 target pathways in the aggregate. Using the same strategy described above, we investigate the ranks of the 51 target pathways obtained from a given method and compare the median rank of the 51 target pathways obtained from the nine individual methods and CGPS.

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002

Compared with the nine individual methods, CGPS ranked the 1st with the lowest median rank value (0.176), which was smaller than that obtained using the 2nd ranked method, i.e., padog (0.200; Fig. 3). Only CGPS ranked more than half (27 of 51) of the target pathways in the top 20%, while the other methods ranked less than half (at most 25) in the top 20%. Moreover, CGPS ranked only 6 target pathways below the 50% mark; in contrast, 13 target pathways were ranked below this mark by padog, which was the best method of the nine individual methods. These observations demonstrated that CGPS was better than the other nine individual methods in prioritizing the target pathways.

To investigate the influence of excluding the best and worst method, we excluded padog or gage and trained two SVMs using the same training sets as CGPS. Then, we applied the two SVMs to the drug-treated datasets. The SVM "noPadog" provided a slightly worse median rank for the 51 target pathways (0.180 *vs.* CGPS's 0.176) and a better median rank than padog (0.200). This finding demonstrated that our method of integrating individual results has the ability to generate a better result than the individual methods alone, and this ability is mainly due to the integration of the individual methods and is not entirely dependent on the best individual method. The SVM "noGage" (median 0.163) performed slightly better than CGPS and much better than gage (Fig. S1). This finding suggested that the worst method could influence the model's performance, but the influence on our drug-treated datasets was small. In general, there was only a slight impact on the testing datasets following the removal of the best or worst method, and the performance of CGPS was primarily attributed to the integration of all methods rather than the best method.

2.4. Comparison of CGPS and ORA using the drug-treated datasets

The ORA method is the first-generation method that has been widely used for a long time. However, one of its limitations is that it considers the most significant gene input using an arbitrary threshold and discards the other genes, while second- and third-generation methods fully use the expression data of all genes. To avoid the biases of the ORA method, we did not incorporate it into CGPS since we have used this method to select the pathways of the training datasets. As the ORA method is the most commonly used GSE method, we also compared it to CGPS (Fig. 3). We used the same drug-treated datasets described above. The median rank value of ORA was 0.223, which is higher than that of CGPS, suggesting that CGPS outperformed ORA in prioritizing the relevant pathways.

2.5. Comparison of CGPS, EnrichmentBrowser, ORA and the nine individual methods using benchmark microarray datasets

To investigate whether CGPS performed consistently well in the prioritization of relevant pathways based on well-accepted benchmark datasets, we use 24 benchmark datasets in the R package *KEGGdzPathwaysGEO* compiled by Tarca et al. (2013) for further comparison. The 24 datasets are subsets of 42 disease datasets, and their target pathways are all listed in the KEGG pathway collection. We use this set of benchmark datasets because it has been widely used in comparisons of GSE methods in many prior reports (Bayerlová et al., 2015; Dong et al., 2016) and is well accepted in this field. Additionally, these datasets are quite different from our drug-treated cell line datasets since they are disease-related gene expression datasets generated from patient samples.

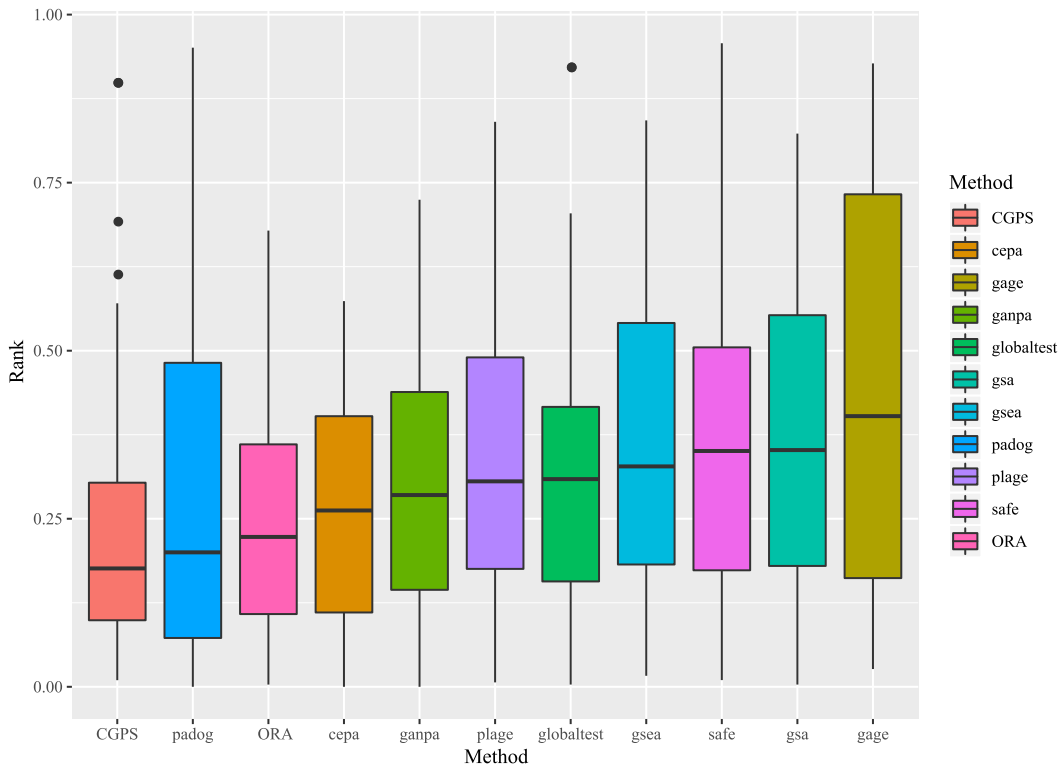


Fig. 3. Comparison of CGPS, ORA and the nine individual methods using all drug-treated datasets. The boxes show the distributions of the ranks of 51 target pathways received from CGPS, ORA and the nine individual methods. All methods were applied to the drug-treated datasets. Lower values indicate that the method has a better prioritization ability. Methods are ordered from best to worst according to their median rank of the 51 target pathways. The median rank of CGPS, padog, and ORA is 0.176, 0.200, and 0.223, respectively. The median rank of the other methods are all above 0.250.

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002

We aimed to investigate whether CGPS could also prioritize the target pathways better than the other methods using such datasets. Thus, we applied all methods mentioned above to these datasets, except for EGSEA, because EGSEA was designed for RNA-Seq data and is unable to process microarray data (Alhamdoosh et al., 2017).

CGPS again achieved the best prioritization ability with the lowest median rank value, i.e., 0.138 (Fig. 4), while padog and EnrichmentBrowser ranked the 2nd and the 3rd with median rank values of 0.143 and 0.190, respectively. This comparison indicated that CGPS provided better prioritization of the relevant pathways, consistently outperforming the other prominent GSE methods using different datasets.

2.6. R score increases as the level of differential expression increases

CGPS is an ensemble method used to integrate results obtained from individual methods into a single measure, i.e., an R score. We demonstrated that ranking pathways by the R score can prioritize pathways relevant to a given experimental condition. Furthermore, to investigate whether the R score can serve as an indicator of relevance, we designed simulated datasets with various levels of differential expression. Here, we only consider the fold change of genes in a pathway and the content of DE genes in a pathway as these factors can reflect the differential expression level in a pathway. We applied CGPS to the simulated datasets with 12 different expression-level configurations.

The configuration of the simulated RNA-Seq datasets involved the following two parameters: FC and DC (i.e., the percentage of genes that are differentially expressed within a pathway). The 12 different configurations of FC and DC are combinations of FC = {1.3, 1.5, 1.8, 2.0} and DC = {20%, 30%, 50%}. Each dataset contains 200 gene sets, and 60 of these gene sets are affected gene sets. For more details regarding the simulated datasets, please refer to the Materials and methods section.

We used the mean value of the top 60 gene sets' R scores as the representative R score value of each configuration. Then, we compared the representative R score value under different configurations of expression levels and analyzed the relationship among the R score, FC and DC. For each configuration, we included 10 batches of sets to calculate the mean representative R score value and standard deviation (Table 1). The representative R score value

Table 1
Mean and standard deviation of representative R scores of differential expression levels in the simulated data.

FC	DC	R score	
		Mean	Sd
1.3	20%	18.53	0.42
1.3	30%	19.72	0.48
1.3	50%	21.63	0.49
1.5	20%	20.75	0.42
1.5	30%	22.88	0.41
1.5	50%	24.88	0.23
1.8	20%	23.81	0.39
1.8	30%	25.30	0.27
1.8	50%	26.48	0.34
2	20%	24.39	0.40
2	30%	25.83	0.36
2	50%	26.66	0.37

FC: fold change of differentially expressed (DE) genes in the simulated dataset. DC (detection call): percentage of DE genes in the gene-affected sets. Representative R score: the mean R score of the top 60 gene sets in each simulation dataset. The mean R score was calculated by 10 datasets involving each configuration of FC and DC. Sd: standard deviation.

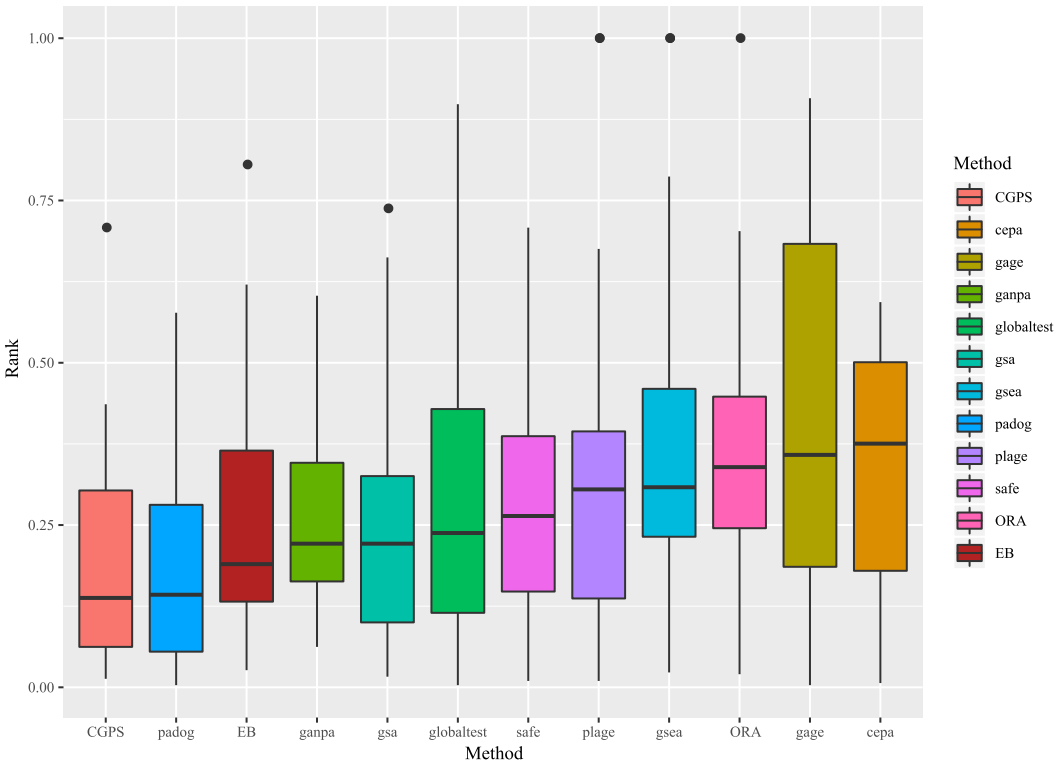


Fig. 4. Comparison of CGPS, EnrichmentBrowser, ORA and the nine individual GSE methods using the *KEGGdzPathwaysGEO* benchmark datasets. The boxes show the distributions of the ranks of the 24 target pathways received from the two combinatorial GSE methods (CGPS and EnrichmentBrowser (EB)) and 10 individual methods (including ORA and the 9 methods contained in CGPS). The lower values indicate that the method has a better prioritization ability. Methods are ranked from best to worst according to their median rank. The median rank of CGPS, padog, and EnrichmentBrowser is 0.138, 0.143, and 0.190, respectively.

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002

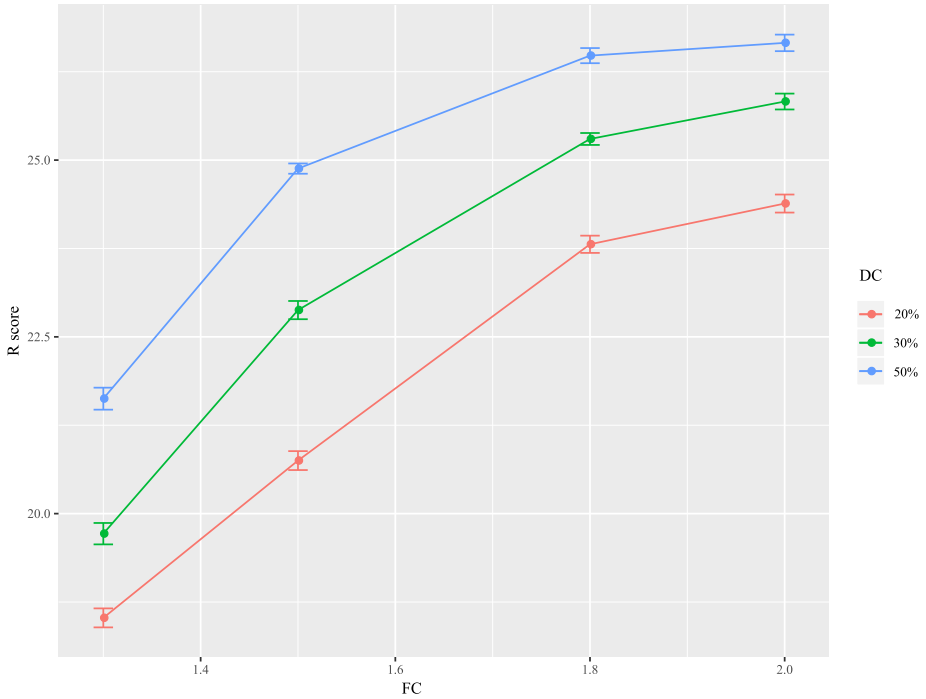


Fig. 5. Line chart of representative R scores under different configurations of FC and DC. FC: fold change of differentially expressed (DE) genes in the simulated dataset, FC = {1.3,1.5,1.8,2.0}. DC (detection call): percentage of DE genes in the gene-affected sets, DC = {20%, 30%, 50%}. Representative R score: the mean R score of the top 60 gene sets in each simulation dataset. For each configuration of FC and DC, 10 simulation datasets were used to calculate the representative R score. The x-axis represents the FC, and the color represents the different DC levels. The R score on the y-axis represents the value of the representative R score. Increasing either the FC or DC increases the R scores, indicating that the R scores are highly correlated with the expression level of the pathway.

changed with the FC and DC (Fig. 5). In Fig. 5, the FC value is labeled on the x-axis, and the different colors represent different DC values, with DC = 20% (red), DC = 30% (green), and DC = 50% (blue). As the value of the FC increased from 1.3 to 2.0, the representative R scores increased, and this increase was particularly dramatical from 1.3 to 1.8 as observed in each setting of DC. For example, at a DC of 30%, the R score increased from 19.72 to 25.83 as the value of the FC increased from 1.3 to 2.0. Moreover, as the value of the DC increased from 20% to 50% in each setting of the FC, the representative R score also increased. The coefficient of the linear regression is 0.90, showing an apparent positive relationship between the R score and the level of differential expression. In general, the R score can reflect the differential expression level of a gene set to some degree. A gene set with a higher differential expression level is assigned a higher R score. Thus, the value of the R score can reflect a pathway's relevance to an experimental condition; thus, the R score has the ability to prioritize relevant pathways.

2.7. Effects of the training datasets and statistical learning method on the R score's performance

The training datasets of CGPS consist of two classes of pathways: the pathways in the positive class were determined to be relevant to the experiment, and the pathways in the negative class were irrelevant and selected by being ranked last in all results provided by the nine methods. A pathway with a high R score is highly likely to belong to the positive class, i.e., to be a relevant pathway. Then, we addressed the following question: “is the prioritization ability due to the training datasets or the statistical learning method?” Thus, we built different training datasets and used another statistical learning model (logistic regression) to investigate the effects of the training datasets and the statistical learning method on the R score's ability to prioritize the relevant pathways and differentiate among the pathways based on their different degrees of relevance. We also used the drug-treated datasets and *KEGGdzPathwaysGEO* benchmark dataset to evaluate the performance.

2.7.1. Training datasets influence the prioritization ability of the R score

To investigate the influence of the training datasets, we first altered the pathways in the positive class or negative class of the training datasets of CGPS and then trained the model using the same methods as CGPS to compare the prioritization ability of the new model with that of the CGPS model based on the drug-treated datasets and *KEGGdzPathwaysGEO* benchmark datasets. The pathways in the positive class were changed to pathways that were less relevant to the experiment compared with those obtained by CGPS, and the pathways in the negative class were changed to pathways that were more relevant to the experiment such that they were more similar to the positive pathways. In this exercise, we built four groups of training datasets, i.e., pos25%, pos50%, neg50%, and neg75%. The training datasets of CGPS, pos25%, and pos50% included the same pathways in the negative class but represented the best to worst positive class (i.e., relevance of pathways from high to low in the positive class). The training datasets of CGPS, neg50%, and neg75% included the same pathways in the positive class but represented the best to worst negative class (in the negative class, the lower degree of relevance is better).

Then, we built four models based on these four groups of training datasets and applied these models to the drug-treated datasets and *KEGGdzPathwaysGEO* benchmark datasets. The ranks of the target pathways were calculated by sorted R scores. The

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002

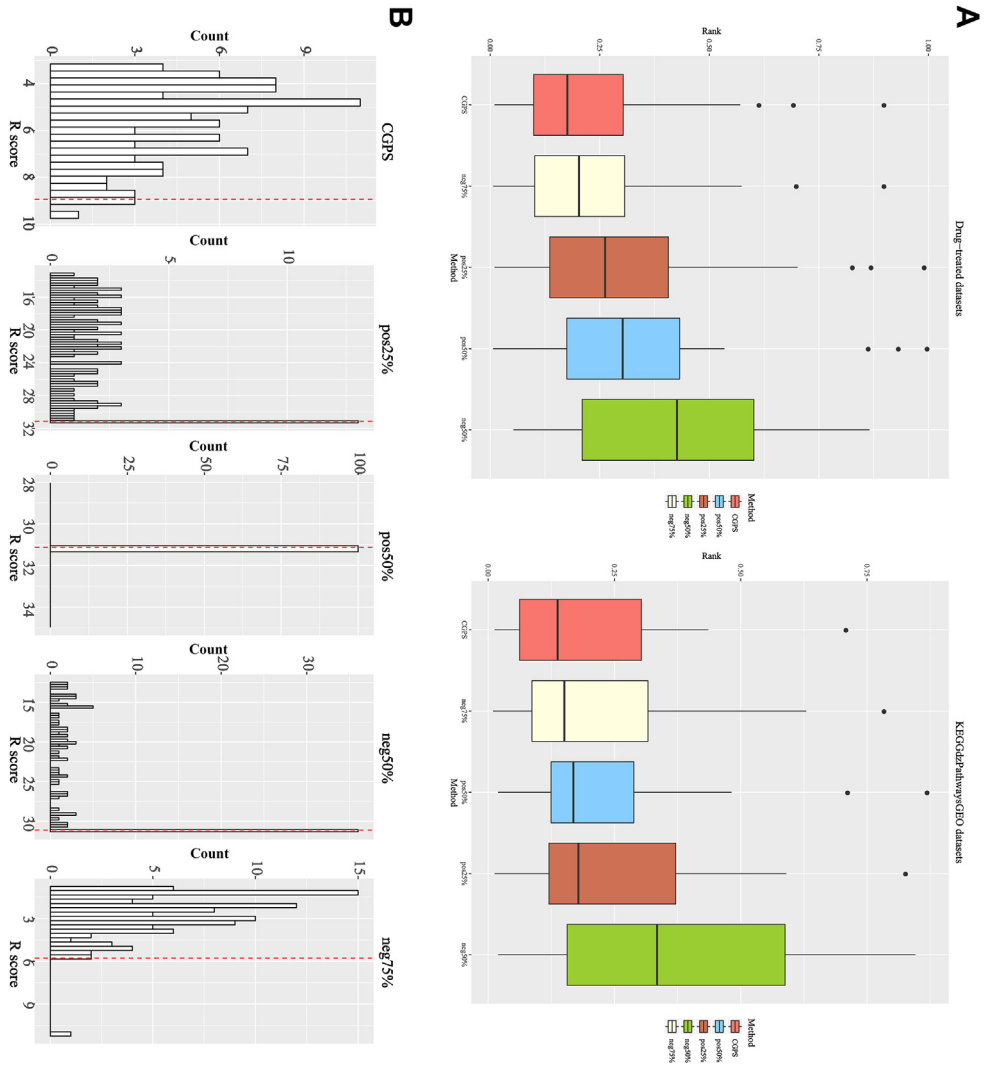


Fig. 6. Comparison of the R score's ability to prioritize target pathways and differentiate pathways based on their relevance to experimental conditions in five models. Five models were trained using different training datasets. CGPS, pos25%, pos50%, neg50%, and neg75% were trained based on different training datasets. Training datasets of the models CGPS, pos25%, and pos50% have the same negative class pathways but have best to worst positive class pathways (for the positive class, a higher degree of relevance is better for performance). Training datasets of the models CGPS, neg75%, and neg50% have the same positive class pathways but have best to worst negative class pathways (for negative class, a lower degree of relevance is better for performance). **A**: Distributions of the ranks of 51 and 24 target pathways using the drug-treated datasets (left) and *KEGGdzPathwaysGEO* benchmark datasets (right). The rank of a target pathway is calculated by ranking all pathways by the R score given by a model. The lower rank of a target pathway shows that the model can prioritize the target pathway better among all pathways. The models are ordered according to the median rank from best to worst. **B**: The distributions of the R scores of the top 100 pathways in the GSE1297 dataset (microarray data of patients with Alzheimer's disease and healthy controls). The count represents the number of pathways with the same range of R scores within the same bar. The R score of the target pathway of GSE1297 is labeled with a red line. The users need to differentiate Alzheimer's disease from 105 pathways with the same high R score using the pos50% model, 35 pathways using the neg50% model, and 12 pathways using the pos25% model, showing that these models' R scores have a poor ability to differentiate the relevant pathways. Compared with CGPS, the distribution of neg75% was more concentrated in the narrower range, making it difficult for users to differentiate the pathways in the same bar. CGPS outperforms the other four models in terms of the R score's ability to differentiate the relevant pathways.

performances of prioritizing the target pathways within each dataset are compared in Fig. 6A. In both the drug-treated datasets and *KEGGdzPathwaysGEO* benchmark datasets, CGPS had the lowest median of target pathway ranks in both datasets (drug-treated datasets: 0.176 and *KEGGdzPathwaysGEO* benchmark datasets: 0.138) and outperformed all four models built from the four groups of training datasets. Additionally, the worst model was built from the neg50% training dataset, with score of 0.426 for the drug-treated datasets and 0.334 for the *KEGGdzPathwaysGEO* datasets, which were clearly worse than the scores obtained with the other models. The model based on the neg75% training dataset was better than the other new models, with score of 0.203 for the drug-treated datasets and 0.151 for the *KEGGdzPathwaysGEO* datasets. These results showed that the relevance degree of the pathways in both the positive class and negative class influences the R score's ability to prioritize the relevant pathways.

2.7.2. Training datasets influence the R score's ability to differentiate relevant pathways

We also explored the differences among the results provided by the different models described above. The results generated by CGPS, pos50%, neg50%, neg25%, and pos25% using the drug-treated datasets and *KEGGdzPathwaysGEO* benchmark datasets are listed in Supplementary data. Because the feature of the R score distribution within a dataset is similar in all datasets, here, we only considered the GSE1297 dataset (microarray data of patients with Alzheimer's disease and healthy controls). The five models (including the CGPS model) generated very different R scores in the results of GSE1297.

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002

We plotted the distribution of the R scores of the top 100 pathways in the results (Fig. 6B). The three models built from the pos50%, neg50%, and pos25% training datasets showed a similar pattern, i.e., the models assigned the same highest R score to many different pathways. The model built from pos50% assigned all 105 pathways the same highest score (31.135) as shown in Fig. 6B. The model built from neg50% also assigned 35 pathways the highest R score of 31.135 (Fig. 6B); additionally, the model built from pos25% showed a similar pattern, and the 12 R scores were all the largest, as 31.135. The R score of the target pathway of GSE1297 (Alzheimer's disease) is labeled with a red line in Fig. 6B. Users could have difficulty in differentiating this pathway from the other less relevant pathways if they used the model built from pos50%, neg50%, or pos25% because many pathways share the same highest R score. However, this problem does not appear in the CGPS or neg75% models. In the results of CGPS, the pathways with high R scores are easily differentiated. Additionally, the phenomena can be observed in many other datasets in addition to GSE1297, such as GSE15471 and GSE18842. Among the four models with different training datasets, the influence of neg75% is less than that of neg50%, pos50%, or pos25%, although the distribution is less continuous than that of CGPS since we observed a gap between the highest and second highest R score in the results of neg75% (Fig. 6B). In general, the R score's ability to differentiate relevant pathways is also lower than that of CGPS when the relevance of the positive class is decreased or the relevance of the negative class is increased.

In summary, according to the observations presented above, both models trained with pos25% and pos50% performed much worse than CGPS, while neg50% performed worse and neg75% performed slightly worse than CGPS. This finding demonstrated that changing the pathways included in the positive class could have a greater influence on the R score's ability to prioritize and differentiate the relevant pathways than changing the pathways in the negative class.

2.7.3. Logistic regression model can be applied to the framework of CGPS

To investigate whether other types of statistical learning methods could also be applied to the framework of CGPS, we used a logistic regression as a model substitution of SVM. We used balanced datasets randomly selected from the training sets to train the logistic regression. Then, we evaluated the prioritization ability of the logistic regression and CGPS using the rank of the target pathways in the drug-treated datasets and *KEGGdzPathwaysGEO* datasets (Fig. S2 A and B). Both boxplots show that the prioritizing abilities of the SVM in CGPS and the logistic regression are almost the same. This finding demonstrates that the SVM can be substituted by a logistic regression in the CGPS framework, although the value of the R score could be changed. Nevertheless, we used the SVM as a model of CGPS to demonstrate this framework.

In addition, since the logistic regression model has the advantage of interpretation over SVM, we investigated the coefficients in the logistic regression to determine whether we can estimate the contribution of each method through the interpretation of the coefficients, such as the odds ratio. The odds ratio was calculated to estimate how much the odds increase if the *p*-value and rank decrease from 1.0 (worse) to 0.0 (better) (see Materials and methods section). A large odds ratio suggests that the method may have a large contribution to the classification. The coefficients and odds ratios are shown in Table S2. globaltest, gsea, padog and cepa show relatively high odds ratio (139, 134, 109 and 79), and the two methods with the smallest odds ratio are gsa and gage (0.43 and 1.81). These findings are coordinated with Fig. 3: padog, cepa, gsea and globaltest show relatively low rank values in our training set (lower is better), while gsa and gage show larger rank values. This interpretation of the odds ratio is helpful for understanding the system of CGPS; however, this interpretation is based on the logistic regression model, which may differ from the core of SVM.

2.8. Application of CGPS to the panobinostat (LBH-589) study

To demonstrate the functionality of CGPS using RNA-Seq data, we used a dataset from the RNA-Seq drug-treated datasets. This dataset is related to a drug called panobinostat (LBH-589), which is a histone deacetylase inhibitor approved for use in combination with bortezomib and dexamethasone in patients with relapsed or refractory multiple myeloma (an incurable malignancy of plasma cells) (Wahaib et al., 2016). This drug has been shown to function as a histone deacetylase inhibitor and can induce apoptosis via altered cell-cycle progression and/or cell differentiation (Bolden et al., 2006; Rasheed et al., 2008; Atadja, 2009). We used two samples from the human multiple myeloma cell line U266 without treatment and two samples from the same cell line with 10 nM panobinostat treatment for 4 h. The data are available from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) under accession GSE56623.

We applied the ensemble methods CGPS, EnrichmentBrowser, and EGSEA to this dataset. For EnrichmentBrowser and EGSEA, we used the average rank method to sort the final results. The reads were subjected to counts per million (CPM) normalization of the log form using Bioconductor edgeR (Robinson et al., 2010), and then, the limma pipeline (Law et al., 2014; Smyth, 2005) was used to perform the differential expression analysis.

Table 2 shows the top 10 pathways and target pathways retrieved from the 305 KEGG pathways. First, CGPS identified pathways associated with apoptosis and cell processes, including cell proliferation, growth, and differentiation, that are relevant to cancer. For example, the p53 signaling pathway *(hsa04115)* was ranked the 2nd by CGPS. It has been reported that the hyperacetylation of non-histone proteins, including p53, may play important roles in mediating the anti-tumor effects of histone deacetylase inhibitors (Buchwald et al., 2009). Furthermore, this pathway is an important cancer associated pathway. In contrast, the p53 signaling pathway was ranked the 23rd and 34th by EnrichmentBrowser and EGSEA, respectively. Another signaling pathway, i.e., the TGF-beta signaling pathway *(hsa04350)*, which is an important pathway related to apoptosis and cytostatic effects (Siegel and Massagué, 2003), is also affected by histone deacetylase inhibitors (Glenisson et al., 2007; Liu et al., 2013; Gaarenstroom and Hill, 2014). This pathway was ranked the 5th by CGPS, 47th by EnrichmentBrowser, and 134th by EGSEA.

Furthermore, the three target pathways of panobinostat recorded in the KEGG DRUG database, i.e., the Notch signaling pathway *(hsa04330)*, pathway in cancer *(hsa05200)*, and cell cycle *(hsa04110)*, were identified by CGPS (highlighted in bold in Table 2), and all ranked at the top position and were assigned a high R scores (R score >20.00, rank < 20 (top 6.60%)). Pathways in cancer *(hsa05200)* was ranked the 16th by CGPS, 41st by EnrichmentBrowser, and 58th by EGSEA.

Additionally, the following four pathways associated with cancer were identified by CGPS: viral carcinoma *(hsa05203)*, basal cell carcinoma *(hsa05217)*, transcriptional misregulation in cancer *(hsa05202)*, and nucleotide excision repair *(hsa03420)*, which were all ranked among the top 10 pathways, with R scores > 23.00. In contrast, EnrichmentBrowser and EGSEA missed *hsa05202* and *hsa03420* and ranked these pathways behind 24.

Interestingly, the pathway circadian rhythm *(hsa04710)* ranked first. It has been reported that histone deacetylases are conserved regulators of circadian function and that the knockdown of histone

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002

Table 2
The top 10 pathways and target pathways of panobinostat identified by CGPS from the panobinostat expression data.

Rank	Gene Set ID	Gene Set Name	R Score	EB	EGSEA
1	hsa04710	Circadian rhythm	27.96	7	6
2	hsa04115	p53 signaling pathway	27.65	23	34
3	**hsa04330**	**Notch signaling pathway**	26.55	2	12
4	*hsa05202*	*Transcriptional misregulation in cancer*	26.16	24	92
5	hsa04350	TGF-beta signaling pathway	25.63	47	134
6	*hsa05203*	*Viral carcinogenesis*	24.12	15	74
7	*hsa05217*	*Basal cell carcinoma*	24.11	42	4
8	hsa05130	Pathogenic Escherichia coli infection	24.09	12	75
9	hsa04728	Dopaminergic synapse	23.74	75	122
10	*hsa03420*	*Nucleotide excision repair*	23.47	30	83
16	***hsa05200***	***Pathways in cancer***	22.49	41	58
19	**hsa04110**	**Cell cycle**	20.92	16	18
24	hsa04210	Apoptosis	19.59	10	128.5

These pathways all have R scores > 6.000. The pathways associated with cell apoptosis, proliferation, growth, and differentiation are highlighted with underline.; the target pathways of panobinostat are highlighted in bold; the pathways associated with cancer are highlighted with italics. The ranks given by EnrichmentBrowser (EB) and EGSEA are also listed in the table.

deacetylases in *Drosophila* clock cells dampens circadian function (Fogg et al., 2014).

In summary, CGPS identified important cell processes associated with cancer, which could have been missed using EnrichmentBrowser and EGSEA, highlighting CGPS's usefulness for prioritizing important biologically relevant functions during GSE analysis.

2.9. Application on gene sets from the MSigDB collections using acute lymphoblastic leukemia microarray data

To further demonstrate that CGPS can be applied to gene set databases other than the KEGG pathway, we applies microarray data of acute lymphoblastic leukemia (Chiaretti et al., 2004). The 37 samples from the case group in this study were obtained from patients with the oncogenic fusion gene *BCR-ABL*, which is created by chromosomal translocation resulting in the joining of the *ABL1* gene on chromosome 9 and a part of the *BCR* gene on chromosome 22. The 39 control samples did not harbor the BCR-ABL gene fusion.

The data are available from Bioconductor in the ALL data package (http://bioconductor.org/packages/release/data/experiment/html/ALL.html). The two gene sets were both downloaded from the Molecular Signatures Database (MSigDB) v6.1, which was updated in October 2017, at http://software.broadinstitute.org/gsea/msigdb. These gene sets include: Gene Ontology biological process gene sets (C5, 5917 gene sets) and canonical pathways (C2, 1329 gene sets) collection. Canonical pathways collection is a comprehensive pathway collection containing gene sets from Reactome (Croft et al., 2011), Biocarta (https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways), KEGG, and PID (Schaefer et al., 2009).

We applied CGPS to this acute lymphoblastic leukemia data using the two gene set collections. The results obtained by CGPS using these gene sets are provided in Supplementary data (ftp://ftp.cbi.pku.edu.cn/pub/CGPS_download/dataResults/section2.9_ALL/). Here, we only list the first 10 gene sets in the two tables. We observed that most gene sets listed in Table 3, such as the Wnt signaling pathway, interferon signaling pathway, ATM pathway and Class I PI3K signaling pathway, were strongly correlated with T-cell acute lymphoblastic leukemia. Interestingly, the Wnt pathway and gene sets related to interferon (*REACTOME INTERFERON ALPHA BETA SIGNALING, GO RESPONSE TO TYPE I INTERFERON*) are ranked at the top in both the C2: canonical pathways collection and the GO biological processes gene set collection. The Wnt pathway is a highly conserved signal transduction pathway that governs cell fate decisions and is recruited by the vertebrate immune system to control early lymphopoiesis (Van de Wetering et al., 2002). Moreover, type I interferons constitute a heterogeneous group of cytokines with antiviral, antiproliferative, and immunomodulatory activities (Anguille et al., 2011). Treatment with alpha-interferon improves survival and remission duration in BCR-ABL$^+$ adult acute lymphoblastic leukemia patients (Visani et al., 2000).

These results show that CGPS could be applied to other gene set collections in addition to the KEGG pathway, such as GO gene sets, Reactome, and Biocarta. Additionally, CGPS provides a convenient way to use gene sets from different databases by inputting ".gmt" files, which are available from MSigDB and include more than 6700 annotated gene sets.

3. Discussion

Integrating multiple GSE methods into one ensemble method can outperform individual methods. However, the existing ensemble methods provide too many types of ensemble scores for ranking gene sets as an integrated result. Thus, choosing the most reliable score for optimal and comprehensively integrated results is challenging. In this study, we have developed a new approach named CGPS that integrates the properties of nine GSE tools into one ensemble score (R score), which improves the prioritization of relevant biological functions in GSE analyses. We have compared CGPS with EnrichmentBrowser and four types of ensemble scores from EGSEA and showed that CGPS's R score performs best in prioritizing the target pathways in the RNA-Seq drug-treated datasets. Additionally, the R score provided by CGPS is among the best two ensemble scores in terms of the FDR according to an evaluation of the simulated RNA-Seq datasets, and the low FDR of the R score indicates that the gene sets correlated with the experimental condition could be determined based on a high R score, allowing the relevant pathways to be prioritized among all gene sets and easily identified. CGPS outperforms the individual methods in prioritizing the target pathways in the drug-treated datasets, demonstrating that the approach using CGPS to integrate these individual methods is efficient. In the *KEGGdzPathwaysGEO* benchmark datasets with 24 target pathways, CGPS outperforms EnrichmentBrowser and 10 individual methods, indicating that the R score's prioritization ability is consistently good, even with different types of expression data. In addition to the prioritization ability of the R score, we show that the level of the R score reflects the differential expression level of a gene set. Finally, we have applied CGPS to a drug-treated cell line experiment (panobinostat). CGPS identifies biological functions related to apoptosis and cell proliferation, such as the p53 and TGF-beta signaling pathways, neither of which are ranked near the top by

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002

Table 3
Top 10 gene sets identified by CGPS from the C2 (canonical pathways) and GO biological processes gene sets.

Collection	Rank	Gene set	Brief description	Reference
MSigDB C2 gene set	1	BioCarta hivnef pathway	HIV-I Nef: negative effector of Fas and TNF	Bernhard et al., 2001
	2	PID wnt noncanonical pathway	Noncanonical Wnt signaling pathway	Van de Wetering et al., 2002
	3	Reactome interferon alpha beta signaling	Genes involved in Interferon alpha/beta signaling	Visani et al., 2000
	4	PID toll endogenous pathway	Endogenous TLR signaling	Chiron et al., 2008
	5	PID hiv nef pathway	HIV-1 Nef: Negative effector of Fas and TNF-alpha	Bernhard et al., 2001
	6	Reactome pecam1 interactions	Genes involved in PECAM1 interactions	Akers et al., 2010
	7	BioCarta atm pathway	ATM Signaling Pathway	Gumy-Pause et al., 2004
	8	PID pi3kci pathway	Class I PI3K signaling events	Fumarola et al., 2014
	9	Reactome nephrin interactions	Genes involved in Nephrin interactions	Takahashi et al., 2017
	10	PID netrin pathway	Netrin-mediated signaling events	Ranganathan et al., 2014
GO (Gene Ontology)	1	Actin filament bundle organization	-	Desouza et al., 2012
	2	Wnt signaling pathway calcium modulating pathway	-	Van de Wetering et al., 2002
	3	Endothelial cell development	-	Pitt et al., 2015
	4	Negative regulation of i kappab kinase nf kappab signaling	-	Staal and Langerak, 2008
	5	Response to type i interferon	-	Visani et al., 2000
	6	Regulation of cysteine type endopeptidase activity	-	Patel et al., 2009
	7	Glomerulus development	-	Luciano and Brewster, 2014; Yetgin et al., 2004
	8	Establishment of endothelial barrier	-	Livrea et al., 1985
	9	Regulation of monooxygenase activity	-	Mayerhofer et al., 2004
	10	Regulation of i kappab kinase nf kappab signaling	-	Staal and Langerak, 2008

The C2 gene set collection consists of 1329 gene sets from several databases, including REACTOME, PID, BioCarta, and KEGG pathway. The GO biological processes gene set collection consists of 4436 gene sets.

EnrichmentBrowser or EGSEA. These lines of evidence demonstrate that the R score provided by CGPS based on the SVM has the capacity to better prioritize the relevant pathways and helps users identify important pathways that may be missed using other tools.

CGPS can be widely used with many types of annotated gene set databases and expression data. First, CGPS can be applied without limitation to annotated gene set databases. As we have demonstrated in the case of acute lymphoblastic leukemia, the gene sets related to the Wnt signaling pathway are recovered from two different gene set databases (C2 and GO_BP) by CGPS. In practice, CGPS also provides a convenient way to apply any type of annotated gene set databases by allowing to input ".gmt" format files. Second, CGPS can be applied without limitation to expression data. Although CGPS is trained on drug-treated datasets, it can be applied to other datasets, such as the *KEGGdzPathwaysGEO* and simulation datasets. These two attributes of CGPS are logical because the SVM of CGPS learns the *p*-value and rank patterns of "relevant" pathways (the positive class of pathways in the training dataset) and "irrelevant" pathways (the negative class of pathways in the training dataset) produced by the nine individual methods. Because CGPS can integrate results obtained from nine individual methods and the nine methods can be widely applied to different gene set databases and different types of datasets, CGPS can also be applied to these gene set databases and datasets.

The training datasets play an important role in the R score's ability to prioritize the relevant pathways and differentiate pathways based on the different degrees of their relevance to the experiment. The training datasets consisted of two classes of pathways, which are selected based on a priori knowledge of the pathways' relevance to the experimental condition, and the two classes act as two poles. The positive class represents pathways relevant to the data, and the negative class represents irrelevant pathways. The R score represents the probability of a positive class, which is used to measure the degree of relevance of the pathways according to what the model has learned from the two poles of the training dataset (a large value indicates a high degree of relevance). Thus, to measure pathways with a broader range of degree of relevance, we could increase the degree of relevance of the pathways in the positive class of the training datasets and decrease the degree of relevance of the pathways in the negative class in building the training datasets.

We have investigated the impact of the relevance of the pathways in the two classes of the training datasets on the R score's ability to both prioritize and differentiate pathways by using different training datasets. First, we have demonstrated that the R score's ability to prioritize can be decrease by either decreasing the relevance of the pathways in the positive class of pathways or increasing the relevance of the pathways in the negative class of pathways. As shown in Fig. 6, based on 51 target pathways in the drug-treated dataset and 24 target pathways in the *KEGGdzPathwaysGEO benchmark datasets*, CGPS outperforms all models built from the new training datasets in prioritizing the target pathways. Second, we have demonstrated that the R score's ability to differentiate the relevant pathways can be decreased by decreasing the relevance of the positive class of pathways in the training dataset since we observe that many pathways shared the same highest R score in the results of the pos50% and pos25% models (Fig. 6B), likely because the models had not learned the features of the pathways with a higher degree of relevance than those in the positive class of pathways. Thus, the models tends to assign these highly relevant pathways a capped probability of being in the positive class. So we recommend that the positive class of pathways in the training dataset should consist of pathways that are highly relevant to the data. A clear and direct causal relationship is better than an unclear correlation between target pathways and experimental conditions. For example, the drug's target pathway and samples only treated by the drug could be used. In contrast, regarding the effect of the negative class, the R score's ability to differentiate the relevant pathways can also be decreased by increasing the relevance of the negative class of pathways in the training set. As demonstrated in Fig. 6B, many pathways shared the same highest R score in the results based on neg50%. We have observed that the influence of neg75% is less than that of neg50%, pos50%, or pos25% and have speculated that the pathways actually

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002

relevant to the experimental condition are in a minority in GSE1297. In general, we recommend that the pathways in the negative class consist of the pathways with the poorest ranks among all the nine methods, i.e., the pathway with the largest average rank of the nine methods. In summary, the training datasets should use the most relevant pathways in the positive class, and the pathways with the poorest ranks among all the nine methods should comprise the negative class.

In this study, we have provided a framework for integrating GSE methods based on the relationship between the pathways and phenotype. We also have discussed the general rules for selecting the training datasets, i.e., a positive class with more relevant pathways is better, and the negative class should comprise the most irrelevant pathways. We also demonstrate that the model can be substituted with a logistic regression, and researchers could also try to apply other statistical learning models to the framework to seek better performance. Importantly, the collected datasets are not massive, although we observe the power of this method. Furthermore, we propose the following two indicators for the evaluation of the R score: the prioritizing ability and differentiating ability using benchmark datasets. However, the benchmark datasets need to be enlarged. In the future, we intend to implement additional types of models for optimization, enlarge the training datasets for more accurate predictions, and enlarge the benchmark datasets for a more accurate evaluation. Moreover, we aim to add more user-friendly modules for the visualization of the results.

Combining multiple GSE methods can outperform individual methods, which has been demonstrated by methods, such as EnrichmentBrowser and EGSEA. The combination strategies differ among EnrichmentBrowser, EGSEA, and CGPS. EnrichmentBrowser applies naïve statistics to ranks, and EGSEA uses a more complex statistical ensemble method to integrate the *p*-value in the results. CGPS uses a different strategy. It includes a priori knowledge of the relationship between the phenotype and pathways into the integrating method. Thus, CGPS not only is a statistical ensemble model but also is a model that learns based on biological knowledge. We propose that some information could not be captured by pure statistical methods, and by learning from the relationship between known target pathways and treated samples. We demonstrated that CGPS has the power to prioritize relevant pathways and outperforms EnrichmentBrowser and EGSEA in tests involving both real and simulated data.

To the best of our knowledge, this study represents the first application of the relationship between pathways and phenotypes as a priori knowledge to integrate gene set enrichment methods. We expect users will benefit from this approach by discovering important biologically relevant functions that may be missed using other GSE methods.

4. Materials and methods

4.1. Drug-treated dataset collection

We aimed to collect gene expression data with known target pathways relevant to given experimental conditions. Thus, we collected gene expression data from 105 experiments related to drug-treated cell lines from human or mouse. Among the 105 experiments, there were 37 drugs in total, which were all under FDA approval. Each drug in a given experiment had one or several target pathways recorded in the KEGG DRUG database (http://www.genome.jp/kegg/drug/) that were considered relevant pathways to each experiment, and the target pathways were all in the KEGG pathway collection. Among the 105 datasets, 52 datasets were generated by RNA-Seq, and the other datasets were based on microarray technology. Across all datasets, 255 target pathways were obtained. All gene expression data are accessible in ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) (Parkinson et al., 2007) and GEO (https://www.ncbi.nlm.nih.gov/geo/) (Edgar et al., 2002). Information regarding the 105 datasets and their target pathways can be downloaded from ftp://ftp.cbi.pku.edu.cn/pub/CGPS_download/. The KEGG pathway collection was downloaded on August 2016. The total number of pathways was 305 in human and 301 in mouse.

4.2. Processing of gene expression data in CGPS

There are two types of gene expression data in the 105 datasets, i.e., RNA-Seq data and microarray data. We treated these data differently in the normalization process. The microarray data were normalized using the quantile method in the limma package (Smyth, 2005). For the RNA-Seq read-count data, because it has been shown that log CPM normalized for sequencing in-depth data, such as Reads Per Kilobase per Million mapped reads (RPKM), could be input into a well-established microarray analysis pipeline, such as the limma package (Law et al., 2014), the log CPM method was applied using the edgeR package (Robinson et al., 2010). If the data had been normalized using the RPKM method, we applied a log function to the data. The process was compiled in the script of CGPS so that CGPS could consider the gene expression data input.

4.3. Running the individual methods and extracting the results

We applied seven set-based methods (gsea, gsa, padog, plage, gage, globaltest, and safe) and two network-based methods (cepa and ganpa) to the 105 drug-treated datasets. The R package EnrichmentBrowser was used to run these methods, except for plage and gage. The plage method is available in the GSVA package (Hänzelmann et al., 2013), and the gage method is available in the gage package. For the network-based methods, the gene regulatory networks were compiled using the KEGGgraph package (Zhang and Wiemann, 2009) based on all KEGG pathways of a given species. For a given dataset, we extracted the target pathways' rank and *p*-value from the results of these methods. The rank was an absolute value, which was then divided by the total number of pathways to generate a decimal value ranging from 0 to 1. Thus, we obtained the performance (rank/*p*-value) of all relevant pathways for further processing.

4.4. Preparing data for training and testing the SVM

To train a model to identify the pathways relevant to the experimental conditions among all pathways using a supervised learning method, we needed two classes of pathways: one class contains the relevant pathways (positive class), and the other class contains pathways that are not relevant to the experimental condition (negative class). The 255 known drug-targeted pathways involved in the 105 drug-treated datasets were used as positive class pathways (labeled as +1 class). To identify the negative class pathways, we made two assumptions. First, the pathways probably not relevant to the experimental conditions (i.e., the negative class of pathways) should show little differential expression between the case and control samples; thus, most genes in these pathways show little difference in expression in comparisons of case and control samples. Second, the negative class of pathways should be consistently ranked low by multiple GSE methods. Thus, we used two methods to identify the negative class pathways. Based on the first assumption, genes with high *p*-values (*p*-value >0.90) in the differentially expressed gene analysis of each expression dataset were selected, and we then applied the ORA method to these genes. All pathways in the results of the ORA analysis were selected. Based

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002

on the second assumption, we calculated the average rank of each pathway provided by the nine methods and sorted the pathways based on their average rank to prioritize the pathways that were consistently ranked low by the nine methods. Finally, we filtered the pathways that were not in the ORA results and obtained 10 pathways with the largest average rank value (10 most lagging pathways according to the nine methods) based on the second assumption. We also checked all pathways to ensure that the pathways in the positive and negative classes did not intersect. The workflow of this process is presented in Fig S3. In total, across all 105 datasets, we identified 1038 "irrelevant" pathways for training the model. We also extracted the rank and *p*-value of these pathways following the methods described above.

4.5. Design of the SVM's feature vector and partition of the datasets

In total, we extracted 1293 positive and negative class pathways as described above. These pathways were assigned gene set rankings and *p*-values by applying the nine GSE methods. These data were shaped as a 1293 × 18 matrix for training and testing the SVM model. First, the 1293 pathways were divided into two parts: one part was used to train the SVM model, and the other part was used to test the performance of the SVM model. Because several pathways were involved in the same datasets and to ensure a maximum of independence between the training and testing sets, we divided the pathways based on the datasets in which they were involved rather than the pathways themselves. In total, 1035 pathways, including 204 target pathways, from 84 (80%) randomly selected datasets were used to train the model. The training method is described below. The other 258 pathways, including 51 target pathways, from 21 datasets were used to test the model. Additionally, we used 51 target pathways in the testing datasets to compare CGPS with the other methods in terms of prioritization ability. In these 21 datasets, 11 RNA-Seq datasets with 29 target pathways designated RNA-Seq drug-treated datasets were used to evaluate CGPS, EnrichmentBrowser, and EGSEA. Notably, we also investigated the Kendal correlation between pairs of the 18 features (Fig. S4), demonstrating that most pairs of features had a correlation as low as 0.1, and only 5 pairs had correlations greater than 0.6 (153 pairs of features). Based on the performance presented in the Results section, we deduced that multicollinearity has little influence on performance.

4.6. Construction of the SVM classifier

Based on the 1035 pathways used for training, an SVM classifier was trained with a linear kernel function, implemented using scikit-learn (http://scikit-learn.org/) and confirmed via five-fold cross-validation. The SVM classifier with the best cross-validation accuracy was used as the final classifier. The parameters of the SVM classifier, including the kernel type ("linear", "poly", "rbf", and "sigmoid"), penalty parameter C of the error term (10^{-3} - 10^{3}), and kernel coefficient gamma (10^{-4},10^{-3}, 10^{-2}), were chosen using a grid search to maximize the accuracy of the classification in the cross-validation process.

The performance of this model in classifying the positive and negative class pathways indicated an accuracy = 98.06%, recall = 94.12%, precision = 96.00%, and specificity = 99.03%. The R score is considered as a variable with a positive relationship to the probability of the outcome class as follows: R score = -lg (1-p), where p is the probability of belonging to the positive class. Additionally, we output the distance D of each pathway to the separating hyperplane. The final results were sorted by the R score in descending order. If two pathways had the same R score, they were sorted by the distance D in descending order only for display as we propose that two pathways with same probability of belonging to the positive class have almost the same relevance.

4.7. General workflow of CGPS

The following three steps are performed in running CGPS: processing to acquire the input for the SVM, running the SVM, and outputting the results.

4.7.1. Processing to acquire the input for the SVM

This stage considered gene expression as the input and processes the expression data following the methods described above. The nine methods were then applied, and the *p*-value and rank of each pathway given by each method were extracted as described above. Finally, we generated a table of the pathways' results, and each row represented a pathway's result, while each column contained the *p*-value or rank of a given method (Fig. 1).

4.7.2. Running the SVM

The table obtained during the first stage is entered into the SVM, and the results, including the R score of each pathway, were generated.

4.7.3. Outputting the results

Each pathway was assigned an R score. A higher R score indicates that the pathway is more relevant to the experiment condition. All pathways are ranked by the R score in descending order such that the relevant pathways are placed near the top position in the resulting table.

Notably, although CGPS is an SVM classifier, regarding the classifier's result, we focus on the order of the pathways sorted by the R score instead of the class of each pathway because pathways with a low R score are not accurately defined as relevant or irrelevant. Additionally, the cut-off suitable for all cases is not easy to define. However, the R score generated by the SVM is still useful for prioritizing the relevant pathways.

4.8. Generating simulated RNA-Seq datasets

4.8.1. Simulation setups

The simulated methods generally followed the procedures described by Rahmatallah et al. (2016). In brief, modeling the count for gene i in sample j by a random variable Y_{ij} with a negative binomial (NB) distribution

$$Y_{ij} \sim NB\left(mean = \mu_{ij},\ var = \mu_{ij}\left(1 + \mu_{ij}\varphi_{ij}\right)\right) = NB\left(\mu_{ij}, \varphi_{ij}\right)$$

where μ_{ij} and φ_{ij} are the mean count and dispersion parameter of gene i in sample j, respectively. For each gene, we randomly selected the vectors for the mean count, dispersion, and gene length from a pool of vectors derived from the real RNA-Seq dataset. Here, we used RNA-Seq data from The Cancer Genome Atlas (The Cancer Genome Atlas Research Network et al., 2013) for Urothelial Bladder Carcinoma (TCGA-BLCA) and selected 36 samples from these data (19 cases *vs.* 17 controls). We then filtered the genes with mean read counts lower than 1 and higher than 1000 to avoid extreme read count values. In total, we obtained 11,638 genes to use as a gene pool. For each gene in this pool, the mean count, dispersion, and gene length were calculated. The dispersion parameters of the individual genes were estimated using the Bioconductor edgeR package.

In each simulation, the gene total was 10,000, and 200 non-overlapping gene sets containing 50 genes per set were generated from these 10,000 genes. These 10,000 genes were randomly

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002

selected from the gene pool 10 times, and each time, the obtained 10,000 genes were used to generate 200 gene sets. These 10 batches of gene sets were then used in the subsequent steps to generate read counts in each configuration (Fig. S5).

The expression data were generated using 20 control samples and 20 case samples. To mimic the differential gene expression between the case and control samples, the case samples were induced with parameters that represented the differential gene expression level. For example, given the mean count μ_i of gene i, the mean count μ_i in the cases are several times (gene fold change) higher than the mean count in the controls for some genes, so that the expression values of some genes were increased only by a particular amount. The genes that had an increased expression in the cases were called DE genes. In a given gene set, we used DC to describe the percentage of DE genes within a gene set. The DC was proposed in an assessment method to identify differentially expressed pathways (Tripathi and Emmert-Streib, 2012).

4.8.2. Variable parameters and processing of read count generation

As described above, we used two variable parameters to describe the expression level of the gene set. These variable parameters are FC and DC. The four levels of FC, i.e., FC = {1.3, 1.5, 1.8, 2.0}, reflect expression changes in DE genes between the case and control groups. The three levels of DC were {20%, 30%, 50%}. These values resulted in 12 distinct configurations of the parameters. Each parameter configuration was examined in 10 runs to obtain stable values.

The gene sets were classified into the following three classes: the Class 1 gene set (C1) was defined as the 60 gene sets that contained as many as {10, 15, 25} DE genes according to the DC configuration; the Class 2 gene set (C2) was defined as the 70 gene sets that contained only two DE genes (4%) to mimic random noise; and the Class 3 gene set (C3) was defined as the 70 gene sets that contained zero DE genes. In the simulation analysis, the C1 gene sets were up-regulated; thus, we used the C1 gene sets as the truly affected gene sets.

There are four steps to generate the datasets of each configuration. The process of the read count generation is shown in Fig. S5.

4.9. Generating the four groups of training datasets

First, we used all pathways in the ORA results based on the genes with p-values >0.90 in the differential expression analysis within each dataset, which is the same process. We then calculated the sum of the ranks of each pathway provided by the nine methods. Within each dataset, we ranked all pathways in ascending order with the relevant pathways ranked near the top of this list (L pathways). For this list, we set the following three points: the first quartile (Q1), the second quartile (Q2), and the third quartile (Q3). At each point, $p = L \times k/4$, $k = 1, 2, 3$; we selected 10 pathways in this range [p-5, p+5] within each dataset. To generate the new training sets based on the training sets of CGPS, we exchanged either the positive class pathways or negative class pathways with selected pathways near the three quartile points. In practice, we generated the following four groups of training sets: pos25%, neg75%, pos50%, and neg50%. For example, pos25% has positive class pathways represented by 10 pathways within each dataset near Q1 selected by the method described above and the negative class pathways were the same as those in the training set of CGPS; neg75% has negative class pathways represented by pathways near Q3, and the positive class pathways are the same as those in the training set of CGPS.

"Pos" or "neg" represents the variable class in the training dataset compared with the training dataset of CGPS. The number "25%" represents how a pathway is ranked among all pathways. For example, the model "pos25%" was trained using a dataset consisting of 10 pathways ranked at Q1 (1st quartile) and the same negative class as that used in the training dataset.

We then applied the four models built from the four groups of datasets to the drug-treated datasets and the *KEGGdzPathwaysGEO* datasets. The models were trained using the same methods as CGPS.

4.10. Logistic regression classifier

We used the same training datasets as CGPS but first balanced the samples in the two classes as 200: 200 by subsampling the pathways in the negative class. The parameters were as follows: Cs = [10^{-3}, 10^{-2}, 10^{-1},10,10^2,10^3], penalty = "l1", and solver = "saga". The model was trained using five-fold cross validation. We used "LogisticRegressionCV" in the sci-kit learn python package.

We investigated the coefficients of the logistic regression and calculated the odds ratio of each method (Table S2). The odds ratio is calculated as follows: the other features are fixed, and we divide the odds when given p-value = 0.0 and rank = 0.0 by the odds when p-value = 1.0 and rank = 1.0, see the formula.

$$\text{Odds ratio} = \frac{odds\ when\ p_1 = 0,\quad r_1 = 0}{odds\ when\ p_1 = 1.0,\quad r_1 = 1.0} = \frac{e^{w_0 + w_1 \times 0.0 + w_2 \times 0.0}}{e^{w_0 + w_1 \times 1.0 + w_2 \times 1.0}} = e^{-(w_1 + w_2)}$$

where w_1 and w_2 are the coefficients of p-value p_1 and rank r_1, and w_0 represents the factor related to the other fixed features.

If the odds ratio is greater than 1.0, the small p-value and rank makes it more likely to be classified as a positive class pathway (enriched pathway). For example, to interpret the contribution of cepa, we calculated the odds ratio of cepa. If the p-value of cepa is p_1 and the rank of cepa is r_1, the coefficients of p_1 and r_1 are w_1 and w_2. We fixed the p-value and rank value of the other methods. If the odds ratio of cepa is 71, the odds ratio of enrichment increase 70 times from $p_1 = 1.0$ and $r_1 = 1.0$ to $p_1 = 0$ and $r_1 = 0$ if the features of the other methods are fixed.

4.11. Using EnrichmentBrowser and EGSEA

EnrichmentBrowser integrates seven methods, including gsea, gsa, padog, plage, globaltest, and safe. EGSEA integrates eight methods, including safe, gage, padog, globaltest, camera, zscore, ssgsea, and ORA. For EnrichmentBrowser, we used the average rank to sort all pathways in ascending order. For EGSEA, we used the following four methods to sort the pathways: average rank, vote rank, combined p-value, and significance score. When using the first three methods in EGSEA, we sorted the pathways in ascending order, but to obtain the significance score, we sorted the pathways in descending order because pathways with a high significance score are assumed to be more relevant to the experimental condition.

4.12. Implementation

We provided a pipeline in python for running CGPS. We relied on the R packages limma, edgeR, gage, and EnrichmentBrowser. All scripts are available at ftp://ftp.cbi.pku.edu.cn/pub/CGPS_download/cgps-1.0.0.tar.gz.

Acknowledgments

We thank Prof. Ge Gao and Prof. Liping Wei for critical feedback regarding this work, Lan Ke for giving suggestions on the model

training, Jingyi Li for feedback regarding the model training, De-Chang Yang for testing the code, Yang Ding for assistance with testing the R scripts, Adam Yongxin Ye for suggestions regarding the simulation design, August Yue Huang and Xiaoxu Yang for feedback regarding the comparison analysis, and Juan Wu for helping with the collection of information regarding the datasets. This work was supported by the National Key Research and Development Program of China (2017YFC1201200, 2017YFC0908404, 2016YFC0901603, 2016YFB0201700), National High-tech R&D Program of China (863 Program) (2015AA020108), and the State Key Laboratory of Protein and Plant Gene Research.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jgg.2018.08.002.

References

Akers, S.M., O'Leary, H.A., Minnear, F.L., Craig, M.D., Vos, J.A., Coad, J.E., Gibson, L.F., 2010. VE-cadherin and PECAM-1 enhance ALL migration across brain microvascular endothelial cell monolayers. Exp. Hematol. 38, 733–743.

Alhamdoosh, M., Ng, M., Wilson, N.J., Sheridan, J.M., Huynh, H., Wilson, M.J., Ritchie, M.E., 2017. Combining multiple tools outperforms individual methods in gene set enrichment analyses. Bioinformatics 33, 414–424.

Anguille, S., Lion, E., Willemen, Y., Tendeloo, V.F.I.V., Berneman, Z.N., Smits, E.L.J.M., 2011. Interferon-α in acute myeloid leukemia: an old drug revisited. Leukemia 25, 739.

Atadja, P., 2009. Development of the pan-DAC inhibitor panobinostat (LBH589): successes and challenges. Canc. Lett. 280, 233–241.

Barry, W.T., Nobel, A.B., Wright, F.A., 2005. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinforma. Oxf. Engl. 21, 1943–1949.

Bayerlová, M., Jung, K., Kramer, F., Klemm, F., Bleckmann, A., Beißbarth, T., 2015. Comparative study on gene set and pathway topology-based enrichment methods. BMC Bioinformatics 16, 334.

Bernhard, D., Skvortsov, S., Tinhofer, I., Hübl, H., Greil, R., Csordas, A., Kofler, R., 2001. Inhibition of histone deacetylase activity enhances Fas receptor-mediated apoptosis in leukemic lymphoblasts. Cell Death Differ. 8, 1014.

Bolden, J.E., Peart, M.J., Johnstone, R.W., 2006. Anticancer activities of histone deacetylase inhibitors. Nat. Rev. Drug Discov. 5, 769–784.

Buchwald, M., Krämer, O.H., Heinzel, T., 2009. HDACi–targets beyond chromatin. Cancer Lett. 280, 160–167.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., Foa, R., 2004. Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. Blood 103, 2771–2778.

Chiron, D., Bekeredjian-Ding, I., Pellat-Deceunynck, C., Bataille, R., Jego, G., 2008. Toll-like receptors: lessons to learn from normal and malignant human B cells. Blood 112, 2205–2213.

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., Stein, L., 2011. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 39, D691–D697.

Desouza, M., Gunning, P.W., Stehn, J.R., 2012. The actin cytoskeleton as a sensor and mediator of apoptosis. BioArchitecture 2, 75–87.

Dong, X., Hao, Y., Wang, X., Tian, W., 2016. LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. Sci. Rep. 6, 18871.

Edgar, R., Domrachev, M., Lash, A.E., 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 207–210.

Efron, B., Tibshirani, R., 2007. On testing the significance of sets of genes. Ann. Appl. Stat. 1, 107–129.

Fang, R., Xiao, T., Fang, Z., Sun, Y., Li, F., Gao, Y., Feng, Y., Li, L., Wang, Y., Liu, X., Chen, H., Liu, X.-Y., Ji, H., 2012a. MicroRNA-143 (miR-143) regulates cancer glycolysis via targeting hexokinase 2 gene. J. Biol. Chem. 287, 23227–23235.

Fang, Z., Tian, W., Ji, H., 2012b. A network-based gene-weighting approach for pathway analysis. Cell Res. 22, 565–580.

Fogg, P.C.M., O'Neill, J.S., Dobrzycki, T., Calvert, S., Lord, E.C., McIntosh, R.L.L., Elliott, C.J.H., Sweeney, S.T., Hastings, M.H., Chawla, S., 2014. Class IIa histone deacetylases are conserved regulators of circadian function. J. Biol. Chem. 289, 34341–34348.

Fumarola, C., Bonelli, M.A., Petronini, P.G., Alfieri, R.R., 2014. Targeting PI3K/AKT/mTOR pathway in non small cell lung cancer. Biochem. Pharmacol. 90, 197–207.

Gaarenstroom, T., Hill, C.S., 2014. TGF-β signaling to chromatin: how Smads regulate transcription during self-renewal and differentiation. Semin. Cell Dev. Biol. 32, 107–118.

Geistlinger, L., Csaba, G., Zimmer, R., 2016. Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- & network-based enrichment analysis. BMC Bioinf. 17, 45.

Glenisson, W., Castronovo, V., Waltregny, D., 2007. Histone deacetylase 4 is required for TGFβ1-induced myofibroblastic differentiation. Biochim. Biophys. Acta BBA - Mol. Cell Res. 1773, 1572–1582.

Goeman, J.J., Bühlmann, P., 2007. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinforma. Oxf. Engl. 23, 980–987.

Goeman, J.J., van de Geer, S.A., de Kort, F., van Houwelingen, H.C., 2004. A global test for groups of genes: testing association with a clinical outcome. Bioinforma. Oxf. Engl. 20, 93–99.

Gu, Z., Wang, J., 2013. CePa: an R package for finding significant pathways weighted by multiple network centralities. Bioinforma. Oxf. Engl. 29, 658–660.

Gumy-Pause, F., Wacker, P., Sappino, A.-P., 2004. *ATM* gene and lymphoid malignancies. Leukemia 18, 238.

Hänzelmann, S., Castelo, R., Guinney, J., 2013. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinf. 14, 7.

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M., 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38, D355–D360.

Khatri, P., Sirota, M., Butte, A.J., 2012. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. 8, e1002375.

Law, C.W., Chen, Y., Shi, W., Smyth, G.K., 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 15, R29.

Liu, N., He, S., Ma, L., Ponnusamy, M., Tang, J., Tolbert, E., Bayliss, G., Zhao, T.C., Yan, H., Zhuang, S., 2013. Blocking the class I histone deacetylase ameliorates renal fibrosis and inhibits renal fibroblast activation via modulating TGF-beta and EGFR signaling. PLoS One 8, e54001.

Livrea, P., Trojano, M., Simone, I.L., Zimatore, G.B., Logroscino, G.C., Pisicchio, L., Lojacono, G., Colella, R., Ceci, A., 1985. Acute changes in blood-CSF barrier permselectivity to serum proteins after intrathecal methotrexate and CNS irradiation. J. Neurol. 231, 336–339.

Luciano, R.L., Brewster, U.C., 2014. Kidney involvement in leukemia and lymphoma. Adv. Chron. Kidney Dis. 21, 27–35.

Luo, W., Friedman, M.S., Shedden, K., Hankenson, K.D., Woolf, P.J., 2009. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinf. 10, 161.

Mayerhofer, M., Florian, S., Krauth, M.-T., Aichberger, K.J., Bilban, M., Marculescu, R., Printz, D., Fritsch, G., Wagner, O., Selzer, E., Sperr, W.R., Valent, P., Sillaber, C., 2004. Identification of heme oxygenase-1 as a novel BCR/ABL-dependent survival factor in chronic myeloid leukemia. Cancer Res. 64, 3148–3154.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., Brazma, A., 2007. ArrayExpress–a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 35, D747–D750.

Patel, N., Krishnan, S., Offman, M.N., Krol, M., Moss, C.X., Leighton, C., van Delft, F.W., Holland, M., Liu, J., Alexander, S., Dempsey, C., Ariffin, H., Essink, M., Eden, T.O.B., Watts, C., Bates, P.A., Saha, V., 2009. A dyad of lymphoblastic lysosomal cysteine proteases degrades the antileukemic drug l-asparaginase. J. Clin. Invest. 119, 1964–1973.

Pitt, L.A., Tikhonova, A.N., Hu, H., Trimarchi, T., King, B., Gong, Y., Sanchez-Martin, M., Tsirigos, A., Littman, D.R., Ferrando, A.A., Morrison, S.J., Fooksman, D.R., Aifantis, I., Schwab, S.R., 2015. CXCL12-producing vascular endothelial niches control acute T cell leukemia maintenance. Cancer Cell 27, 755–768.

Rahmatallah, Y., Emmert-Streib, F., Glazko, G., 2016. Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. Briefings Bioinf. 17, 393–407.

Ranganathan, P., Mohamed, R., Jayakumar, C., Ramesh, G., 2014. Guidance cue Netrin-1 and the regulation of inflammation in acute and chronic kidney disease. Mediat. Inflamm. 2014, 525891.

Rasheed, W., Bishton, M., Johnstone, R.W., Prince, H.M., 2008. Histone deacetylase inhibitors in lymphoma and solid malignancies. Expert Rev. Anticancer Ther. 8, 413–432.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., Buetow, K.H., 2009. PID: the pathway interaction database. Nucleic Acids Res. 37, D674–D679.

Siegel, P.M., Massagué, J., 2003. Cytostatic and apoptotic actions of TGF-β in homeostasis and cancer. Nat. Rev. Cancer 3, 807–820.

Smyth, G.K., 2005. Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Statistics for Biology and Health. Springer, New York, NY, pp. 397–420.

Staal, F.J.T., Langerak, A.W., 2008. Signaling pathways involved in the development of T-cell acute lymphoblastic leukemia. Haematologica 93, 493–497.

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., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550.

Takahashi, Y., Ikezumi, Y., Saitoh, A., 2017. Rituximab protects podocytes and exerts anti-proteinuric effects in rat adriamycin-induced nephropathy independent of

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002

B-lymphocytes. Nephrol. Carlton Vic. 22, 49–57.
Tarca, A.L., Bhatti, G., Romero, R., 2013. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. PLoS One 8.
Tarca, A.L., Draghici, S., Bhatti, G., Romero, R., 2012. Down-weighting overlapping genes improves gene set analysis. BMC Bioinf. 13, 136.
The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., 2013. The cancer genome Atlas Pan-cancer analysis project. Nat. Genet. 45, 1113–1120.
Tomfohr, J., Lu, J., Kepler, T.B., 2005. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinf. 6, 225.
Tripathi, S., Emmert-Streib, F., 2012. Assessment method for a power analysis to identify differentially expressed pathways. PLoS One 7, e37510.
Van de Wetering, M., de Lau, W., Clevers, H., 2002. WNT signaling and lymphocyte development. Cell 109, S13–S19.
Visani, G., Martinelli, G., Piccaluga, P., Tosi, P., Amabile, M., Pastano, R., Cavo, M., Isidori, A., Tura, S., 2000. Alpha-interferon improves survival and remission duration in P-190BCR-ABL positive adult acute lymphoblastic leukemia. Leukemia 14, 22.
Wahaib, K., Beggs, A.E., Campbell, H., Kodali, L., Ford, P.D., 2016. Panobinostat: a histone deacetylase inhibitor for the treatment of relapsed or refractory multiple myeloma. Am. J. Health-Syst. Pharm. AJHP Off. J. Am. Soc. Health-Syst. Pharm. 73, 441–450.
Yetgin, S., Olgar, S., Aras, T., Cetin, M., Düzova, A., Beylergil, V., Akhan, O., Oğuz, O., Saraçbaşi, O., 2004. Evaluation of kidney damage in patients with acute lymphoblastic leukemia in long-term follow-up: value of renal scan. Am. J. Hematol. 77, 132–139.
Zhang, J.D., Wiemann, S., 2009. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. Bioinformatics 25, 1470–1471.

Please cite this article in press as: Ai, C., Kong, L., CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways, Journal of Genetics and Genomics (2018), https://doi.org/10.1016/j.jgg.2018.08.002