

Integrative Construction of Gene Signatures Based on Fusion of Expression and Ontology Information

Wojciech Łabaj and Andrzej Polanski

Abstract Gene signatures are lists of genes used for summarizing high-throughput gene expression profiling experiments. Various routines for obtaining and analyzing gene signatures in molecular biology studies exist, including statistical testing with false discovery corrections and annotations by gene ontology keywords. Despite the presence of well established routines there are still challenges in efficient application of gene signatures, which include gene signature instability, problems in defining optimal sizes and possible unreliability of inference results. Therefore there are continuous attempts towards improving algorithms for constructing meaningful gene signatures. In this paper we are introducing a methodology for constructing gene signatures, based on the fusion of information coming from statistical tests for differential gene expression analysis and resulting from statistical tests for GO terms enrichment analysis. On the basis of the DNA microarray datasets we are demonstrating that the proposed algorithm for fusion of expression and ontology information leads to improvement of the composition of gene signatures.

Keywords Gene expression · Gene signature · Gene ontology · Functional analysis

1 Introduction

DNA microarrays originally developed to study differential gene expression using diverse populations of RNA undergo continuous refinements and developments, such as disease diagnosis, gene discovery, drug discovery or toxicological research, with regard to gene expression profiling, comparative genomic hybridization, SNP detection and many others. DNA microarray technology also becomes an important element of integrative ‘omics’ studies, complementary to other methods of high throughput molecular measurements, e.g., sequencing [21, 27].

W. Łabaj (✉) · A. Polanski

Institute of Informatics, Silesian University of Technology, Gliwice, Poland
e-mail: wojciech.labaj@polsl.pl

A. Polanski

e-mail: andrzej.polanski@polsl.pl

DNA microarrays measure expression of tens of thousands of genes or gene products simultaneously. They are often summarized by gene signatures [9], which are lists of genes exhibiting certain patterns of expression across experiments. Gene signatures are frequently treated as a starting point for further downstream analysis, where biological conclusions from experimental results are drawn. Therefore, analyzes of gene signatures are an important area of research in bioinformatics. Searching for information concerning genes in signatures can be based annotations stored in dedicated databases, Gene Ontology (GO) [3], KEGG Pathways [17], motifs from InterPro database [15], keywords describing entries from the UniProt database [22] and many others [11].

Gene Ontology is a special source of gene annotation information, due to its hierarchical structure and controlled vocabulary, used by biologists as a standardized nomenclature for many specialized, large biological databases. GO terms are divided into three domains: biological process (BP), molecular function (MF) and cellular component (CC). Genes and gene products have been annotated to the categories of GO using the best currently available knowledge [7]. A hierarchical structure of GO terms, described by the directed acyclic GO graph (DAG) allows for representing biological knowledge at different levels of detail (GO edition from 30/11/2014 has 15 levels in the BP, 15 levels in MF and 13 levels in CC DAG).

A very common approach is that the functional analysis based on GO terms is performed for the domain of interest and level selected arbitrarily. However, two terms located at the same level in the GO graph can provide very different level of detail. Therefore, in the present article the ratio of Information Content (IC) [5] is used to characterize the accuracy of a GO term instead of the graph level.

There is a wide range of tools and algorithms to determine which GO terms are significantly overrepresented considering the given list of gene signatures. The basic approach implemented in many tools for GO terms enrichment/depletion evaluation is based on computation of frequencies, separately for each of the GO terms. In this approach, named *classic* [2] no dependencies related to the topology of the GO graph are taken into account. On the contrary, more advanced algorithms take into account dependencies resulting from the topology of the GO graph. These types of algorithms address the inheritance problem, related to the fact that annotations assigned to more general, ancestral GO terms are inherited from more specific, descendant GO terms, which can mislead the biological interpretation of GO signatures. To overcome this problem a new approach to GO signatures analyses includes additional steps of decorrelation of the GO DAG graph (named *elim*, *weight*, *weight01*, *lea*, *parentchild* [2, 12]).

All these algorithms for enrichment/depletion analysis of GO terms take into account exclusively binary information on gene selection (genes in the signature, all annotated genes, all genes of the considered organism etc.), but drop the full information from expression analysis from the previous steps. In this paper we want to point out that the gene signature construction can be stated as the problem of integrating data derived from gene expression analysis and GO terms assigned to genes.

In this paper we are presenting an approach, where information on the statistical significance of the differential gene expression is fused with the information concerning enrichment/depletion of GO terms assigned to genes. Such an approach does not overlook the valuable information that has been obtained during the differential expression analysis, namely the p-value for each gene. It also allows us to filter out GO terms of low Information Content (IC), which is crucial for many applications, e.g. tumors classification. Unbiased comparison of methods for determining the overrepresentation of GO Terms and assessing the quality of the results is a challenging task. Therefore, we have applied multiple complementary approaches. To this end biological consistency, the number of reduced GO terms and the characteristics of the IC ratio in the remaining GO terms were assessed. In addition to providing a robustness of our comparison we have focused on testing the accuracy of the classification and stability of gene signatures. A side effect of reducing the list of enriched GO terms is limiting the size of the gene signature. We have thus also assessed the impact of the used methods on the size of the gene signatures and its repeatability.

2 Data

Unbiased comparison of our method with the state-of-the-art algorithms for GO Terms enrichment analysis has been performed on two real datasets. We have used multiple DNA microarray experiments related to astrocyte cancer and leukemia.

2.1 Dataset I

Astrocytic brain tumors are cancers of the primary central nervous system (CNS), which develop from astrocytes and are most common glial tumors. They can be divided into two groups according to the way of growth, diffused or localized. In our study we are focusing on the astrocytic brain tumors with the diffused growth, which give poorer prognosis and are assigned to a higher grade according to the World Health Organization (WHO) [20]. We are further confining the research to the two most common tumors from this group, namely anaplastic astrocytoma (AA) and glioblastoma multiforme (GBM). Additionally, there are two different forms of the GBM, primary GBM arising de novo and secondary GBM arising from lower grade diffuse astrocytoma [23]. Primary and secondary glioblastomas are histologically indistinguishable, except the facts that the frequency of extensive necrosis is higher for the primary GBM and the frequency of oligodendroglioma components is higher for the secondary GBM [14]. Similar histopathology of glioblastomas may be due to similarity of genetic alterations behind their growths.

On the basis of the biological and clinical characteristics of astrocytoma and primary and secondary glioblastomas, it seems an interesting issue to design an experiment, aimed to highlight both the differences and similarities between these

cancers. We have explored the Gene Expression Omnibus (GEO) database [4] for experiments where gene expression profiling corresponding to the above-mentioned tumors and their comparisons to normal tissues was performed. The datasets found in GEO originate from three studies [6, 13, 19], which relied on comparisons of the three types of brain tumors, AA, primary GBM (GBM.P) and secondary GBM (GBM.S) with the normal brain tissues (NBT).

2.2 Dataset II

Although the first data set is very compelling, it lacks the sufficient number of samples for statistically significant classification procedure. In order to be able to perform such classification another data set has been compiled. Here we have focused on a more frequently studied disease (leukemia), so that we were able to collect sufficient number of microarray samples to perform the classification.

Leukemia is a group of cancers that usually begins in the bone marrow and results in a high numbers of abnormal white blood cells, which are called blasts. It is part of a broader group of neoplasms which affect the blood, bone marrow, and lymphoid system, known as tumors of the hematopoietic and lymphoid tissues [26]. But the exact cause of developing leukemia is unknown, it is even believed that each type of leukemia has a different cause [16].

There are four main types of leukemia: acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL) and chronic myeloid leukemia (CML), as well as a number of less common types [25]. In this article we have focused on studies where ALL and AML were investigated.

ALL is the most common type of leukemia in young children. This disease also affects adults, especially those aged 65 and older, whereas AML occurs more commonly in adults than in children.

We searched the GEO database for the experiments related to these two types of leukemia. The gene expression datasets originating from three studies were selected [10, 18]. We have chosen 84 samples from these experiments—42 samples for each type of leukemia respectively.

3 Methods

3.1 DATA Normalization and Differential Expression Analysis

All microarray analyses were performed with use of state of the art academic processing software. A custom CDF file for RefSeq annotation from BrainArray [8, 24] was used as it provides the latest genome and transcriptome information. GC Robust Multi-array Average (GCRMA) was used as background adjustment on Affymetrix

microarray probe-level data. In this method the sequence information is summarized by base types at each probe position, in a more complex way than the simple GC content. Variance Stabilization and Normalization (VSN) has been used for normalizing microarray intensity. It ensures that the variance remains nearly constant over the whole intensity spectrum. VSN has been used for normalization within experiments and normalization of all microarrays together (also to remove batch effect). As a summarization step the Affymetrix Probe-Level Modeling (affyPLM) was used. The next step was to create a model of comparison between case and control, after which data was fitted to the model. For this purpose Linear Models for Microarray Data (limma) were used, which provide the ability to analyze comparisons between many RNA targets simultaneously. Holm correction has been applied in order to correct statistical test results for multiple comparison, with threshold equal to 0.05. As a result, gene signatures were obtained, which allow for enrichment analysis of GO terms.

3.2 *Enrichments Analysis of GO Terms*

R package '*topGO*' [1] is an effective tool for semi-automatic enrichment analysis of GO terms. The package includes a set of ready-to-use functions for carrying out GO enrichments analysis. Not all combinations between algorithms and statistical tests currently supported by topGO are allowed [1].

The *elim* and *weight* algorithms were introduced by Alexa et al. [2]. Both methods investigate the nodes in the GO DAG from bottom to top and introduce weights for genes. In *elim* method the weighting process boils down to assigning 0 or 1 to genes. This means that the *elim* method eliminates the genes from the list of significant genes in ascendant GO terms, once they have been found to be associated with a statistically significant GO term. In *weight* method the process of assigning weights to genes is described as a function which always attains a value in the interval [0, 1]. The default algorithm used by the *topGO* package is *weight01*, which is a mixture of the *elim* and the *weight* algorithms. The *parentChild* algorithm was introduced by Grossmann et al. [12]. It measures overrepresentation of the GO term in the context of gene annotations to the parents of the term, not to the root.

During our analysis Fisher's Exact Test was used exclusively for simplicity of the comparison. It belongs to the class of exact tests, because the significance of the deviation from a null hypothesis can be calculated exactly, rather than relying on an approximation, as with many statistical tests.

3.3 *Enrichment Analysis by Fusion*

Our new heuristic approach for measuring overrepresentation of GO terms can be described as a method for filtering results obtained by the *classic* annotation

routine. The method requires as input data results of the *classic* algorithm and results of differentiation tests (in the form of a list of genes with corresponding p-values). Output data for each GO term are two indexes, namely enrichment index (EI) and differentiation index (DI). Fisher’s combined probability test was used as a pattern of conduct to transform p-values from the enrichment test for the GO term and to fuse p-values from differentiation tests for genes annotated to this GO term. For standardization purposes indexes for each GO term are scaled to 0–1 range, which provides the method’s applicability for any dataset. The methods of indexes’ calculation are shown in formulae (1) and (2).

$$EI_{GO\text{Terms}_i} = -2\ln (ET_{GO\text{Term}_i}) \tag{1}$$

where,
ET—enrichment test (*classic* method).

$$DI_{GO\text{Term}_i} = \frac{-2 \sum_{j=1}^n \ln (DT_{gene_j})}{n} \tag{2}$$

where,
DT—differentiation test and
n—number of annotation genes to i-th GO term.

The filtering condition is the distance on two-dimensional plane of indexes between each GO term and the point (1,1). Exemplary result of our new approach are shown in the diagrams below (Fig. 1).

Our goal was to fuse information coming from the enrichment analysis of GO terms and differentiation analysis. Therefore, the cut-off was set as on arc of a circle of

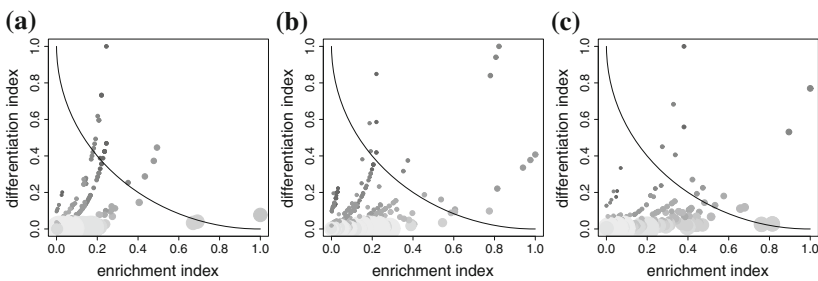


Fig. 1 Diagram presents an example of results for the new method. The x-axis is the enrichment index (EI) and y-axis is the differentiation index (DI). Color and size of the dots describe the information content (IC) ratio (*dark grey*—highest, *light grey*—lowest) and the number of significant genes associated with the GO term (the more genes the larger dot). Filtering condition is distance between each GO Term and the point (1,1) which correspond to perfect, maximal enrichment. Cut-off value has been chosen empirically and set to 1 (GO terms with higher distance are filtered out). The black line is the boundary cut-off. **a** AA, **b** GBM.P, **c** GBM.S

radius equal to 1. This contributes to the rejection of GO terms with low information content. However, in the analyzed cases, there were a few GO terms, which, despite low information content (IC) had a high rate of EI. They can be an integral part of the particular case and the method allows to retain them.

3.4 Criteria for Comparison of Algorithms

The Dataset I was used to investigate the characteristic of results of each method for enrichment analysis of GO Terms. First step was to check the cohesion of results with biological knowledge. More precisely, we assume that both tumors GBM.P and GBM.S should have a common part of the significant genes as well as have a common part of enriched GO Terms, which is expected to be larger than the common parts of GBM and AA.

Another step in the comparison is to check the quality of the results by examining the information content (IC) of enriched GO Terms. Dataset I will also provide us information on the number of reduced GO Terms, which entails a reduction in the size of the gene signature.

The Dataset II was used then for classification and stability analysis of gene signatures. As a validation method k-fold cross validation algorithm was selected. Therefore Dataset II was divided into 6 subsets, each of which contained 7 samples of both types of leukemia. We used this kind of division to maintain equal participation of both leukemia samples from different experiments. Training and validation sets were then subjected to data normalization, differential expression analysis, enrichment analysis of GO Terms and re-annotation of the significant GO Terms back to the genes to obtain gene signatures (features). K-nearest neighbors' algorithm was used as a classification method. Classification was carried out for different neighborhood sizes (from 1 to 6) and accuracy was the parameter determining the quality of a classifier.

The second metric calculated on the basis of Dataset II is gene signature stability. It assesses the reproducibility of the gene signature at all steps of classification. We chose two primary methods: the basic method of calculating this measure and the method, which introduces the adjustment related to the size of the reduced genes [9]. As this article presents new approaches for reduction of gene signatures based on GO and comparing it with other commonly used methods, therefore, the use of the second measure gene signature stability seems to be more appropriate.

4 Results

4.1 Reduction Assessment

Differential expression analysis (see. Methods) of Dataset I has provided a list of significant genes for each tumor. Not all of these genes are annotated to GO Terms therefore, these reduced gene signatures were used for comparison of *classic* routine of GO terms enrichment analysis with selected reduction methods for AA, GBM.P and GBM.S tumor experimental data.

The *classic* approach gives results in line with biological backgrounds of diseases, with numerous GO terms shared by GBM.P and GBM.S (Table 1) and few GO terms shared between AA and GBM. Our method completely removes common GO terms between AA and GBM, where the other methods with decorrelation step preserve a certain proportion of these terms. Some of the GO terms may be linked to different tumors, which is associated with low biological information. Removal of such GO terms can be useful in the tumor classification tasks, or for better understanding the characteristics of the specific tumor. As a result, the filtering performed by our new algorithm leaves only GO terms specific to the particular case.

The summarization of the major changes in resulted lists of GO terms for GBM tumors can be seen in Table 1. We present here what fraction of overrepresented GO terms is reduced for GBM.P, GBM.S and the common part. Additionally, it may be seen that our method radically reduced the common part of GBM, which can be helpful in tumour classification or retrieving important information related to the inspected diseases. In this way, it can be helpful eg. in drug discovery.

More details about influence of the reduction method can be seen on distribution of information content (IC, see Methods) for each of the used methods (Fig. 2). Table 2 presents complementary information with the percentage of enriched GO Terms which have IC ratio greater than 0.7. This threshold has been selected empirically

Table 1 Summary of size and reduced fraction of overrepresented GO terms for GBM.P, GBM.S and the common part

Method	GBM.P	Reduced GBM.P (%)	Common part of GBM	Reduced common part of GBM (%)	GBM.S	Reduced GBM.S (%)
Classic	409	–	160	–	251	–
Elim	154	62.3	41	74.4	69	72.5
Weight	101	75.3	26	83.8	41	83.7
Weight01	114	72.1	32	80	45	82.1
Lea	129	68.5	35	78.1	55	78.1
Parentchild	244	40.3	79	50.6	149	40.6
New	20	95.1	5	96.9	16	93.6

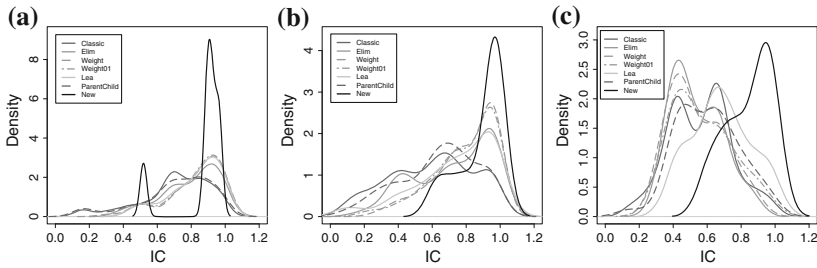


Fig. 2 Probability density function for all methods for each tumor. To estimate the probability density function of the random variable we used nonparametric methods, known as Kernel density estimation (KDE). **a** AA, **b** GBM.P, **c** GBM.S

Table 2 Table contains the percentage of enriched GO terms which have IC ratio greater than 0.7 for all methods and tumors

IC > 0.7	Classic (%)	Elim (%)	Weight (%)	Weight01 (%)	Lea (%)	Parentchild (%)	New (%)
AA	48.2	71.4	78.3	76.6	76.6	41.0	85.7
GBM.P	32.4	53.6	65.0	62.8	56.8	37.4	87.5
GBM.S	20.1	37.3	45.0	43.2	33.3	21.7	86.7
Mean	33.6	54.1	62.8	60.9	55.6	33.4	86.6

and reflects the separation of GO terms with high information content from ones with low information content.

It is clear that the shape of PDF for our method is distinct and the least similar to other methods. It has a clear shift of the main peak towards higher IC values, which reflects the reduction of the fraction of GO terms with low information content. This can be clearly seen in Table 2.

Reduction of GO Terms as a side effect leads to reduction of gene signature sizes, which is convenient as it reduces dimensionality of any follow-up analysis. Table 3 presents a summary of gene signature sizes and their reduction for GBM.P, GBM.S and the common part.

4.2 Classification Quality

Comparison based on Dataset II was performed on the reduced size of the gene signatures. Data after summarization steps was used as input features for classification and validation. (see. Methods) Accuracy of classification was presented in the form of a table (Table 4).

Analysis of classification accuracy has shown that the method hierarchy is highly dependent on the neighborhood size we have used (Table 4). Interestingly, our method

Table 3 Summary of size and reduced fraction of gene signatures for GBM.P, GBM.S and the common part

Method	GBM.P	Reduced GBM.P (%)	Common part of GBM	Reduced common part of GBM (%)	GBM.S	Reduced GBM.S (%)
Base	66	–	36	–	48	–
Classic	64	3.0	29	19.4	41	14.6
Elim	53	19.7	23	36.1	35	27.1
Weight	49	25.8	21	41.7	33	31.3
Weight01	51	22.7	22	38.9	34	29.2
Lea	60	9.1	23	36.1	35	27.1
Parentchild	62	6.1	25	30.6	37	22.9
New	27	59.1	9	75.0	25	47.9

Table 4 Summary of classification results of accuracy for each method and different neighborhood sizes

	Classic (%)	Elim (%)	Weight (%)	Weight01 (%)	Lea (%)	Parentchild (%)	New (%)
1	64.3	75.0	72.6	64.3	64.3	64.3	85.7
2	70.2	76.2	77.4	71.4	70.2	70.2	83.3
3	82.1	84.5	84.5	81.0	83.3	82.1	81.0
4	84.5	85.7	84.5	83.3	85.7	84.5	81.0
5	85.7	84.5	84.5	85.7	85.7	85.7	79.8
6	91.7	88.1	86.9	88.1	90.5	89.3	86.9

has the smallest variability of classification accuracy for all investigated sizes of the neighborhood. It is important to note that accuracy for the neighborhood equal to or greater than 3 for all methods is very similar. Therefore, based just on the accuracy metric, it is hard to determine which method gives the optimal signature for classification purposes.

For this reason, an additional approach was examined—stability of the gene signatures. This measure allows to assess reproducibility of gene signatures, taking into account all steps of the validation process. Results for all methods are shown in Table 5.

The highest score for adjusted stability (stabA) was obtained with our method. The best signature stability together with the lowest accuracy sensitivity to the neighborhood size and with not worse overall accuracy level, indicate that applying our new method provides an improvement toward the composition of the gene signatures.

Table 5 Summary for gene signature stability measures

	Classic (%)	Elim (%)	Weight (%)	Weight01 (%)	Lea (%)	Parentchild (%)	New (%)
StabA	26.18	30.21	30.70	28.16	27.41	26.58	32.69

5 Discussion

In the post genomic era, there is an enormity of publicly available molecular biology data, which originates from a variety of biological research projects. These data often include overlapping fields of science. The present state of molecular biology knowledge indicates the possibility of using comprehensive and integrative approaches to understand the biological processes. Therefore, methods for high-throughput data integration are a hot topic of research for molecular biology.

In this article, we have compared our new method with the reference, state-of-the-art algorithms for GO terms enrichment analysis. All reference methods take into account solely information concerning the size of sets (all investigated genes, significant genes and annotated number of genes from first and second group to each GO term). In contrast, our approach takes into account additional information originating from analysis of differential expression of genes. Thanks to this approach we do not omit the valuable information that has been obtained during the differential expression analysis, namely the p-value for each gene.

As previously mentioned there is no well-established procedure for comparison of methods for enrichment analysis of GO terms. Therefore, we have applied several possible approaches aiming at more robust conclusions. On Dataset I we have verified the characteristic of results for each method by means of the cohesion of the results with biological knowledge and quality of the results by examining the information content (IC) of enriched GO terms. Impact of GO terms reduction on gene signatures size reduction was also investigated. We have shown that the all of the methods for enrichment analysis of GO terms with decorrelation step both reduce number of enriched GO terms and maintain fraction of GO terms shared between AA and GBM. The common part of GO terms shared between AA and GBM is completely removed by our method, which in addition filters out GO terms of low Information Content (IC), leaving in only the GO terms which are specific to the particular case. In the end, it can help in tumor classification or can provide more information about characteristic of the tumors by reducing the dimensionality of the follow up analysis.

One of the major aims in molecular biology is to obtain a stable gene signature, which will be helpful in many areas of research. Still there are existing problems such as defining the sizes (numbers of genes) of gene signatures and possible unreliability of results of inference based on gene signatures. For this reason, in the last stage of our comparison we were investigating Dataset II for the classification accuracy as well as the stability of gene signatures. We have shown that our method is characterized by the best signature stability. Together with the lowest sensitivity of classification

accuracy to the used neighborhood size and with not worse overall accuracy level, it indicates that the proposed reduction approach provides an improvement in the composition of the gene signatures in comparison to other methods.

We have shown that our method reduces the number of enriched GO terms, focusing on the meaningful ones. Thus, it facilitates the analysis and biological interpretation. It also contributed to reducing the dimensionality of gene signatures, which ultimately improved accuracy of classification and stability of gene signatures.

Acknowledgments The authors are grateful to Pawel P. Labaj and Anna Papiez for helpful discussions. The work was financially supported by SUT - BKM/525/RAU-2/2014. Calculations were carried out using the infrastructure supported by POIG.02.03.01-24-099/13 grant: GeCONiI - Upper-Silesian Center for Scientific Computation.

References

1. Alexa, A., Rahnenführer, J.: Gene set enrichment analysis with topGO (2009). http://rgm.ogalab.net/RGM-files/R_BC/download/topGO/inst/doc/topGO.pdf
2. Alexa, A., Rahnenführer, J., Lengauer, T.: Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**(13), 1600–1607 (2006)
3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**(1), 25–29 (2000)
4. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al.: NCBI GEO: archive for functional genomics data sets update. *Nucleic Acids Res.* **41**(D1), D991–D995 (2013)
5. Bérard, S., Tichit, L., Herrmann, C.: ClusterInspector: a tool to visualize ontology-based relationships between biological entities. *Actes des Journées Ouvertes Biologie Informatique Mathématiques*, pp. 447–457. Lyon (2005)
6. Bielen, A., Perryman, L., Box, G.M., Valenti, M., de Haven Brandon, A., Martins, V., Jury, A., Popov, S., Gowan, S., Jeay, S., et al.: Enhanced efficacy of IGF1R inhibition in pediatric glioblastoma by combinatorial targeting of PDGFR α/β . *Mol. Cancer Ther.* **10**(8), 1407–1418 (2011)
7. Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler, R.: The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Res.* **32**(suppl 1), D262–D266 (2004)
8. Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., et al.: Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**(20), e175–e175 (2005)
9. Davis, C.A., Gerick, F., Hintermair, V., Friedel, C.C., Fundel, K., Küffner, R., Zimmer, R.: Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics* **22**(19), 2356–2363 (2006)
10. Figueroa, M.E., Wouters, B.J., Skrabanek, L., Glass, J., Li, Y., Erpelinck-Verschueren, C.A., Langerak, A.W., Lowenberg, B., Fazzari, M., Greally, J.M., et al.: Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features. *Blood* **113**(12), 2795 (2009)
11. Galperin, M.Y., Rigden, D.J., Fernández-Suárez, X.M.: The 2015 nucleic acids research database issue and molecular biology database collection. *Nucleic Acids Res.* **43**(D1), D1–D5 (2015)

12. Grossmann, S., Bauer, S., Robinson, P.N., Vingron, M.: Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis. *Bioinformatics* **23**(22), 3024–3031 (2007)
13. Grzmil, M., Morin, P., Lino, M.M., Merlo, A., Frank, S., Wang, Y., Moncayo, G., Hemmings, B.A.: MAP kinase-interacting kinase 1 regulates SMAD2-dependent TGF- β signaling pathway in human glioblastoma. *Cancer Res.* **71**(6), 2392–2402 (2011)
14. Homma, T., Fukushima, T., Vaccarella, S., Yonekawa, Y., Di Patre, P.L., Franceschi, S., Ohgaki, H.: Correlation among pathology, genotype, and patient outcomes in glioblastoma. *J. Neuro-pathol. Exp. Neurol.* **65**(9), 846–854 (2006)
15. Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., et al.: InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**(suppl 1), D211–D215 (2009)
16. Hutter, J.J.: Childhood leukemia. *Pediatr. Rev.* **31**(6), 234–241 (2010)
17. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., et al.: KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**(suppl 1), D480–D484 (2008)
18. Krupp, M., Itzel, T., Maass, T., Hildebrandt, A., Galle, P.R., Teufel, A.: Cell LineNavigator: a workbench for cancer cell line analysis. *Nucleic Acids Res.* **41**(D1), D942–D948 (2013)
19. Liu, Z., Yao, Z., Li, C., Lu, Y., Gao, C.: Gene expression profiling in human high-grade astrocytomas. *Comp. Funct. Genomics* (2011)
20. Louis, D.N., Ohgaki, H., Wiestler, O.D., Cavenee, W.K., Burger, P.C., Jouvet, A., Scheithauer, B.W., Kleihues, P.: The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathol.* **114**(2), 97–109 (2007)
21. Nat. Biotechnol. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. **32**(9), 903–914 (2014)
22. *Nucleic Acids Res.* The universal protein resource (UniProt) in 2010. **38**(suppl 1), D142–D148 (2010)
23. Ohgaki, H., Kleihues, P.: The definition of primary and secondary glioblastoma. *Clin. Cancer Res.* **19**(4), 764–772 (2013)
24. Sandberg, R., Larsson, O.: Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinform.* **8**(1), 48 (2007)
25. Stewart, B.W., Wild, C.P.: World cancer report 2014. IARC Press, International Agency for Research on Cancer, Geneva (2008)
26. Vardiman, J.W., Thiele, J., Arber, D.A., Brunning, R.D., Borowitz, M.J., Porwit, A., Harris, N.L., Le Beau, M.M., Hellström-Lindberg, E., Tefferi, A., et al.: The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood* **114**(5), 937–951 (2009)
27. Wang, C., Gong, B., Bushel, P.R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., et al.: The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.* **32**(9), 926–932 (2014)