BMC Bioinformatics

SOFTWARE

Open Access

PhenoFam-gene set enrichment analysis through protein structural information

Maciej Paszkowski-Rogacz*[1], Mikolaj Slabicki[1], M Teresa Pisabarro[2] and Frank Buchholz*[1]

Abstract

Background: With the current technological advances in high-throughput biology, the necessity to develop tools that help to analyse the massive amount of data being generated is evident. A powerful method of inspecting large-scale data sets is gene set enrichment analysis (GSEA) and investigation of protein structural features can guide determining the function of individual genes. However, a convenient tool that combines these two features to aid in high-throughput data analysis has not been developed yet. In order to fill this niche, we developed the user-friendly, web-based application, PhenoFam.

Results: PhenoFam performs gene set enrichment analysis by employing structural and functional information on families of protein domains as annotation terms. Our tool is designed to analyse complete sets of results from quantitative high-throughput studies (gene expression microarrays, functional RNAi screens, *etc.*) without prior pre-filtering or hits-selection steps. PhenoFam utilizes Ensembl databases to link a list of user-provided identifiers with protein features from the InterPro database, and assesses whether results associated with individual domains differ significantly from the overall population. To demonstrate the utility of PhenoFam we analysed a genome-wide RNA interference screen and discovered a novel function of plexins containing the cytoplasmic RasGAP domain. Furthermore, a PhenoFam analysis of breast cancer gene expression profiles revealed a link between breast carcinoma and altered expression of PX domain containing proteins.

Conclusions: PhenoFam provides a user-friendly, easily accessible web interface to perform GSEA based on high-throughput data sets and structural-functional protein information, and therefore aids in functional annotation of genes.

Background

Analysis of large sets of results derived from high-throughput experiments is a challenging but promising field of study. Enrichment analysis is a very powerful strategy helping researchers in identifying biological processes or pathways related to their studies. Most of the currently available tools (*i.e.* Onto-Express [1], DAVID [2], FatiGO+ [3], ConceptGene [4] and others reviewed in [5]) search for enrichment of Gene Ontology (GO) terms [6], KEGG pathways [7] or other functional properties in a pre-selected subset of genes by contrasting it with the background set, usually a whole genome. This approach strongly relies on a chosen hit selection algorithm and user-defined thresholds. Moreover, the experimental results (*i.e.* level of expression or phenotype strength) are not considered. There are few applications overcoming these limitations by performing gene set enrichment analysis (GSEA) [8]. They search for gene annotations enriched on the top or the bottom of a complete list of genes ranked by their experimental values. This allows even mild effects to contribute to the overall enrichment score. However, to our knowledge, annotations used by available GSEA tools have so far primarily been used in combination with GO terms, pathways or transcription factors, and only few of these applications are web-based (*e.g.* GSEA [9], FatiScan [3], GeneTrail [10]).

In recent years, access to high-resolution protein structural information has increased considerably. Many new structures reveal the presence of domains known from other proteins, and the domain composition of a protein can help forming a hypothesis about its biological func-

* Correspondence: paszkows@mpi-cbg.de, buchholz@mpi-cbg.de
Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307 Dresden, Germany
Full list of author information is available at the end of the article

tion (*e.g.* a homeodomain fold indicates a transcription factor activity involved in cellular differentiation [11]). Moreover, Hahne *et al.* demonstrated, that the domain composition of proteins could be used for predicting their pathway membership [12]. There are many databases classifying and providing information about protein families, domains, regions and functionally relevant sites. InterPro [13] constitutes a repository that integrates a number of the most well established sources of data: PROSITE [14], HAMAP [15], Pfam [16], PRINTS [17], ProDom [18], SMART [19], TIGRFAMs [20], PIRSF [21], SUPERFAMILY [22], Gene3D [23] and PANTHER [24]. We have developed a GSEA web application that can be used for analysing data from large-scale experiments (phenotypes, gene expression, *etc.*). Our tool combines the experimental results with annotations from the databases integrated in InterPro (called 'member databases'), thereby allowing a streamlined structure/function annotation of proteins. Utilization of information about protein domain families in GSEA is a novel approach that can be used in parallel to other enrichment analysis applications.

Implementation

Data management

PhenoFam is a Java web application running on a Tomcat 5.5 server. It uses a MySQL database to store mappings between various protein, gene or probe names and identifiers related to member databases of InterPro. This database is an easily updatable compilation of the current releases of the Ensembl database [25]. Client-server communication is mainly handled by AJAX technologies. User-uploaded data sets and calculation results are stored as session objects on the server side for at least 30 minutes after closing the browser window.

Identifiers association

One of the key features of our application is that it accepts as input a wide range of identifiers used in all genomes integrated in the Ensembl database [25]. Identifiers provided by the user are translated into respective Ensembl (gene, or transcript) identifiers and, using mappings from the InterPro database, linked to none, one or several protein domains or features from different InterPro database members (Figure 1). Reversing the mapping, each protein domain is linked with at least one user identifier and at least one experimental value.

It must be noted that all identifier mappings are based on contents of the Ensembl database, which establishes the links based on sequence similarity of entities stored in remote databases to sequences stored in Ensembl. This approach provides the highest quality of associations. However, care must be taken if gene-related identifiers are used. Due to alternative splicing, different gene products may be composed of different protein domains or even encode different proteins (*i.e.* shift in the reading frame). In such cases, a value associated with the user-provided identifier is mapped to all possible protein features that can be associated with the gene (Figure 1).

Gene set enrichment analysis

To test if a set of values associated with a given domain is significantly higher or lower than the remaining set of values, we use the Mann-Whitney U-test. The U-test is the most powerful nonparametric alternative to the Student's t-test. Its main advantage is that it makes no assumptions about the underlying distributions and is more robust in case of outliers. The U-test is also implemented in other popular GSEA tools, *i.e.* GeneTrail [10] or PANTHER [26,27]. Other applications (*i.e.* GOdist [28], GSEA [9]) implemented the Kolmogorov-Smirnov (KS) test, another non-parametric procedure that checks whether two samples (values associated with a given domain and the other values) may be assumed to come from the same distribution. However, the KS test is also sensitive to differences in the general shapes of the distributions, which limits its use for our PhenoFam application. Parametric analysis, which was proposed by Kim *et al.* and implemented in PAGE [29], is also not suitable for GSEA of protein domains because many domains are associated with small number of proteins (< 10). In those cases, the normality criteria required for parametric tests might not be satisfied. Adjustment for multiple testing is done using the false discovery rate (FDR) control procedure designed by Benjamini and Hochberg [30] and resulting q-values are obtained by applying Storey's algorithm [31,32]. Additionally, we calculate a Herrnstein's ρ statistic [33], which is an unbiased measure of the overlap between distributions of values in the two compared sets. It can reach values between 0 and 1, where 0.5 indicates a complete overlap of the two distributions and both extreme values show a complete separation. This statistic shows how much a median of domain-associated values differs from a median of the other values, and together with the p-value can help identifying domains of interest. We recommend using it for sorting results that passed the significance-threshold criteria. Due to the fact that InterPro is a collection of partially redundant databases, the enrichment analysis and the adjustment for multiple testing procedure are performed for each database independently. Otherwise, treating InterPro as a uniform set of annotations would lead to a significant underestimation of the results.

User interface

To implement the user interface and to ensure compatibility with all major browsers, we used the Google Web Toolkit (GWT) framework. We have designed a simple

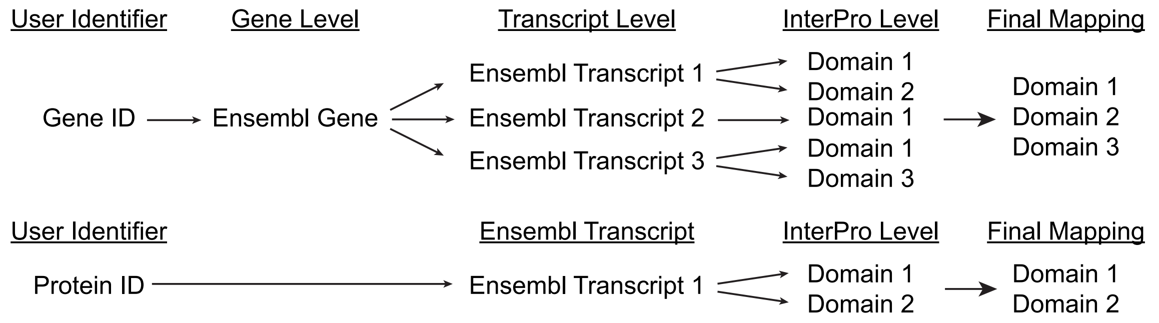
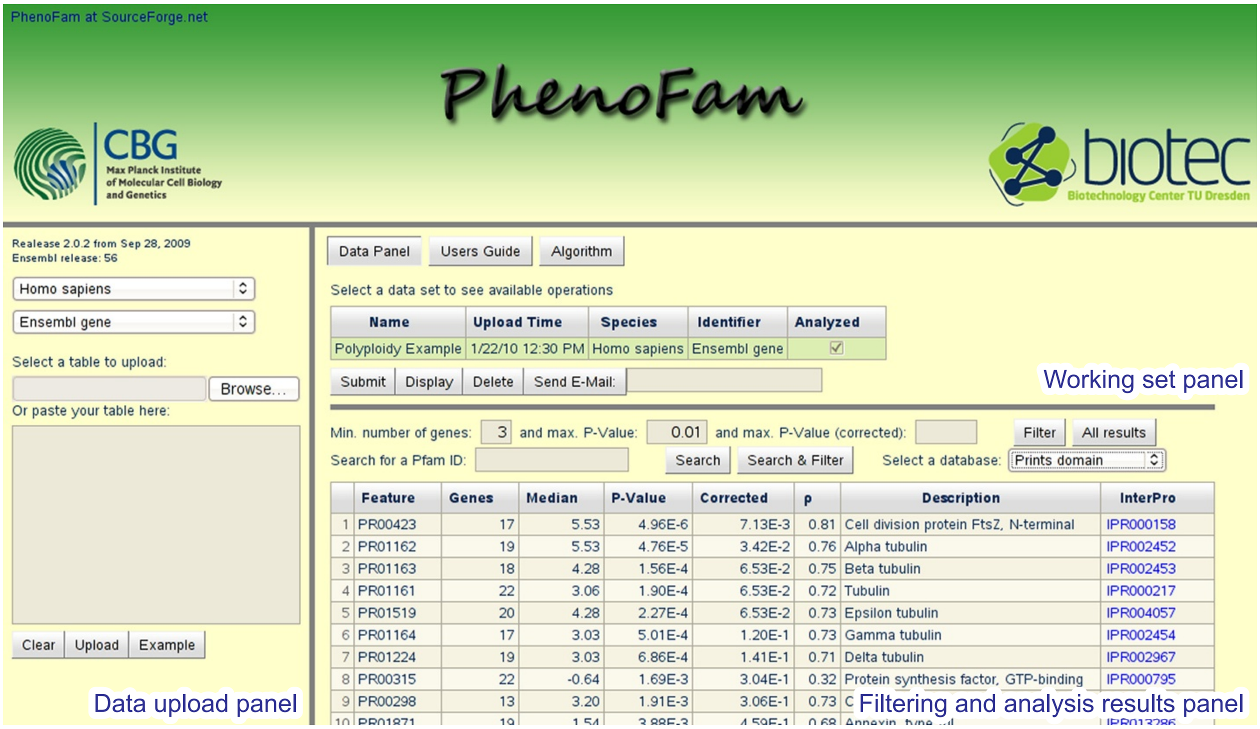


Figure 1 Identifier mapping procedure. Gene-related identifiers (*e.g.* Gene Names) are mapped to Ensembl Gene IDs and further to all protein-coding Ensembl Transcript IDs. Each of the transcripts is associated with protein features from the InterPro database. Redundant identifiers are removed in the final mapping. Protein- or transcript-related identifiers (*e.g.* UniProt IDs) are directly linked to Ensembl Transcript IDs and then to protein features.

and user-friendly data management system for storing uploaded data sets and the analysis results. It allows users to investigate and compare multiple data sets at the same time.

Our GSEA algorithm reports the following information: a member database identifier, the domain description, a number of user identifiers associated with the domain, a median of the values, a p-value reported by the Mann-Whitney U-test, a FDR corrected p-value, a ρ statistic and the InterPro identifier. The results associated with one of the selected InterPro member databases are displayed in a pageable table (Figure 2) that can be sorted and filtered. We also provide a possibility to search for specific domains. For each selected domain, we also show a table of associated values together with original identifiers, UniProt accessions and descriptions. A brief user's guide to PhenoFam is provided in Additional file 1, as well as on the application web site.



Realease 2.0.2 from Sep 28, 2009
Ensembl release: 56

Homo sapiens

Ensembl gene

Select a table to upload:

Or paste your table here:

Select a data set to see available operations

Name	Upload Time	Species	Identifier	Analyzed
Polyploidy Example	1/22/10 12:30 PM	Homo sapiens	Ensembl gene	☑

Min. number of genes: 3 and max. P-Value: 0.01 and max. P-Value (corrected):

Search for a Pfam ID:

Select a database: Prints domain

	Feature	Genes	Median	P-Value	Corrected	p	Description	InterPro
1	PR00423	17	5.53	4.96E-6	7.13E-3	0.81	Cell division protein FtsZ, N-terminal	IPR000158
2	PR01162	19	5.53	4.76E-5	3.42E-2	0.76	Alpha tubulin	IPR002452
3	PR01163	18	4.28	1.56E-4	6.53E-2	0.75	Beta tubulin	IPR002453
4	PR01161	22	3.06	1.90E-4	6.53E-2	0.72	Tubulin	IPR000217
5	PR01519	20	4.28	2.27E-4	6.53E-2	0.73	Epsilon tubulin	IPR004057
6	PR01164	17	3.03	5.01E-4	1.20E-1	0.73	Gamma tubulin	IPR002454
7	PR01224	19	3.03	6.86E-4	1.41E-1	0.71	Delta tubulin	IPR002967
8	PR00315	22	-0.64	1.69E-3	3.04E-1	0.32	Protein synthesis factor, GTP-binding	IPR000795
9	PR00298	13	3.20	1.91E-3	3.06E-1	0.73	C	
10	PR01871	19	1.54	3.88E-3	4.59E-1	0.68	Annexin, type ...	IPR013286

Figure 2 PhenoFam web interface. The main user interface display is divided into three panels. The 'Data upload panel' allows uploading data sets for the GSEA analysis either by pasting the data or by selecting a text file. All uploaded data sets are displayed in the 'Working set panel', where the user can submit data for the analysis, view the results in the browser or send them by e-mail. The sortable table with results is displayed in the 'Filtering and analysis results panel'. The top section of the panel contains a form that provides searching and filtering functionality. The displayed table contains a list of significantly enriched PRINTS domains.

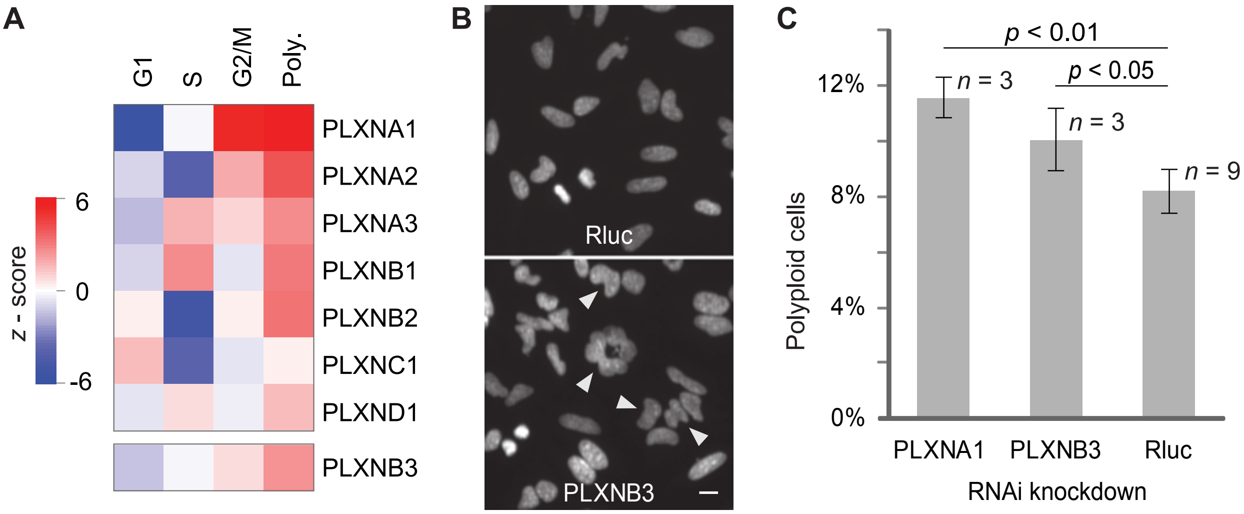


Figure 3 Plexins are enriched for polyploidy RNAi phenotypes. **A**. Cell cycle phenotypic profiles after knockdown of plexins. The top heatmap shows data extracted from the genome-wide RNAi screen [34], and the bottom heatmap shows the cell cycle profile after knock-down of PLXNB3. The PLXNB3 profile was obtained from the automated analysis of microscopy images, which quantifies proportions of cells in different phases of the cell cycle. *z*-scores were calculated by normalization to the mean and standard deviation of respective values obtained from the analysis of negative control images. **B**. Fluorescence microscopy images of HeLa cells 48 hours after transfection of esiRNA (endoribonuclease-prepared siRNA) against Rluc (negative control) and PLXNB3. The images show DAPI-stained nuclei, and arrows indicate cells with polyploidy phenotype. The scale bar represents 10 μm. Investigation of both images shows that the knock-down of PLXNB3 results in a polyploidy phenotype compared to the negative control condition. **C**. Quantification of the image analysis of polyploidy phenotypes among cells treated by different silencing triggers (≈ 5 000 cells per replicate). Error bars represent one SD. Student's *t*-test confirmed that the ratio of polyploid cells is significantly increased after knock-down of indicated plexins, compared to the control condition (Rluc).

Results

PhenoFam allows many data sets as the starting point, such as results of microarray studies, systematic RNA interference (RNAi) screens, ChIP-Chip/ChIP-Seq experiments or comparative mass-spectrometry (*i.e.* SILAC) results. To test the utility of PhenoFam, we analysed a data-set derived from a genome-scale cell cycle progression RNAi screen [34]. In this screen, a genome-wide study of genes was carried out providing *z*-scores for cell cycle progression phenotypes (*i.e.* cells in G1, S, G2/M phases and polyploidy) for each knockdown.

A PhenoFam analysis of the complete RNAi data-set revealed that plexins containing a cytoplasmic RasGAP domain were enriched ($p < 0.005$) for polyploidy phenotypes (Figure 3A, Table 1). Knockdown of most transcripts encoding these genes resulted in an increase of polyploidy cells. Although in the published RNAi screen [34] only genes with the strongest polyploidy phenotypes of *z*-score > 6 were selected for further investigation, the PhenoFam analysis suggests that plexins not passing this criteria might also have a function in cytokinesis.

Moreover, based on this result we predicted that knockdown of the gene PLXNB3, which belongs to the same family, but had not been tested in the screen, would also increase the degree of polyploidy. Indeed, an increased number of polyploid cells were measured after PLXNB3 knockdown (Figure 3), indicating that depletion of this gene, like other plexins with cytoplasmic RasGAP domains, influences proper cytokinesis. This example demonstrates that PhenoFam can be a valuable support for selecting hits from the RNAi screens.

To show that PhenoFam is also suitable for analysis of other large-scale data-sets, we examined publicly available gene expression data that compares transcriptomes of human breast carcinoma and healthy tissue [35]. GSEA of this data-set using GeneTrail [10] showed that genes whose expression is altered in breast cancer are significantly enriched with the 'signal transduction' and 'cell differentiation' gene ontologies, highliting the importance of these biological processes during cellular transformation (data not shown). However, the analysis with GeneTrail did not provide information of enrichment of certain protein domains. In contrast, analysis of the same data-set with PhenoFam showed that among differentially expressed genes, Ras-family proteins and phox (PX) domain-containing proteins were enriched ($p < 0.001$, data not shown).

Ras GTPases are known to play a role in breast cancer development [36] and, therefore, it is not surprising that this group of proteins was enriched in this set. Proteins containing a PX domain are involved in cell signalling, vesicular trafficking, protein sorting and lipid modifica-

Table 1: Normalized values of polyploidy RNAi phenotypes of plexins, from the primary cell cycle progression screen.

Ensembl ID	Gene	Polyploidy (*z*-score)
ENSG00000114554	PLXNA1	13.15
ENSG00000076356	PLXNA2	4.05
ENSG00000130827	PLXNA3	2.72
ENSG00000164050	PLXNB1	3.14
ENSG00000196576	PLXNB2	3.33
ENSG00000004399	PLXND1	1.45
ENSG00000136040	PLXNC1	0.32

tion, and are primarily found in sorting nexins [37]. Previous studies suggest that various sorting nexins are involved in leukemia [38], colon tumorigenesis [39] and, in general, contribute to cell cycle progression in mammalian cells [40]. However, their role in breast cancer has not been described so far. Our PhenoFam anaysis suggests that proteins with PX domains are frequently misregulated in breast cancer. Hence, we propose that these proteins should be investigated for a possible role in breast cancer development.

Conclusions

PhenoFam is a computational tool designed to analyse experimental results by integration of functional and structural information about protein families. The distinct features of our application include a user-friendly interface and a broad range of supported genomes and identifiers. It should also be noted that our algorithm, in contrast to existing software, treats the InterPro repository as a collection of partially redundant databases, which improves the power of our testing procedure. Using a specific example, we show that the application can be used as an additional hit selection tool for functional screens. Typical hit selection procedures (*i.e. z*-score or quantile-based normalization) apply thresholds that can be passed only by genes showing the strongest phenotypes, which often leads to a high false-negatives rate. In case of our GSEA method, a domain may appear to be significantly enriched despite moderate phenotypes of the associated genes. From the potential relationship between the domain and the investigated biological process, genes with moderate phenotypic scores are considered in the list of hits selected from the screen, thereby reducing the false-negative rate.

We also demonstrated that PhenoFam can help forming novel hypothesis based on gene expression data. Accordingly, PhenoFam should be useful in analysing results of other high-throughput experiments, such as ChiP-Chip/ChiP-Seq and comparative mass-spectrometry. In summary, together with other enrichment analysis tools, PhenoFam can assist in annotating genes of unknown function and in discovering new functions of already characterised genes.

Availability and requirements

Project name: PhenoFam

Project homepage: http://www.phenofam.org/

Operating system(s): Platform independent (web-based application)

Programming language: Java

Other requirements: A web browser with JavaScript support

License: GNU GPLv3 http://www.gnu.org/licenses/gpl-3.0.html

Any restrictions to use by non-academics: None

Additional material

Additional file 1 PhenoFam User's Guide. The file contains a PDF version of the User's Guide provided on the PhenoFam home page.

Authors' contributions

MPR implemented the application and drafted the manuscript. MS carried out the experiments and contributed in writing the manuscript. MTP contributed in designing the application and in writing the manuscript. FB coordinated the study and contributed in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Max Planck Society, the Bundesministerium für Bildung und Forschung grant NGFN-Plus (01GS0859), the Deutsche Forschungsgemeinschaft (SFB 655 'Cells into tissues') and by the Klaus Tschira Stiftung gGmbH.

Author Details

[1]Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307 Dresden, Germany and [2]Structural Bioinformatics, BIOTEC TU Dresden, Tatzberg 47-51, 01307 Dresden, Germany

Received: 23 February 2010 Accepted: 17 May 2010
Published: 17 May 2010

References

1. Khatri P, Draghici S, Ostermeier GC, Krawetz SA: **Profiling gene expression using onto-express.** *Genomics* 2002, **79(2)**:266-70.
2. Dennis G, Sherman BT, Hosack Da, Yang J, Gao W, Lane HC, Lempicki Ra: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome biology* 2003, **4(5)**:P3.

3. Al-Shahrour F, Minguez P, Tarraga J, Medina I, Alloza E, Montaner D, Dopazo J: **FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments.** *Nucleic acids research* 2007:W91-6.
4. Sartor MA, Mahavisno V, Keshamouni VG, Cavalcoli J, Wright Z, Karnovsky A, Kuick R, Jagadish HV, Mirel B, Weymouth T, Athey B, Omenn GS: **ConceptGen: a gene set enrichment and gene set relation mapping tool.** *Bioinformatics (Oxford, England)* 2010, **26(4)**:456-63.
5. Huang daW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 2009, **37**:1-13.
6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.
7. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic acids research* 2010:D355-60.
8. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC: **PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.** *Nat Genet* 2003, **34**:267-273.
9. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102(43)**:15545-15550.
10. Keller A, Backes C, Al-Awadhi M, Gerasch A, Küntzer J, Kohlbacher O, Kaufmann M, Lenhof HP: **GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments.** *BMC bioinformatics* 2008, **9**:552.
11. Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF, Resendez-Perez D, Affolter M, Otting G, Wöthrich K: **Homeodomain-DNA recognition.** *Cell* 1994, **78**:211-223.
12. Hahne F, Mehrle A, Arlt D, Poustka A, Wiemann S, Beissbarth T: **Extending pathways based on gene lists using InterPro domain signatures.** *BMC bioinformatics* 2008, **9**:3.
13. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJA, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C: **InterPro: the integrative protein signature database.** *Nucleic acids research* 2009:D211-5.
14. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJA: **The 20 years of PROSITE.** *Nucleic acids research* 2008:D245-9.
15. Lima T, Auchincloss AH, Coudert E, Keller G, Michoud K, Rivoire C, Bulliard V, de Castro E, Lachaize C, Baratin D, Phan I, Bougueleret L, Bairoch A: **HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot.** *Nucleic acids research* 2009:D471-8.
16. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2008, **36**:D281-288.
17. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C: **PRINTS and its automatic supplement, prePRINTS.** *Nucleic acids research* 2003, **31**:400-2.
18. Corpet F, Servant F, Gouzy J, Kahn D: **ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons.** *Nucleic acids research* 2000, **28**:267-9.
19. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: **Recent improvements to the SMART domain-based sequence annotation resource.** *Nucleic acids research* 2002, **30**:242-4.
20. Haft DH, Selengut JD, White O: **The TIGRFAMs database of protein families.** *Nucleic acids research* 2003, **31**:371-3.
21. Wu CH, Yeh LSL, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, Vinayaka CR, Zhang J, Barker WC: **The Protein Information Resource.** *Nucleic acids research* 2003, **31**:345-7.
22. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *Journal of molecular biology* 2001, **313(4)**:903-19.
23. Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C: **The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis.** *Nucleic acids research* 2005:D247-51.
24. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, Kitano H, Thomas PD: **The PANTHER database of protein families, subfamilies, functions and pathways.** *Nucleic acids research* 2005:D284-8.
25. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P: **Ensembl 2009.** *Nucleic Acids Res* 2009, **37**:D690-697.
26. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function.** *Genome research* 2003, **13(9)**:2129-41.
27. Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, Lazareva-Ulitsky B: **Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools.** *Nucleic Acids Research* 2006:W645-W650.
28. Ben-Shaul Y, Bergman H, Soreq H: **Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression.** *Bioinformatics (Oxford, England)* 2005, **21(7)**:1129-37.
29. Kim SY, Volsky D: **PAGE: parametric analysis of gene set enrichment.** *BMC bioinformatics* 2005, **6**:144.
30. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society Series B-Methodological* 1995, **57**:289-300.
31. Storey JD: **A direct approach to false discovery rates.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002, **64(3)**:479-498.
32. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100(16)**:9440-5.
33. Hernstein RJ, Loveland DH, Cable C: **Natural concepts in pigeons.** *J Exp Psychol Anim Behav Process* 1976, **2**:285-302.
34. Kittler R, Pelletier L, Heninger AK, Slabicki M, Theis M, Miroslaw L, Poser I, Lawo S, Grabner H, Kozak K, Wagner J, Surendranath V, Richter C, Bowen W, Jackson AL, Habermann B, Hyman AA, Buchholz F: **Genome-scale RNAi profiling of cell division in human tissue culture cells.** *Nat Cell Biol* 2007, **9**:1401-1412.
35. Cheng ASL, Culhane AC, Chan MWY, Venkataramu CR, Ehrich M, Nasir A, Rodriguez BAT, Liu J, Yan PS, Quackenbush J, Nephew KP, Yeatman TJ, Huang THM: **Epithelial progeny of estrogen-exposed breast progenitor cells display a cancer-like methylome.** *Cancer research* 2008, **68(6)**:1786-96.
36. Li T, Sparano JA: **Inhibiting Ras signaling in the therapy of breast cancer.** *Clinical breast cancer* 2003, **3(6)**:405-16. discussion 417-20
37. Worby CA, Dixon JE: **Sorting out the cellular functions of sorting nexins.** *Nature reviews. Molecular cell biology* 2002, **3(12)**:919-31.
38. Fuchs U, Rehkamp G, Haas OA, Slany R, Kōnig M, Bojesen S, Bohle RM, Damm-Welk C, Ludwig WD, Harbott J, Borkhardt A: **The human formin-binding protein 17 (FBP17) interacts with sorting nexin, SNX2, and is an MLL-fusion partner in acute myelogeneous leukemia.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98(15)**:8756-61.
39. Nguyen LN, Holdren MS, Nguyen AP, Furuya MH, Bianchini M, Levy E, Mordoh J, Liu A, Guncay GD, Campbell JS, Parks WT: **Sorting nexin 1**

down-regulation promotes colon tumorigenesis. *Clinical cancer research: an official journal of the American Association for Cancer Research* 2006, **12(23):**6952-9.

40. Fuster JJ, González JM, Edo MD, Viana R, Boya P, Cervera J, Verges M, Rivera J, Andrés V: **Tumor suppressor p27Kip1 undergoes endolysosomal degradation through its interaction with sorting nexin 6.** *The FASEB journal: official publication of the Federation of American Societies for Experimental Biology* 2010. E-pub

doi: 10.1186/1471-2105-11-254
Cite this article as: Paszkowski-Rogacz *et al.*, PhenoFam-gene set enrichment analysis through protein structural information *BMC Bioinformatics* 2010, **11**:254