JACOBS PUBLISHERS

OPEN ACCESS

Jacobs Journal of Bioinformatics and Proteomics

Short Research Article

Comparison of Cluego and Impala for Integrated Pathway Enrichment Analysis

Akshay Bhat[1,2], Vera Jankowski[3], Antonia Vlahou[4], Harald Mischak[2,5], Jerome Zoidakis[4*]

[1]*Charité-Universitätsmedizin Berlin, Med. Klinik IV, Berlin, Germany*

[2]*Mosaiques diagnostics GmbH, Hannover, Germany*

[3]*Institute for Molecular Cardiovascular Research (IMCAR), Aachen, Germany*

[4]*Biomedical Research Foundation Academy of Athens, Biotechnology Division, Athens, Greece*

[5]*BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK*

**Corresponding author: Dr. Jerome Zoidakis PhD, Biomedical Research Foundation, Academy of Athens, Department of Biotechnology, Soranou Efessiou 4, 11527 Athens, Greece, Tel: 30-210-6597485; Email: izoidakis@bioacademy.gr*

Received: 06-22-2015

Accepted: 07-21-2015

Published: 01-04-2016

Abstract

Background

High-throughput experimental technologies ranging from genomic sequencing and gene/protein profiling are now commonly being used for the molecular characterization of diseases. These techniques produce large datasets of differentially expressed features that include genes, mRNAs, proteins and metabolites. The abundant data defy straightforward intuitive interpretation. Hence the correlation of molecular features to biological pathways may ultimately help in understanding the patho-physiology of a disease. Many available computational tools allow annotating such integrated datasets at the pathway level. Two prominent tools are ClueGO (Cytoscape plug-in) and ImPAla (web based application). Both tools provide advantages in integrating different pathway databases. However, each tool abides by specific statistical and mathematical algorithms in enriching molecular features onto pathway-centric networks.

Materials and Methods

Bladder cancer (BC) specific molecular features were retrieved from literature and omics profiles. The data comprise of differentially expressed DNA-mutations, DNA-methylation, mRNAs, miRNAs, proteins and metabolites. These features were combined and subjected to protein-protein interactions to yield the BC interactome. The features from this interactome were used as the input-list for the pathway enrichment analysis.

Results

292 pathways were obtained from ClueGO and 471 pathways from ImPAla. The resulting pathways were selected according to

the following significance criterion: multiple comparison corrected p-value <0.05. Comparison of the results obtained by the two applications yielded 152 pathway terms with exactly the same name. Moreover, 137 ClueGO pathway terms were similar to 251 ImPAla pathways. Thus, the overall overlap between the two datasets is 289 ClueGO pathways corresponding to 403 ImPAla pathways. ClueGO yielded 3 unique pathway terms whereas in the case of ImPAla 68 unique pathways were obtained. Both datasets contain redundant terms but the ImPAla results are characterized by higher redundancy. In addition, ImPAla yields 12 unique pathways that are not related to BC.

Conclusion

Cytoscape-ClueGO has better performance than ImPAla in pathway enrichment analysis since the output is less redundant and contains all the biologically significant information.

Keywords: Bioinformatics; Systems Biology; Cluego; Impala; Data Integration; Pathway Enrichment

Introduction

Next-generation sequencing and profiling techniques ranging from genomics to transcriptomics, proteomics and metabolomics have transformed biological research by allowing a comprehensive monitoring of biological systems [1]. These technologies yield a vast amount of data, typically as a list of differentially expressed proteins, genes, transcripts, miRNA and metabolites that may have specific roles in a given clinical phenotype [2]. However, these lists of individual features fail in providing a mechanistic insight for the molecular characterization of a disease [3]. Hence, these challenges have led to an advent of new functional annotation approaches in which individual features are grouped together into pathways by statistical or mathematical algorithms [2]. An important advantage of working with molecular pathways rather than individual proteins or genes is the fact that it is often easier and more relevant to predict the function of a module than a function of an individual protein/gene [3]. Prediction of a functional module is possible only if the pathway contains a sufficient number of features known to be associated to that pathway. Such functional module prediction is also known as enrichment analysis; as it builds on the assumption that features could be assigned to a particular pathway or process, grouped and organized in Gene Ontologies (GO) [4]. GO are sub-categorized into cellular component, biological process and molecular pathway. Enrichment analysis determines whether the number of features attributed to a specific pathway is higher than expected by chance. This can be calculated using statistical methods such as $\chi 2$, hypergeometric tests and Fisher's exact tests and have been implemented in many software packages, like R-Project (http://www.r-project.org/). Some of the frequently used software packages and applications for performing enrichment analysis are publically available. Some of the packages include ClueGO [5], BinGO [6], Gorilla [7]a variety of tools have been developed that support exploring and searching the GO database. In particular, a variety of tools that perform GO enrichment analysis are currently available. Most of these tools require as input a target set of genes and a background set and seek enrichment in the target set compared to the background set. A few tools also exist that support analyzing ranked lists. The latter typically rely on simulations or on union-bound correction for assigning statistical significance to the results.\n\nRESULTS: GOrilla is a web-based application that identifies enriched GO terms in ranked lists of genes, without requiring the user to provide explicit target and background sets. This is particularly useful in many typical cases where genomic data may be naturally represented as a ranked list of genes (e.g. by level of expression or of differential expression, Enrichment Map [8], Metscape [9], InCroMap [10], 3Omics [11]proteomics and metabolomics datasets require an intensive knowledge of tools and background concepts. Thus, it is challenging for users to perform such analyses, highlighting the need for a single tool for such purposes. The 3Omics one-click web tool was developed to visualize and rapidly integrate multiple human inter- or intra-transcriptomic, proteomic, and metabolomic data by combining five commonly used analyses: correlation networking, coexpression, phenotyping, pathway enrichment, and GO (Gene Ontology, iPEAP [12] and ImPAla [13]. The aim of this short communication is to evaluate the performance of pathway enrichment analysis obtained from two bioinformatics tools ImPAla and ClueGO. The two applications were selected because ClueGO has more than 120 citations (compared to other programs) whereas 18 articles cited ImPAla (a significantly higher number of citations compared to other web-based enrichment tools.

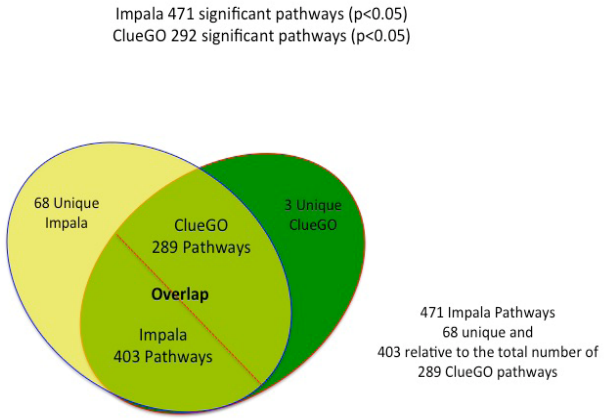
Materials and Methods

The list of bladder cancer associated features that were subjected to protein-protein interactions (PPIs) was obtained from the BcCluster database (www.bccluster.org) [14]. Protein interaction information was acquired by querying PPI databases such as IntAct [15], BioGRID [16], String [17] and Reactome [18]. First, all available PPIs for the human proteome were downloaded into Cytoscape [19] to form the human interactome. The PPIs relevant to the BC-associated proteins were retrieved from this human interactome. Only proteins that had at least one binding partner were retained. This step yielded 435 entries (in official gene-ids) that correspond to proteins from the BC PPI. The list of 435 entries with information regarding the official gene id and gene name, in addition to the differential expression of the feature is provided in Supplementary Table S1. This list was then subjected to pathway enrichment using ClueGO and ImPAla. ClueGO provides an advantage to perform cluster comparisons for pathway enrichment and allows the option to separately input up and down regulated molecules in the software. In addition, ClueGO provides an

optional redundancy reduction feature (Fusion) to assess GO terms that share similar associated features in a parent–child relation. This option was selected in our ClueGO pathway enrichment analysis to eliminate the redundant pathway terms. In contrast, ImPAla does not provide an option of redundancy reduction for pathway terms. The pathway databases selected for enrichment were KEGG [20] and Reactome. The statistical selection criterion taken into account for the enrichment analysis was the corrected for multiple comparisons p-value < 0.05. The overlap assessment between the pathway outputs was performed manually.

Results and Discussion

The overview of the pathway analysis performed by ClueGO and ImPAla is illustrated in Table 1. The total number of KEGG and Reactome pathway terms obtained from ClueGO was 292. ImPAla produced 471 pathways. Additional information for the raw enriched pathway results is also provided in Supplementary Table S2. This information includes number of input genes, corrected p-value scores, Gene Ontology source and the number of genes held in each pathway. By comparing the pathway results, 152 pathway terms exactly overlapped in ClueGO and ImPAla. 137 pathway terms from ClueGO were highly similar to 251 ImPAla pathway terms. Therefore, the total calculated overlap of pathways between the two tools equalled to 289 ClueGO pathways that correspond to 403 ImPAla enriched pathways. In addition, the software programs also produced unique pathway terms. There were 3 unique pathways from the total 292 ClueGO pathway terms whereas 68 pathways were unique from ImPAla. Both the enrichment tools yielded redundancy in the output results, however results from ImPAla were characterized by higher redundancies in pathway terms (for e.g. the pathway terms “DNA replication”, “synthesis of DNA”). Moreover, from the unique set of 68 ImPAla pathway terms, 12 pathway terms were not related in the context of BC. Some of these pathways include alcoholism, amphetamine addiction, inflammatory bowel disease (IBD), malaria, viral myocarditis and prion diseases. On the contrary, the 3 unique pathways obtained by ClueGO were relevant to BC. It was also noted that the overlapping pathway terms from ImPAla and ClueGO contained pathway names that are not relevant in the context of BC. These common terms totalled to 34 ImPAla and 30 ClueGO pathway terms. The common pathway terms included oocyte meiosis, tuberculosis, type II diabetes mellitus, circadian clock and shigellosis. In addition, there were some common terms that were very general in the description, for e.g. “Disease and Developmental biology”. The resulting raw pathway outputs, exactly overlapping pathways, highly similar pathways, unique pathways and unrelated pathways retrieved from the two programs are provided in Supplementary Table S2. In addition, the comparison of significant overlapping pathways obtained from ClueGO and ImPAla is represented as a Venn diagram in Figure 1.



Software	Availability	User input	p-value correction method	Total pathway output	Reference
ClueGO	Cytoscape plugin	435 entries	Bonferroni	292	[5]
ImPAla	Web-based	435 entries	Benjamini Hochberg	471	[13]

Table 1. General information for the results obtained from the pathway enrichment analysis.

In conclusion, output by both software tools provided a significant set of pathways for the enrichment analysis. However, ImPAla produced a significantly higher number of pathways than ClueGO. This is due to the fact that ImPAla does not have a pathway term redundancy reduction feature. Hence this could be the reason for the high redundancy observed in the ImPAla pathway term list (many common pathways with different names). For e.g. the pathway term Chagas disease (American trypanosomiasis) was retrieved from ClueGO, whereas Chagas disease (American trypanosomiasis) and African trypanosomiasis, were retrieved from ImPAla. In our previous publication [14], we had filtered 292 ClueGO pathways to make them non-redundant for the database storage. This non-redundant pathway list equals 90 BC specific pathways. ImPAla retrieves more redundant pathways than ClueGO. Thus, the effort to manually eliminate redundant terms from the 471 ImPAla derived pathways is significantly higher. In addition, the 68 unique pathways retrieved by ImPAla contain 12 pathways not related to BC. Both pathway enrichment tools allow the input of regulation information as numerical values (fold change) and p-values. However, ClueGO has the additional feature of allowing text input for regulation. Since our data set contained regulation information in the form of text (up/down) (Table S1), ClueGO was able to incorporate this feature in the pathway enrichment analysis. It should be noted that if the user in ImPAla does not provide numerical regulation values, the

tool considers the protein/gene as differentially expressed but does not assign a specific trend (up or down). ClueGO provides additional options that allow the user to define the stringency of pathway selection. The options include: Kappa statistics in order to generate pathway network visualizations, mid-p-values and doubling p-values in order to retrieve significant pathways based on user-defined threshold p-values, setting specific limits for ratios of differentially expressed genes relative to the total number of genes present in a pathway in order to consider a pathway as significant. In contrast, ImPAla does not provide these features. Moreover, ClueGO provides a more descriptive data output that contains significant additional information when compared to ImPAla data output. The common columns shared among the two enrichment tools include pathway name, pathway source, number of input genes present in the pathway, total number of genes present in the pathway and corrected p-values. Additional columns provided by ClueGO were genes down-regulated in input set, genes up-regulated in input set, Gene Ontology ID, and percentage input Genes present in pathway. In addition, ClueGO is more user-friendly when compared to ImPAla since it offers better help options.

In the study by Jaakkola MK et al. the performance of six enrichment tools were tested on experimental datasets from six renal-cell carcinoma and four type-1 diabetes samples. The software programs tested included, SPIA, CePa, DAVID, NetGSA, GSEA and Pathifier. From the resulting enrichment outputs, the authors noted that significant pathways were different according to different enrichment methods, and the number of significant findings depended on the enrichment method. Hence, they conclude that the selection of the enrichment method had a large impact on the pathway output results [21]. Our attempt was to compare the performance for significant pathway outputs yielded from ClueGO and ImPAla in the context to bladder cancer. It could be stated that ImPAla can provide advantages over ClueGO by integrating many more pathway database resources for the comprehensiveness in pathway information. However, we only selected two widely used and up-to-date pathway database resources, KEGG and Reactome. Adding more databases in our analysis would introduce higher redundancy in pathway outputs. In regard to manually updating database sources, the ClueGO application allows users to update individual pathway database source within Cytoscape in order to obtain latest data whereas ImPAla is an omics-integration (with a focus towards metabolomics integration) and pathway enrichment application that contains the latest update of January 2015. In addition, ImPAla also allows the incorporation of differential expression information for molecules such as fold change and p-values. Nevertheless, ImPAla does not offer the option to input separately up and down regulated genes and does not make predictions on the activation/deactivation of an affected pathway in contrast to ClueGO. In addition, Cytoscape provides various plug-ins for analysing different omics datasets such as genes, mRNAs, proteins, SNPs, metabolites and miRNAs. This gives the advantage in using one analysis and visualization tool for all high-throughput sequencing and profiling experiments. Furthermore, having a single analysis tool also helps to prevent errors due to compatibility when transferring data between different software applications. Therefore, we conclude that Cytoscape-ClueGO is preferable to ImPAla for pathway enrichment and in the comprehensive characterization of molecular diseases.

Acknowledgement

HM, VJ, AV and JZ designed and coordinated the study, and AB performed data retrieval, the analysis and drafted the manuscript. All authors contributed to the interpretation of the results and drafted the publication along with reading and approving the final manuscript. This work was supported in part by the BCMolMed grant PITN-GA-2012-317450. BCMolMed is funded by the European Commission.

References

1. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. 2012, 8(2): e1002375.

2. Bhat A, Heinzel A, Mayer B, Perco P, Mühlberger I et al. Protein interactome of muscle invasive bladder cancer. PLoS One. 2015, 10(1): e0116404.

3. Bhat A, Dakna M, Mischak H. Integrating proteomics profiling data sets: a network perspective. Methods Mol Biol. 2015, 1243: 237-253.

4. Blake JA, Dolan M, Drabkin H, Hill DP, Li N et al. Gene Ontology annotations and resources. Nucleic Acids Res. 2013, 41: D530-535.

5. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009, 25(8): 1091-1093.

6. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005, 21(16): 3448-3449.

7. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009, 10: 48.

8. Isserlin R, Merico D, Voisin V, Bader GD. Enrichment Map - a Cytoscape app to visualize and explore OMICs pathway enrichment results. F1000Research. 2014, 3: 141.

9. Karnovsky A, Weymouth T, Hull T, Tarcea VG, Scardoni G et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics. 2012, 28(3): 373-380.

10. Eichner J, Rosenbaum L, Wrzodek C, Häring H-U, Zell A et al. Integrated enrichment analysis and pathway-centered visualization of metabolomics, proteomics, transcriptomics, and genomics data by using the InCroMAP software. J Chromatogr B Analyt Technol Biomed Life Sci. 2014, 966: 77-82.

11. Kuo T-C, Tian T-F, Tseng YJ. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Syst Biol. 2013, 7: 64.

12. Sun H, Wang H, Zhu R, Tang K, Gong Q et al. iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. Bioinformatics. 2014, 30(5): 737-739.

13. Kamburov A, Cavill R, Ebbels TMD, Herwig R, Keun HC. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics. 2011, 27(20): 2917-2918.

14. Bhat A, Mokou M, Zoidakis J, Jankowski V, Vlahou A, Mischak H. BcCluster: A bladder cancer database at the molecular level. Bl. Cancer. 2015.

15. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2012, 40: D841-846.

16. Stark C, Breitkreutz B-J, Chatr-Aryamontri A, Boucher L, Oughtred R et al. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 2011, 39: D698-704.

17. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013, 41: D808-815.

18. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011, 39: D691-697.

19. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003, 13(11): 2498-2504.

20. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004, 32: D277-280.

21. Jaakkola MK, Elo LL. Empirical comparison of structure-based pathway methods. Brief Bioinform. 2015, bbv049.