Bioinformatics, 37(21), 2021, 3981–3982
doi: 10.1093/bioinformatics/btab559
Advance Access Publication Date: 6 August 2021
Applications Note

Data and text mining

LipiDisease: associate lipids to diseases using literature mining

Piyush More 1,2,*, Laura Bindila[3], Philipp Wild[4], Miguel Andrade-Navarro[2] and Jean-Fred Fontaine[2]

[1]Department of Pharmacology, University Medical Center, 55131 Mainz, Germany, [2]Faculty of Biology, Johannes Gutenberg University of Mainz, 55128 Mainz, Germany, [3]Clinical Lipidomics Unit, Institute of Physiological Chemistry, University Medical Center, 55131 Mainz, Germany and [4]Center for Thrombosis and Hemostasis (CTH), University Medical Center, 55131 Mainz, Germany

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on April 1, 2021; revised on July 10, 2021; editorial decision on July 26, 2021

Abstract

Summary: Lipids exhibit an essential role in cellular assembly and signaling. Dysregulation of these functions has been linked with many complications including obesity, diabetes, metabolic disorders, cancer and more. Investigating lipid profiles in such conditions can provide insights into cellular functions and possible interventions. Hence the field of lipidomics is expanding in recent years. Even though the role of individual lipids in diseases has been investigated, there is no resource to perform disease enrichment analysis considering the cumulative association of a lipid set. To address this, we have implemented the LipiDisease web server. The tool analyzes millions of records from the PubMed biomedical literature database discussing lipids and diseases, predicts their association and ranks them according to false discovery rates generated by random simulations. The tool takes into account 4270 diseases and 4798 lipids. Since the tool extracts the information from PubMed records, the number of diseases and lipids will be expanded over time as the biomedical literature grows.

Availability and implementation: The LipiDisease webserver can be freely accessed at http://cbdm-01.zdv.uni-mainz.de:3838/piyusmor/LipiDisease/.

Contact: piyusmor@uni-mainz.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Lipids are a diverse group of organic compounds having characteristic solubilities in nonpolar solvents. They exhibit essential roles in cellular assembly, signaling and energy storage. Over the past few years, there has been an increasing interest in untargeted lipidomics, largely driven by the technological advancements in lipid profiling and functional analysis improving, thus, our understanding of lipid biology. The lipidomic analysis encompasses lipid quantification and identification, allowing both insights into changes into the structural heterogeneity and into individual lipid levels in a biospecimen. It is evident that the lipidome is altered during disease development and comparative lipidomics could identify lipid biomarkers for disease prognosis or even intervention (Ghosh and Nishtala, 2017; Lydic and Goo, 2018). For this reason, it is important to develop methods to identify lipids that could be predictive of disease type, progression and risk using the study of lipidomics data.

One of the ways to identify disease-relevant biomarkers is to associate them with diseases using pre-existing literature data. Enrichment analysis is a computational procedure to assist this task in an automated and exhaustive manner. Enrichment analysis is performed to associate a set of biomolecules, for example, derived from a high-throughput experiment, with a biological trait. The analysis compares the association of biomolecules with a background set to identify significantly associated candidates with biological traits. The background association is derived from biological information and is typically obtained from experimental datasets and information databases. There is a plethora of such resources available for genomics and proteomics data that perform pathway enrichment, disease association and more (Alfoldi and Lindblad-Toh, 2013; Conesa *et al.*, 2016; Fontaine and Andrade-Navarro, 2016; Schmidt *et al.*, 2014). Such databases and tools are severely lacking in the case of lipids because of the stronger focus of the omics field on DNA, RNA and protein biology (Stephenson *et al.*, 2017), and consequently, the enrichment analysis of lipids is limited to their categorization and pathway identification (Acevedo *et al.*, 2018; Clair *et al.*, 2019; Martin *et al.*, 2013; Molenaar *et al.*, 2019), while their functional enrichment is lacking. While much information on metabolite association with diseases can be accessed and retrieved from databases such as MarkerDB (Wishart *et al.*, 2021) and Human

Metabolome Database (Wishart *et al.*, 2007), a dedicated database for lipids is missing. LipidPedia is a resource specialized in associating lipids with biomedical information (Kuo and Tseng, 2018). It utilizes full-text mining focused on individual lipids, but lacks statistical filtering and does not allow manual inspection of the underlying basis for disease associations. Furthermore, currently, there is no way to analyze a set of lipids and consider their cumulative association, which is more relevant in human diseases.

The LipiDisease web server aims to address this by providing a web app performing disease enrichment analysis using a set of lipids and ranking them for prioritizing the diseases. It utilizes the PubMed database listing more than 26 million biomedical records with their manual associations with chemicals and diseases. We expect that the web server will contribute to a better understanding of lipid-disease associations and will inspire the development of many such tools expanding our knowledge about lipids beyond membranes and signaling.

2 Implementation

LipiDisease is built around the biomedical literature from the PubMed database. With a focus on avoiding false positive, instead of full-text mining, it utilizes the manual annotations of the PubMed records using the Medical Subject Headings thesaurus MeSH. Disease terms were extracted from the branch ‘C’ of the MeSH. The list of lipids and corresponding PubChem Compound ID numbers (CIDs) were obtained from Lipid Maps (Fahy *et al.*, 2009). Articles corresponding to the lipid PubChem CIDs were obtained from PubMed using NCBI’s Entrez Programming Utilities. The entire data was locally stored in a MySQL database (details in Supplementary Information).

The interactive web app was built using the Shiny package in R (R Core Team, 2020). JavaScript functionalities were implemented using the shinyjs package. The required input for the web server is either a set of PubChem CIDs for lipids or MeSH Unique IDs for diseases depending on the analysis. There are the following four types of analysis:

- Lipid-set enrichment;
- Lipids to Diseases;
- Diseases to Lipids;
- Lipid-set enrichment (with fold changes).

The Lipid-set enrichment option considers a list of lipids. These lipids are considered collectively for disease enrichment analysis. Lipids to Diseases and Diseases to Lipids consider individual lipids and diseases, respectively, and identify associations using individual entries; using either of these two analyses is close to browsing our underlying database of associations. Lipid-set enrichment (with fold changes) is similar to Lipid-set enrichment with the additional consideration of lipid-level statistics (e.g. fold changes for the lipid expression between two biological conditions). To understand the functionality of the web server, users can use the test cases by clicking on the ‘example’ link next to every analysis type.

An analysis is done in seconds using the pre-indexed data. The output is represented in table and plot form, which can be downloaded locally for further exploration. The output table can be downloaded in TSV format (entries separated by TAB) and can be accessed using text or spreadsheet programs. The table provides hyperlinks to corresponding lipid PubChem CIDs, disease MeSH terms and PubMed records.

The predicted enrichment is derived from the manual annotations of the PubMed records with the lipid PubChem CIDs using a computational procedure employed before for literature mining of gene to disease associations (see Supplementary Information for details; Fontaine and Andrade-Navarro, 2016). In short, lipid enrichment is performed by identifying the over-representation of PubMed articles using one-tailed Fisher’s exact test. The results are then ranked according to false discovery rate calculated by Benjamini and Hochberg method (Benjamini and Hochberg, 1995). The type of association (for example, positive or negative) cannot be derived from this. However, the links from the results to the underlying data that was used to derive the associations, particularly the connections to the relevant literature through PubMed records, facilitate post-analysis of the results.

3 Conclusion

An online web server, LipiDisease, was constructed to predict significant associations between lipids and diseases using literature mining. LipiDisease is the first tool performing disease enrichment using a set of lipids and providing a ranked list of statistically significant associations. The tool provides a user-friendly web interface and outputs results in tabular as well as graphical form, with links to the underlying data used in the literature mining procedure, making it easier to generate hypotheses about the involvement of lipids in disease.

Acknowledgements

The authors gratefully acknowledge the help from the IT group at the Johannes Gutenberg University Mainz. They thank Prof. Leszek Wojnowski from the Department of Pharmacology, University Medical Center Mainz for his insightful comments.

Funding

This work was supported by funding from the Bundesministerium für Bildung und Forschung (BMBF), Germany, grant program Forschungskernen für Massenspektrometrie in der Systemmedizin (MSCoreSys) to project Data-Independent Acquisition-based Systems Medicine: Mass spectrometry for high-throughput deep phenotyping of the heart failure syndrome (DIASyM) to M.A. (funding number: 031L0217B).

Conflict of Interest: none declared.

References

Acevedo,A. *et al.* (2018) LIPEA: lipid pathway enrichment analysis bioinformatics. bioRxiv, 274969.

Alfoldi,J. and Lindblad-Toh,K. (2013) Comparative genomics as a tool to understand evolution and disease. *Genome Res.*, **23**, 1063–1068.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.*, **57**, 289–300.

Clair,G. *et al.* (2019) Lipid Mini-On: mining and ontology tool for enrichment analysis of lipidomic data. *Bioinformatics*, **35**, 4507–4508.

Conesa,A. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.

Fahy,E. *et al.* (2009) Update of the LIPID MAPS comprehensive classification system for lipids. *J. Lipid Res.*, **50**, S9–14.

Fontaine,J. and Andrade-Navarro,M. (2016) Gene Set to Diseases (GS2D): disease enrichment analysis on human gene sets with literature data. *Genomics Comput. Biol.*, **2**, e33.

Ghosh,A. and Nishtala,K. (2017) Biofluid lipidome: a source for potential diagnostic biomarkers. *Clin. Transl. Med.*, **6**, 22.

Kuo,T.-C. and Tseng,Y.J. (2018) LipidPedia: a comprehensive lipid knowledgebase. *Bioinformatics*, **34**, 2982–2987.

Lydic,T.A. and Goo,Y. (2018) Lipidomics unveils the complexity of the lipidome in metabolic diseases. *Clin. Transl. Med*, **7**, 4.

Martin,S.S. *et al.* (2013) Very large database of lipids: rationale and design. *Clin. Cardiol.*, **36**, 641–648.

Molenaar,M.R. *et al.* (2019) LION/web: a web-based ontology enrichment tool for lipidomic data analysis. *GigaScience*, **8**, giz061.

R Core Team. (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schmidt,A. *et al.* (2014) Bioinformatic analysis of proteomics data. *BMC Syst. Biol.*, **8**, S3.

Stephenson,D.J. *et al.* (2017) Lipidomics in translational research and the clinical significance of lipid-based biomarkers. *Transl. Res. J. Lab. Clin. Med.*, **189**, 13–29.

Wishart,D.S. *et al.* (2007) HMDB: the human metabolome database. *Nucleic Acids Res.*, **35**, D521–D526.

Wishart,D.S. *et al.* (2021) MarkerDB: an online database of molecular biomarkers. *Nucleic Acids Res.*, **49**, D1259–D1267.