LipiDisease: associate lipids to diseases using literature mining

Piyush More[1,2,*], Laura Bindila[3], Philipp Wild[4], Miguel Andrade-Navarro[2], 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
[4]Center for Thrombosis and Hemostasis (CTH), University Medical Center, 55131 Mainz, Germany

[*] Author to whom correspondence should be addressed

Supplementary Information

Computational methods

The data retrieved from PubChem and PubMed was filtered to limit PubMed records corresponding to at least one disease and one lipid. Among these, the significant associations of diseases and lipids were identified by performing one-tailed Fisher's exact test as described before (Fontaine and Andrade-Navarro, 2016). In short, articles associated with diseases as well as articles associated with lipids were represented in a 2-by-2 contingency matrix to identify the over-representation of the articles associated with both diseases and lipids. All computations were limited to the resulting 709,038 PubMed records associated with 4,488 diseases and 2,771 lipids.

To evaluate the enrichment of a set of lipids to a particular disease, one-tailed Fisher's exact test was performed by creating the contingency matrix depicted in Table S1.

Table S1. The contingency matrix used for performing a one-tailed Fisher's exact test. *S* represents a lipid set under investigation and *D* represents a disease. The number of lipids in the groups *a*, *b*, *c*, and *d* were used to identify the over-representation.

Lipids	From the *S*	Not from the *S*
Associated with the *D*	*a*	*b*
Not associated with the *D*	*c*	*c*

The resulting p-value was corrected for multiple testing by calculating the false discovery rate (FDR) by Benjamini and Hochberg method (Benjamini and Hochberg, 1995) using the R statistical environment (R Core Team, 2020).

Lipid-level statistics were also considered for disease enrichment. A ranked list of lipids was used for disease enrichment using the fast gene-set enrichment algorithm. This was implemented using the fgsea package in R (Korotkevich *et al.*, 2019).

Database contents

The current version of LipiDisease consists of a total of 709,038 PubMed records associated with 4,488 diseases and 2,771 lipids. Sterol lipids, polyketides, and fatty acyls are the most represented categories by those articles (Fig. S1). Among diseases, 23% of the PubMed articles discuss neoplasms with lipids. Figure S1 shows the top 10 diseases associated with any lipids as described in the literature.

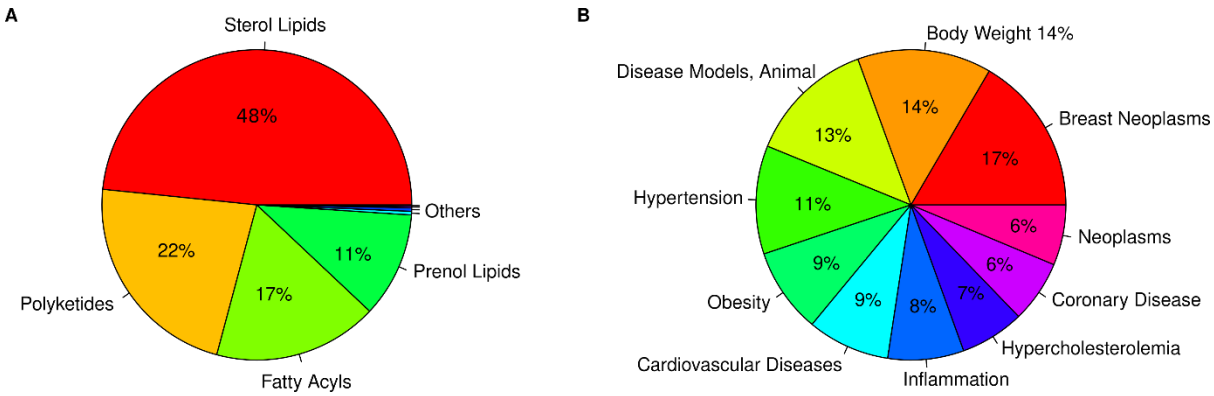

Fig. S1. Percentage of PubMed articles annotated with (**A**) different lipid classes and (**B**) diseases.

Comparison with LipidPedia

LipidPedia is one of the few tools specialized in associating lipids with biomedical information (Kuo and Tseng, 2018). It performs full-text mining to extract lipid-relevant information including the biomedical associations. However, it lacks statistical filtering and only provides diseases associated with individual lipids. Hence, we obtained lipids associated with coronary disease, type II diabetes mellitus, dyslipidemias, heart failure, and hyperlipidemias from LipidPedia and compared with the results of "Diseases to Lipids" analysis from LipiDisease performed at the default settings (minimum 5 citations associating a disease with a lipid, and 0.05 FDR). In general, LipidPedia reported high number of lipids associated with diseases compared to LipiDisease (Table S2). This could be because LipidPedia performs full-text mining to find these associations, while LipiDisease only considers MesH-curated citations. Furthermore, LipidPedia lacks statistical filtering making LipiDisease results more stringent. The output of our tool LipiDisease facilitates manual lookup of the PubMed articles from which associations are derived (Figure S2). This is not possible with LipidPedia.

Table S2. Comparison between LipiDisease and LipidPedia. The number of lipids associated with selected diseases obtained from both resources and overlap between them.

Disease	LipiDisease	LipidPedia	Common
Coronary disease	35	70	12
Diabetes Mellitus, type II	60	87	25
Dyslipidemias	13	7	2
Heart failure	21	36	4
Hyperlipidemias	26	73	10

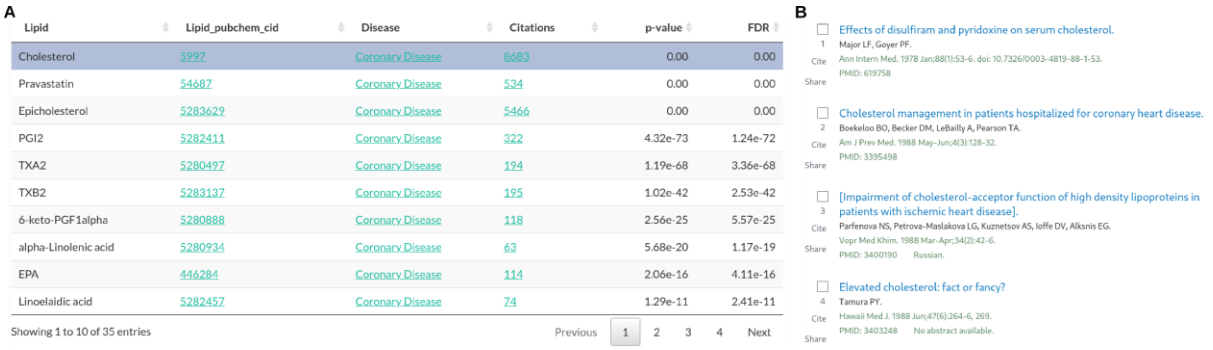

Fig. S2. Output of the "Diseases to Lipids" analysis from LipiDisease. (**A**) top 10 lipids associated with Coronary Disease and (**B**) articles from PubMed associating the lipid Cholesterol with Coronary Disease.

References

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.

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.

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

Korotkevich,G. *et al.* (2019) Fast gene set enrichment analysis. *BioRxiv*.

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