MetaFun: Unveiling sex differences in multiple omics studies through comprehensive functional meta-analysis

Pablo Malmierca-Merlo[1], Rubén Sánchez-Garcia[1], Rubén Grillo-Risco[1], Irene Pérez-Díez[2], José F. Català-Senent[1], María de la Iglesia-Vayá[2], Marta R. Hidalgo[1*], Francisco Garcia-Garcia[1*]

[1]Bioinformatics and Biostatistics Unit, Principe Felipe Research Center (CIPF), 46012 Valencia, Spain.
[2]Biomedical Imaging Unit FISABIO-CIPF, Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana, 46012 Valencia, Spain.

*To whom correspondence should be addressed.

Abstract

Summary: Sex and gender differences in different health scenarios have been thoroughly acknowledged in the literature, and yet, very scarcely analyzed. To fill the gap, here we present MetaFun, a web-based tool to meta-analyze multiple omics datasets with a sex-based perspective, and to combine different datasets to gain major statistical power and to assist the researcher in understanding these sex differences in the diseases under study. Metafun is freely available at bioinfo.cipf.es/metafun

Availability and implementation: MetaFun is available under http://bioinfo.cipf.es/metafun. The backend has been implemented in R and Java and the frontend has been developed using Angular.

Contact: fgarcia@cipf.es, mhidalgo@cipf.es

Supplementary information: R code available at https://gitlab.com/ubb-cipf/metafunr

1 Introduction

The existence of sex and gender differences in different health scenarios has been thoroughly acknowledged in the literature [1,2], and yet, in many cases, not exhaustively analyzed. Many times, the importance of such differences has been neglected, when not denied, and the sex variable has not been taken into account in the experimental design of studies, leading in the extreme to experiments with samples of just one sex. As a result, most of the underlying reasons for such differences have not been yet established.

Fortunately, in the past few years the scientific community has made great efforts to improve

this situation, and researchers are beginning to include a sex/gender perspective in their scientific approaches. However, there is still a vast amount of generated data stored in public databases (such as GEO [3] or GDC [4]) which has not been analyzed with this perspective. The information in these databases is a powerful tool which should not be wasted.

When exploiting these resources with a particular objective, we often find multiple studies trying to answer similar questions, sometimes with different and even contradictory results. The question of which one to trust has no optimal answer, and a solution might be to integrate all different datasets in the same analysis. Defined for this purpose, the meta-analysis is a statistical methodology which takes into account the relative importance of different studies in order to combine them in a single analysis and extract results based on more evidence and samples [5,6,7]. However, the application of advanced statistical techniques as the meta-analysis is often out of reach for biomedical researchers aiming to analyze their data in an easy way.

MetaFun has been designed to simplify the process and facilitate the application of the meta-analysis to researchers working with omics data which may not be familiar with it, allowing to meta-analyze functionally multiple omics datasets with or without a sex/gender perspective, and to combine them to gain major statistical power and soundness. MetaFun is a complete suite which allows analyzing transcriptomics data and exploring its results at all levels, performing single-dataset exploratory analysis, differential expression, pathway analysis, gene set functional enrichment and, finally, combining functional results in a functional meta-analysis.

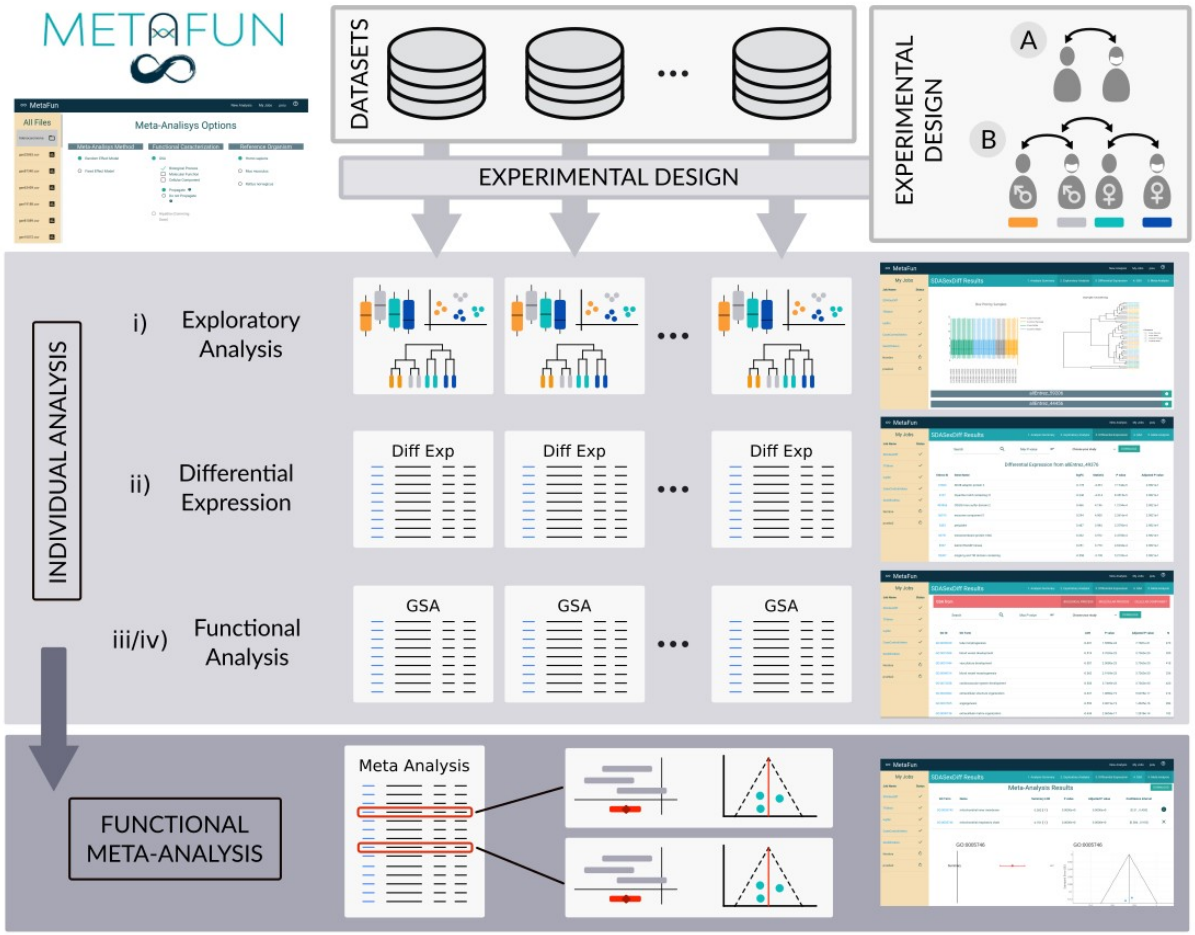


Figure 1: Metafun pipeline. First, datasets and experimental designs must be uploaded as CSV and TSV files, respectively. Available comparisons include (A) a classical *Case vs. Control* comparison, and (B) the sex-perspective comparison (*Female case* vs. *Female control*) vs. (*Male case* vs. *Male control*). Then, single-experiment analysis including i) an exploratory analysis, ii) a gene differential expression and iii) a functional analysis are performed on each dataset. Finally, functional results are integrated into a functional meta-analysis. The tool allows the user to explore all results generated in the process.

2 Methods

Metafun is available under https://bioinfo.cipf.es/metafun. Help may be found under https://gitlab.com/metafundev/metafun/-/wikis/MetaFun-Help.

2.1 Input data and experimental design

MetaFun takes as input a set of at least 2 CSV expression files and 2 TSV experimental design files. CSV expression files must include already normalized transcriptomics data which must come from comparable studies with assimilable experimental groups. Columns must be the samples in the study, and rows must be the analyzed genes as ENTREZ_ID. The first row will be the names of the samples. TSV experimental files must define the class to which each sample of the study belongs, by including two columns: the names of the samples and the class to which they belong. Accepted reference organisms are, for the

moment, human (*Homo sapiens*), mouse (*Mus musculus*) and rat (*Rattus norvegicus*). The analysis will be made with respect to a comparison which must be applicable to all datasets. Options are the classical comparison *Case* vs. *Control* (Fig.1A), or the sex-perpective comparison (*Male case* vs. *Male control*) vs. (*Female case* vs. *Female control*) (Fig.1B), in which the effect under study is compared between sexes.

2.2 Single-dataset analyses

After the selection of the studies and the experimental design, MetaFun analyzes each dataset separately with an individual analysis consisting of: i) an exploratory analysis including boxplots, PCA and cluster plots using *plotly* library [8], ii) a gene differential expression analysis, using *limma* package [9], and iii) a Gene Set Enrichment Analysis (GSEA) [10] based on Gene Ontology (GO) [11], from *mdgsa* package [12]. Figures and tables resulting from these analyses may be explored and downloaded from the Results area once the job is ready. Links to NCBI and QuickGO databases are present to go into detail about the results.

2.3 Functional meta-analysis

After the single-set analyses, MetaFun combines the gene set functional enrichments of all datasets in a meta-analysis with the same experimental design, using the *metafor* package [13]. Forest and funnel plots are generated by means of the *plot.lyJS* library [8]. Figures and tables resulting from this meta-analysis are interactive and may be explored and downloaded from the Results area once the job is ready.

2.4 Implementation

MetaFun back-end has been written using Java and R, and is supported by a non relational database (*MongoDB* [14]) which stores the files, users and jobs information. The front-end has been developed using the Angular Framework [15]. All graphics generated in this webtool have been implemented with *Plot.ly* [8] except for the exploratory analysis cluster plot which uses *ggplot2* package [16].

2.5 Study Cases

MetaFun includes as example three sets of pre-selected study cases, one for each accepted species: human, mouse and rat. The study cases can be executed directly from the webtool and allow to explore the functionalities of the tool easily. Human study case includes 9 studies from lung cancer patients [6].

Acknowledgements

The authors thank the Principe Felipe Research Center (CIPF) for providing access to the cluster, which is co-funded by European Regional Development Funds (FEDER) in Valencian Community 2014–2020.

Funding

This work was supported by GV/2020/186, and partially funded by the Institute of Health

Carlos III (project IMPaCT-Data, exp. IMP/00019), co-funded by the European Union, European Regional Development Fund (ERDF, “A way to make Europe”).
Conflict of Interest: none declared.

References

1. Ballestri S. et al. NAFLD as a Sexual Dimorphic Disease: Role of Gender and Reproductive Status in the Development and Progression of Nonalcoholic Fatty Liver Disease and Inherent Cardiovascular Risk. Advances in Therapy. 2017; 34: 1291–1326. URL http://link.springer.com/10.1007/s12325-017-0556-1
2. Yuan Y, Liu L, Chen H, et al. Comprehensive Characterization of Molecular Differences in Cancer between Male and Female Patients. Cancer cell. 2016; 29(5): 711-722. doi:10.1016/j.ccell.2016.04.001
3. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5
4. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine. 2016; 375:12, 1109-1112. doi: 10.1056/NEJMp1607591
5. Casanova Ferrer F, Pascual M, Hidalgo MR, Malmierca-Merlo P, Guerri C, García-García F. Unveiling Sex-Based Differences in the Effects of Alcohol Abuse: A Comprehensive Functional Meta-Analysis of Transcriptomic Studies. Genes. 2020; 11(9):1106. https://doi.org/10.3390/genes11091106
6. Pérez-Díez I, Hidalgo MR, Malmierca-Merlo P, Andreu Z, Romera-Giner S, Farràs R, de la Iglesia-Vayá M, Provencio M, Romero A, García-García F. Functional Signatures in Non-Small-Cell Lung Cancer: A Systematic Review and Meta-Analysis of Sex-Based Differences in Transcriptomic Studies. Cancers. 2021; 13(1):143. https://doi.org/10.3390/cancers13010143
7. Català-Senent JF, Hidalgo MR, Berenguer M, Parthasarathy G, Malhi H, Malmierca-Merlo P, de la Iglesia-Vayá M, García-García F. Hepatic Steatosis and Steatohepatitis: A Functional Meta-analysis of Sex-based Differences in Transcriptomic Studies. Biology of Sex Differences 2021, 12:29; https://doi.org/10.1186/s13293-021-00368-1
8. Plotly Technologies Inc. Collaborative data science. Montréal, QC, 2015. https://plot.ly
9. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015; Volume 43, Issue 7, 20, Page e47, https://doi.org/10.1093/nar/gkv007
10. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SC, 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 Oct 2005, 102 (43) 15545-15550; doi: 10.1073/pnas.0506580102
11. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-29. doi:10.1038/75556
12. Montaner D, Dopazo J. Multidimensional Gene Set Analysis of Genomic Data. PLoS One. 2010; 5(4), e10348. doi: 10.1371/journal.pone.0010348
13. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software. 2010; 36(3), 1–48. https://www.jstatsoft.org/v36/i03/
14. https://www.mongodb.com/
15. Jain N, Mangal P & Mehta D. AngularJS: A modern MVC framework in JavaScript. Journal of Global Research in Computer Science. 2014; 5(12), pp.17–23
16. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. 2016. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Supplementary material for

MetaFun: Unveiling sex differences in multiple omics studies through comprehensive functional meta-analysis

Pablo Malmierca-Merlo, Rubén Sánchez-Garcia, Rubén Grillo-Risco, Irene Pérez-Díez, José F. Català-Senent, María de la Iglesia-Vayá, Marta R. Hidalgo, Francisco Garcia-Garcia

S1. Web Tool overview

The web tool can be used with an anonymous user or with a registered one. Registered users will keep their data and jobs stored from one session to the other, while data and jobs from anonymous users will not be saved after leaving the session. The general design of the web tool includes an upper right menu with the basic functionalities of the tool, a left side panel with specific submenus, and a central panel from which to interplay with the web. After logging in, the user is directed to the form launching a new job, which can be otherwise accessed through the *New Analysis* button on the top right menu. The *New Analysis* form goes through a series of steps asking for different information which must be filled in (see Supplementary Section S2 for details), and allows to launch a new meta-analysis. After the launchment and execution of the job, it will be listed on the jobs area, which can be accessed through the *My jobs* button located on the top right menu. There, all created jobs are listed and can be accessed through the left side submenu to visualize their results. Through the top right panel the user can also access the user area, through the button named after his or her user name. The user area includes a browser of the user folder and information about all the launched jobs. The user area submenu allows a series of actions related to personal settings and deleting options. The top right panel includes also an exit icon button which logs the user out, and a question mark icon button which opens the documentation of the web tool, accessible also through https://gitlab.com/ubb-cipf/metafunweb/-/wikis/Summary.

S2. Input data

All datasets included in a same meta-analysis should be comparable, including similar experimental designs and individuals with similar conditions. At least two datasets must be included in a meta-analysis. Input data consists of one expression matrix and one experimental design file for each of the datasets in the meta-analysis. The expression matrix must have been normalized, with samples in the columns and EntrezID genes in the rows. The experimental design file must indicate the original group to which each sample belongs, with samples in the rows and groupings in the columns. More than one grouping per file is accepted. Accepted file formats are CSV or TSV for both the expression matrix and the experimental design files.

S3. Launching a meta-analysis

The *New Analysis* button on the top right menu directs to the form to launch a new meta-analysis. The first tab of the form, labeled *Files*, includes a browser of the user's files and allows the user to upload and manage the datasets to analyze. Tab *Options* allows the user to specify the Effect Model to random or fixed, to select the reference organism among Homo sapiens, Mus musculus and Rattus norvegicus, and to define the Gene Ontology ontologies to analyze (Biological Process, Molecular Function and Cellular Component) and whether to propagate the annotation. Tab *Studies* is the interface to select the studies to meta-analyze and their experimental design. The selection is done dragging the files from the right panel entitled *My Files* to the columns *Expression* or *Experimental Design*, depending on the case. Matched studies and experimental designs must be placed in the same row, and a verification of their compatibility is performed to avoid future errors. In the *Comparison* tab the user can specify the kind of comparison to perform. Two different options are available: the classical *Case vs. Control*, where we compare the effect of a variable, and the comparison with sex perpective *(Case Female vs. Control Female) vs. (Case Male vs. Control Male)*, in which we compare the effect of a variable in females with respect to the effect in males. In the second case, significant results will just refer to differential effects between males and females, and may not coincide with the results of the first comparison. After the selection of the comparison, the user must indicate which samples of each study are included in each canonical compared group (*Case*, *Control*, *Case Female*, etc). This is done through the assignment of one of the classes in the experimental design of each study to these canonical groups. Finally, in the *Launch* tab, a summary of the defined meta-analysis is shown, and, after the assignment of a name, it may be launched through the button *Launch job*.

S4. Analysis summary

After the execution of the job, and its selection in the left side panel of the *My Jobs* panel, the *Analysis summary* tab will show a summary of its main results. This summary includes the selected analysis options (name, comparison, effect model, functional profile, and reference organism), a table and an interactive barplot describing the number of samples per dataset and per group, a table describing the number of differentially expressed genes in each dataset (per columns, the studies, number of total analyzed genes, number of total significant genes, number of significant up-regulated genes, and number of significant down-regulated genes), a table including the same columns describing the number of significant functional profile items in each dataset (either enriched functions or differentially activated subpathways, depending on the selected functional profile), and a table also including the same columns describing the number of significant functional terms in each ontology (BP for Biological Process, MF for Molecular Functional and CC for Cellular Component) of the meta-analysis.

S5. Exploratory analysis

The *Exploratory analysis* tab contains the figures resulting from the unsupervised exploratory analysis performed on each dataset in the meta-analysis. This analysis includes a boxplot representation of the expression of the samples, a clustering of the samples and a Principal

Components Analysis (PCA) plot representing the first two components of the PCA. All samples are colored by the experimental design selected in the meta-analysis.

S6. Differential expression

The differential expression analysis is performed with library *limma* [9], applying *lmFit, contrast.fit* and *eBayes* functions, and taking into account whether the samples are paired or not. Results are displayed as a table in the *Differential expression* tab of the job once it is ready. The table shows the EntrezID, Gene Name, logarithm of the fold-change (logFC), test statistic, raw p-value and Bonferroni-Holm adjusted p-value of each analyzed feature. The table is initially ordered by the raw p-value, but buttons on the names of the columns allow the user to order the table differently. Links from the EntrezID column direct to the NCBI gene database of the specific gene. Different tools allow the user to search, download and filter the table by a maximum p-value.

S7. Gene Set Enrichment Analysis

The functional analysis consists of a Gene Set Enrichment Analysis (GSEA) [10] based on the Biological Process, Molecular Function and Cellular Component ontologies from Gene Ontology (GO) [11] to the user's wish. The pipeline, performed with the *mdgsa* library [12], splits the ontologies, propagates the annotation, filters too generic (more than 500 annotated genes) or too specific (less than 10 annotated genes) annotations, transforms the p-value into an index and performs the corresponding contrasts. Results are displayed as a table in the *GSA* tab of the job once it is ready. Three sub-tabs on the top right of the table show the results separately for the three different ontologies. For each ontology, the table shows the GO ID, GO term, logarithm of the odds-ratio (LOR), raw p-value, Bonferroni-Holm adjusted p-value and number of genes included in each analyzed feature. The table is initially ordered by the raw p-value, but buttons on the names of the columns allow the user to order the table differently. Links from the GO ID column direct to the QuickGO entry of the specific term. Different tools allow the user to search, download and filter the table by a maximum p-value.

S8. Meta-analysis

The functional meta-analysis integrates the results of the functional analysis and is performed using the *rma* function in the *metafor* package [13]. For each of the functions, a meta-analysis is carried out that combines the level of overrepresentation (LOR) of that function in the different studies. Two methods have been implemented to perform the meta-analysis: the fixed effects models (FE) and the random effects models (DL DerSimonian & Laird; HS Schmidt & Hunter; Hedges, HE) [13]. The fixed effect model has been designed for similar studies (i.e. with the same technology, platform and in similar times), while the random effect model allows to compute more variability. Results are displayed as a table in the *Meta-Analysis* tab of the job once it is ready. The table shows the GO ID, GO term, LOR, confidence interval of the LOR, raw p-value, and Bonferroni-Holm adjusted p-value of each analyzed feature. The table is initially ordered by the raw p-value, but buttons on the names of the columns allow the user to order the table differently. Links from the GO ID column direct to the QuickGO entry of the specific term. Different tools allow the user to search, download and filter the table by a maximum p-value.

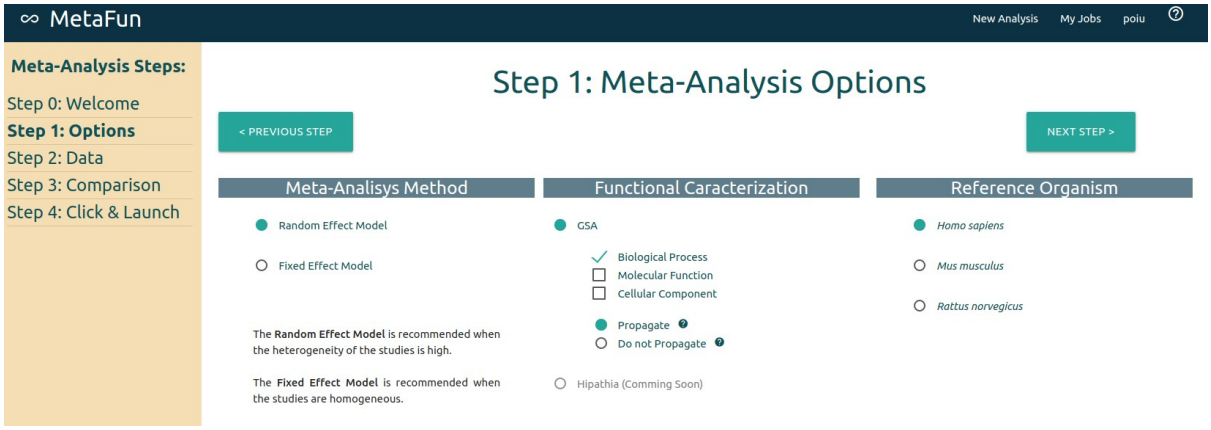
S9. Study cases

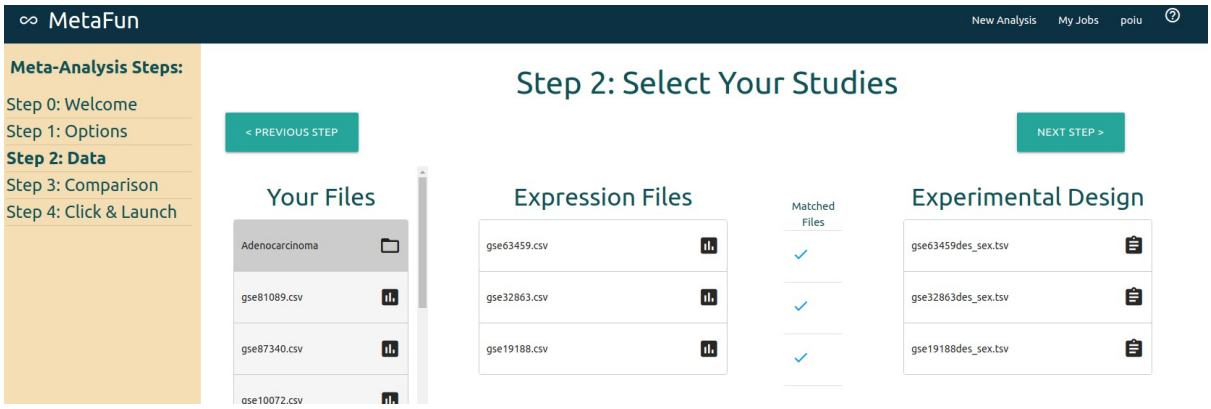
The following case describes the potential use of MetaFun in the characterization of sex differences in lung adenocarcinoma. The results obtained were published in Cancers. 2021 Jan 5;13(1):143. doi: 10.3390/cancers13010143.

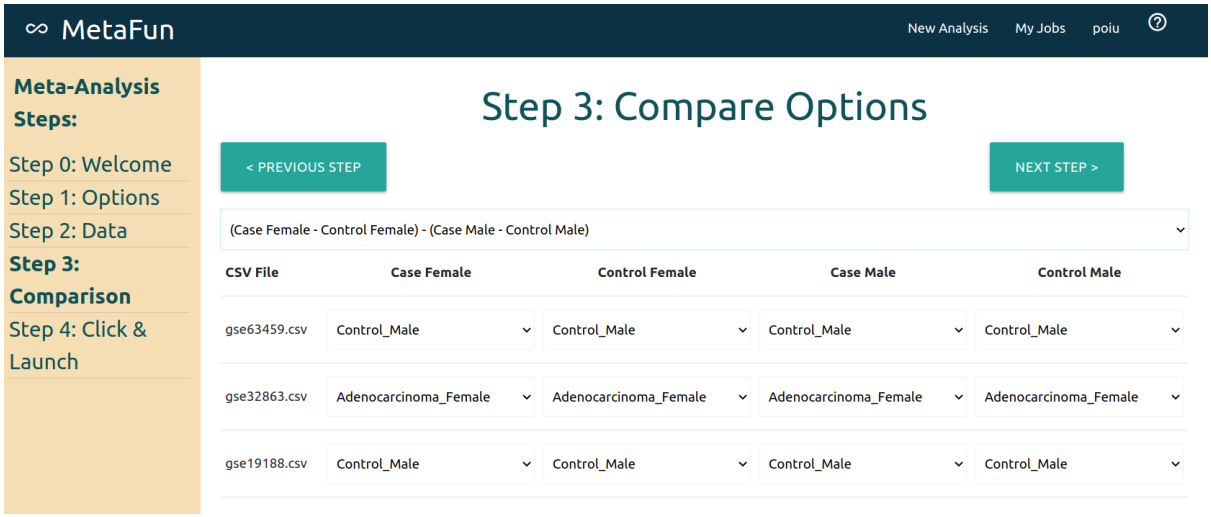
Input data:

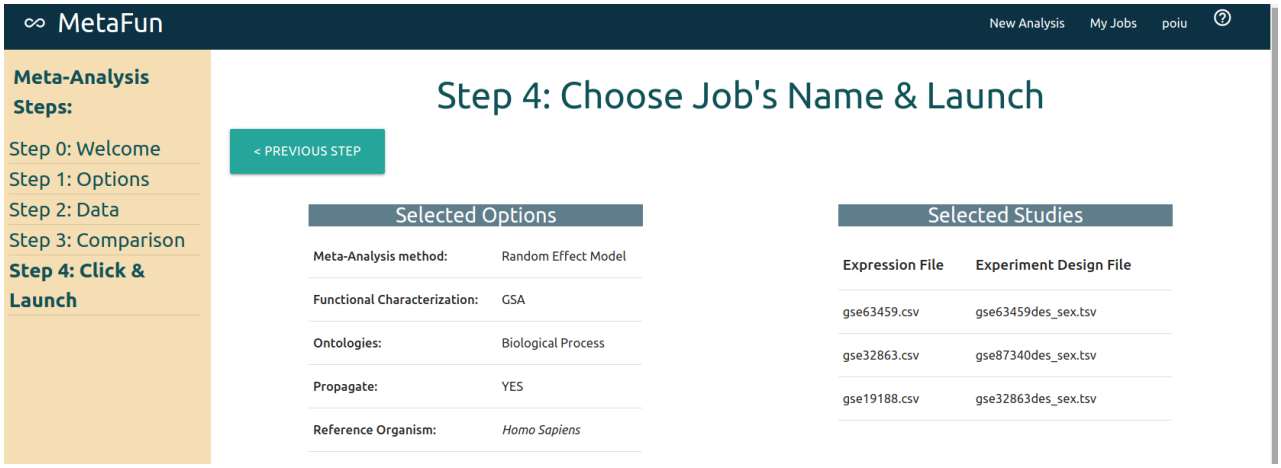
For each of the studies we will need two files: a first one with the expression data and a second one with the description of the experimental groups to which each sample belongs, indicating the sex of the participant. In this link you can download the files corresponding to this use case:
https://gitlab.com/ubb-cipf/metafunpipeline/-/blob/master/metafun_sample_data.tar

4 easy steps to launch the meta-analysis job:









Results:

For each of the sections described above (1. Analysis Summary, 2. Exploratory Analysis, 3. Differential Expression, 4. Gene Set Analysis, 5. Meta-Analysis), we show the results generated by MetaFun in this use case:

1. Analysis Summary

A summary of the results at the different stages of the bioinformatics analysis strategy is detailed:

SexDifAdeno Results | 1. Analysis Summary | 2. Exploratory Analysis | 3. Differential Expression | 4. GSA | 5. Meta-Analysis

Analysis Summary

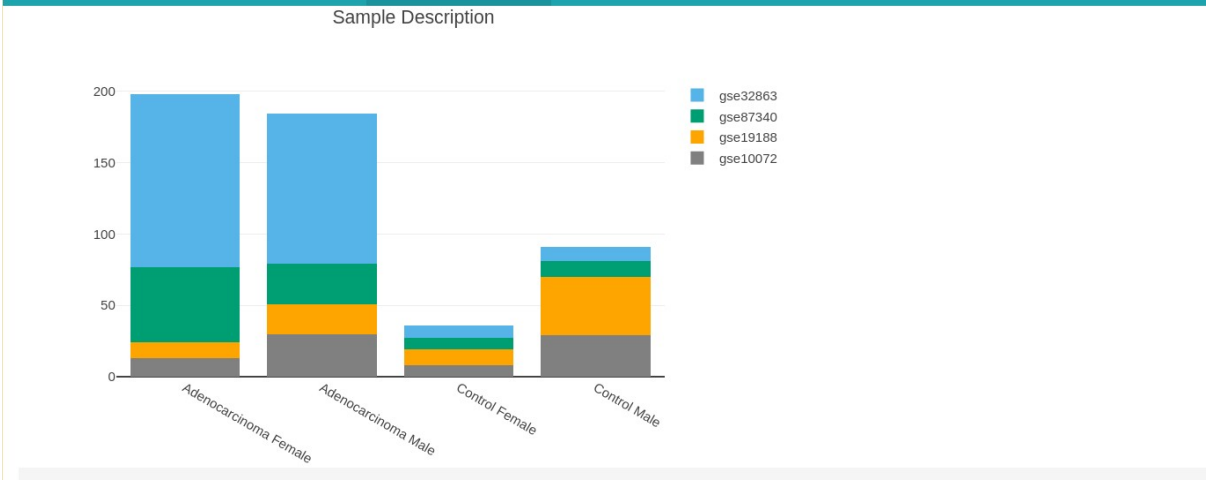
Job Options

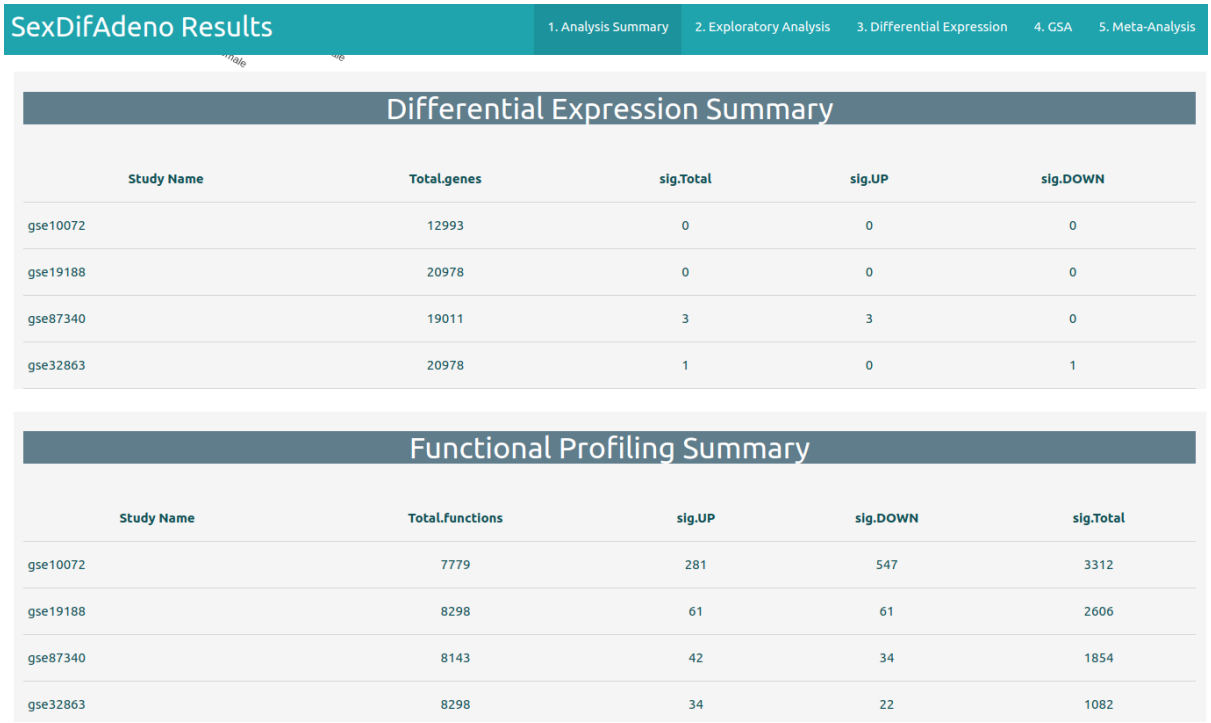
Name	Contrast	Effect Model	Functional Profiling	Reference Organism
SexDifAdeno	(ME-MS)-(HE-HS)	Random Effect Model	GSA	hsa

Samples Description

Study Name	Adenocarcinoma Female	Adenocarcinoma Male	Control Female	Control Male	Total
gse10072	13	30	8	29	80
gse19188	11	21	11	41	84
gse87340	53	28	8	11	100
gse32863	121	105	9	10	245
Total	198	184	36	91	509

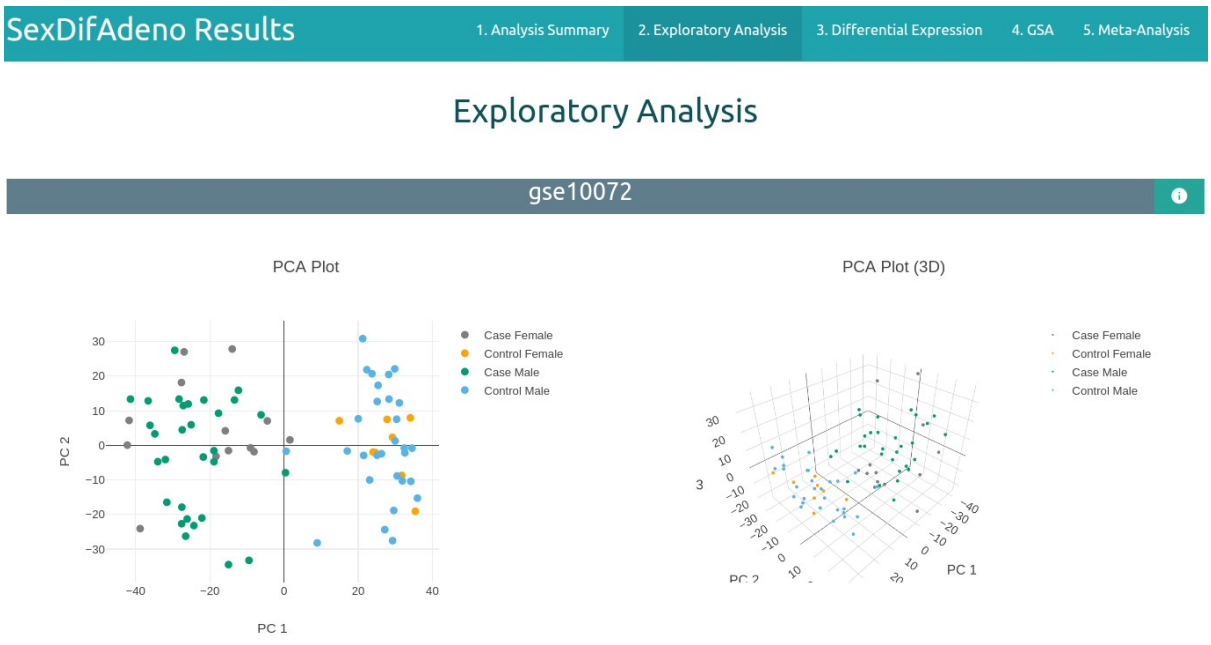
SexDifAdeno Results | 1. Analysis Summary | 2. Exploratory Analysis | 3. Differential Expression | 4. GSA | 5. Meta-Analysis

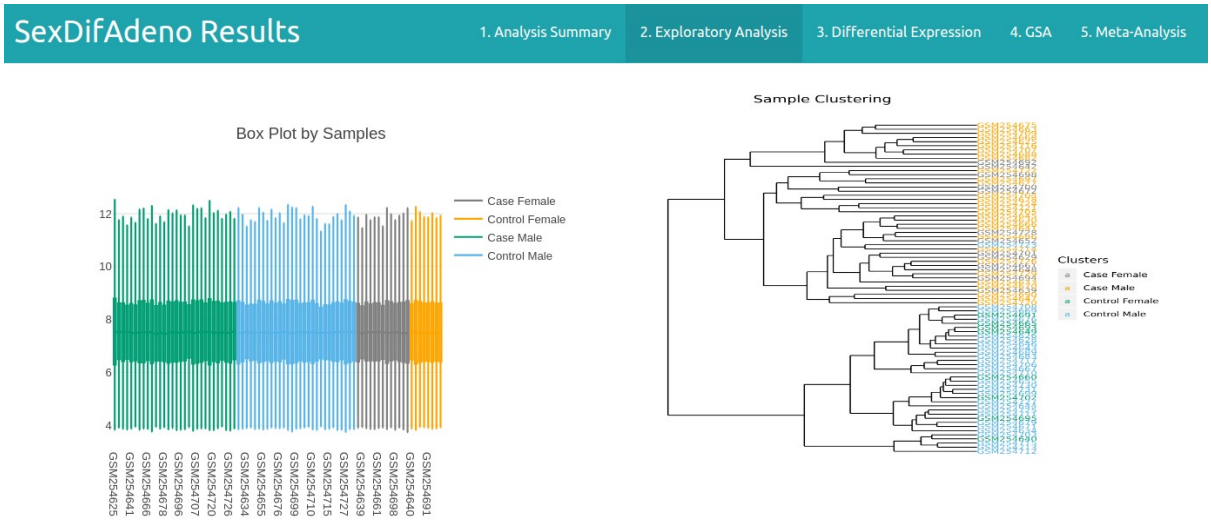




2. Exploratory Analysis

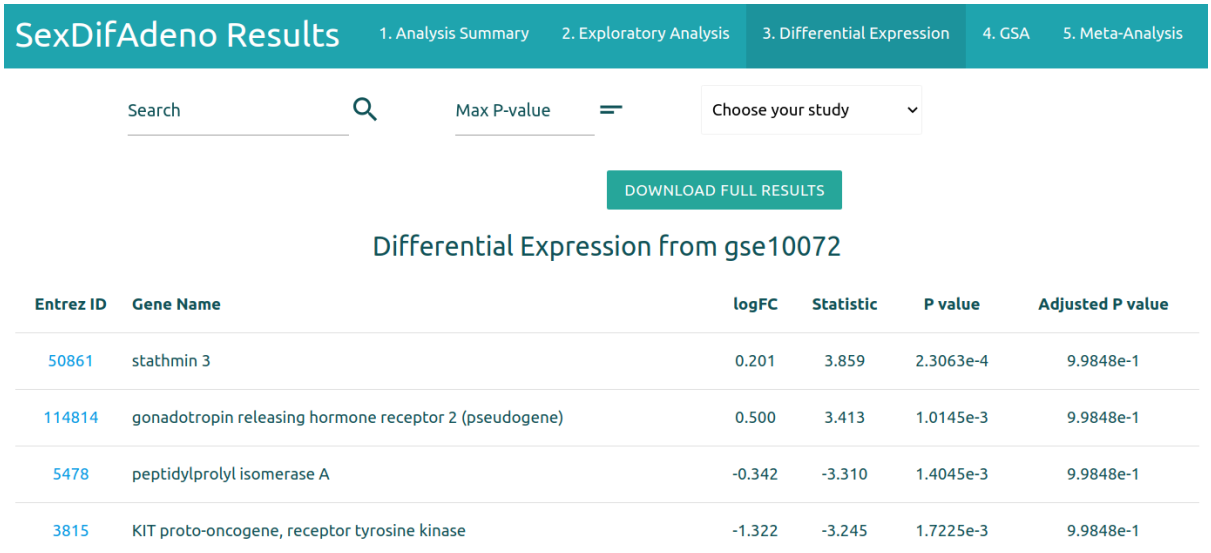
Principal component analysis, clustering and boxplots are used to explore the expression levels of each of the samples in the selected studies:





3. Differential Expression

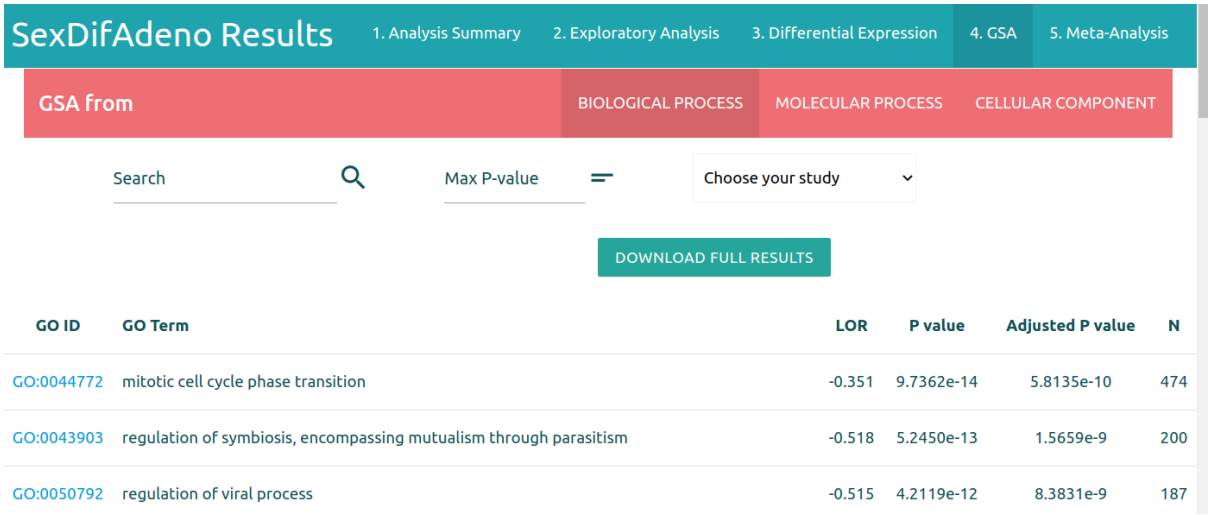
For each of the studies, we identified which genes show a differential expression pattern by sex in the disease. By clicking on each of the links to the gene identifiers, you can expand their biological information:



4. Gene Set Analysis (GSA)

Functional characterization of the differential expression results will identify which functions

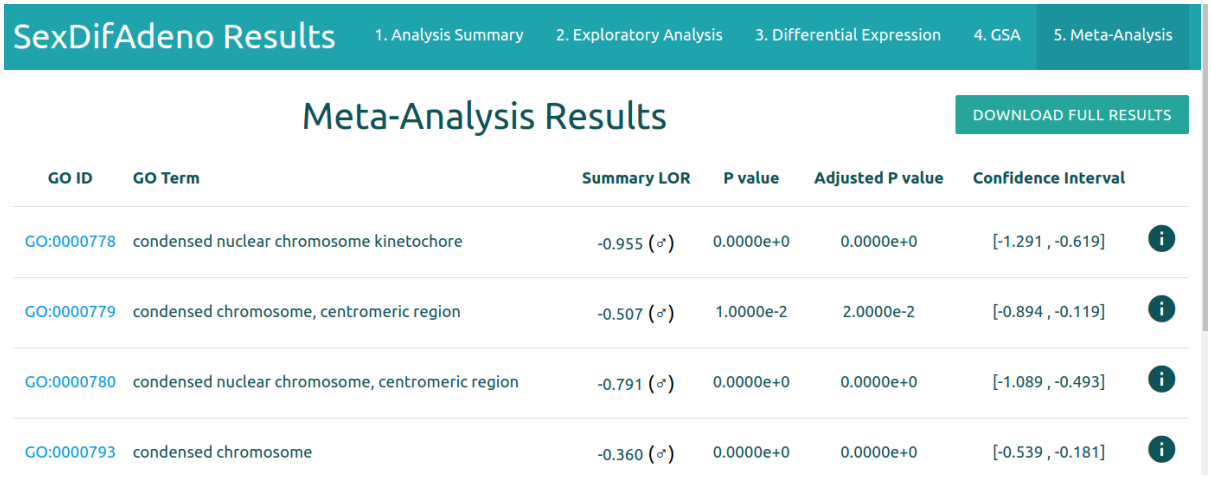
are more active in males and females. The information for each of the significant functions can be expanded by clicking on the link to its identifier.



GO ID	GO Term	LOR	P value	Adjusted P value	N
GO:0044772	mitotic cell cycle phase transition	-0.351	9.7362e-14	5.8135e-10	474
GO:0043903	regulation of symbiosis, encompassing mutualism through parasitism	-0.518	5.2450e-13	1.5659e-9	200
GO:0050792	regulation of viral process	-0.515	4.2119e-12	8.3831e-9	187

5. Meta-Analysis

Finally in this section, MetaFun shows the functions and pathways that are activated in the set of studies evaluated. If we click on the information icon, we will obtain detailed information on each of these significant functions:



GO ID	GO Term	Summary LOR	P value	Adjusted P value	Confidence Interval
GO:0000778	condensed nuclear chromosome kinetochore	-0.955 (♂)	0.0000e+0	0.0000e+0	[-1.291 , -0.619]
GO:0000779	condensed chromosome, centromeric region	-0.507 (♂)	1.0000e-2	2.0000e-2	[-0.894 , -0.119]
GO:0000780	condensed nuclear chromosome, centromeric region	-0.791 (♂)	0.0000e+0	0.0000e+0	[-1.089 , -0.493]
GO:0000793	condensed chromosome	-0.360 (♂)	0.0000e+0	0.0000e+0	[-0.539 , -0.181]