MetaPhOR

Emily Isenhart

Introduction

MetaPhOR was developed to enable users to assess metabolic dysregulation using transcriptomic-level data (RNA-sequencing and Microarray data) and produce publication-quality figures. A list of differentially expressed genes (DEGs), which includes fold change and p value, from DESeq2 (Love, Huber, and Anders (2014)) or limma (Ritchie et al. (2015)), can be used as input, with sample size for MetaPhOR, and will produce a data frame of scores for each KEGG pathway. These scores represent the magnitude and direction of transcriptional change within the pathway, along with estimated p-values (Rosario et al. (2018)). MetaPhOR then uses these scores to visualize metabolic profiles within and between samples through a variety of mechanisms, including: bubble plots, heatmaps, and pathway models.

Installation

This command line can be used to install and load MetaPhOR.

if (!require("BiocManager", quietly = TRUE))
install.packages("BiocManager") BiocManager::install("MetaPhOR")

```
library(MetaPhOR)
```

Data Preparation

Minimal data preparation is required to run MetaPhOR. DEGs may be loaded into R in the form of .csv or .tsv for use with this package. The DEG file must contain columns for log fold change, adjusted p-value, and HUGO gene names. By default, MetaPhOR assumes DESeq2 header stylings: "log2FoldChange" and "padj". In any function that assumes these headers, however, the user can define column names for these values. Below is a sample DEG file resulting from limma:

```
exdegs <- read.csv(system.file("extdata/exampledegs.csv", package = "MetaPhOR"),
                   header = TRUE)
```

X	logFC	AveExpr	t	P.Value	adj.P.Val	B
TATDN1	0.0996779	6.112478	6.074387	0e+00	0.0000523	10.736686
ZNF706	0.1055866	6.446976	5.845307	0e+00	0.0000958	9.237467
DCAF13	0.0882855	6.393943	5.513711	1e-07	0.0002918	7.531541
TRMT12	0.1093866	6.071099	5.434470	1e-07	0.0003550	7.377948
IGF2BP1	0.9256589	4.454204	5.631783	0e+00	0.0002065	7.273836
MAP6D1	0.1619178	5.719428	5.342911	1e-07	0.0004612	7.090578

Pathway Analysis

"pathwayAnalysis" first assigns scores and their absolute values using log fold change and p value to each gene (Rosario et al. (2018)). These transcript-level scores, along with sample size, are then utilized to calculate both scores (directional change) and absolute value scores (magnitude of change) (Rosario et al. (2018)) for each KEGG Pathway (Kanehisa and Goto (2000)). We then utilize a bootstrapping method, to randomly calculate 100,000 scores per pathway, based on the number of genes in that pathway and model the distribution. This distribution can then be used to evaluate where the actual score for that pathway sits in relation to the distribution, and can assign a p-value to the achieved score.

For example, if the polyamine biosynthetic pathway contains 13 genes, we can get a score for the sum of the 13 genes that exist within that pathway. Using bootstrapping, we can then randomly sample, with replacement, 13 genes to create scores, 100,000 times. We use these random samples (100,000) to generate a distribution, and we can calculate a p-value dependent on where the score that consists of the 13 genes that actually exist within the pathway falls within the distribution.

Taken together, the scores and p-values resulting from "pathwayAnalysis" provide a measure for both the biological and statistical significance of metabolic dysregulation.

Note: A seed MUST be set before utilizing this function to ensure reproducible results by bootstrapping. It is NECESSARY that the seed remain the same throughout an analysis.

pathwayAnalysis() requires:

- The file path to the DEG list of interest
- Correct headers for fold change and p value columns (as indicated above)
- The name of the column containing HUGO gene names
- The sample size of the DEG analysis

A partial output of the pathway analysis function can be seen as follows:

```
set.seed(1234)

brca <- pathwayAnalysis(system.file("extdata/BRCA_DEGS.csv",
            package = "MetaPhOR"), "X", 1095, headers = c("logFC", "adj.P.Val"))
```

	Scores	ABSScores	ScorePvals	ABSScorePvals
Cardiolipin.Metabolism	0.0293332	0.0299031	0.23132	0.60257
Cardiolipin.Biosynthesis	0.0031671	0.0031671	0.43061	0.91996
Cholesterol.Biosynthesis	0.2400430	0.2614076	0.06227	0.35879
Citric.Acid.Cycle	0.0314352	0.1469312	0.48484	0.92836
Cyclooxygenase.Arachidonic.Acid.Metabolism	0.0004634	0.0080444	0.51948	0.92329
Prostaglandin.Biosynthesis	-0.0392858	0.1046875	0.79542	0.35333

bubblePlot

The metabolic profile determined by pathway analysis can be easily visualized using "bubblePlot." Scores are plotted on the x-axis, while absolute value scores are plotted on the y-axis. Each point represents a KEGG pathway, where point size represents p-value (the smaller the p value, the larger the point) and point color is dictated by scores. Negative scores, which indicate transcriptional downregulation, are blue, and positive scores, which indicate transcriptional upregulation, are red. The top ten points, either by smallest

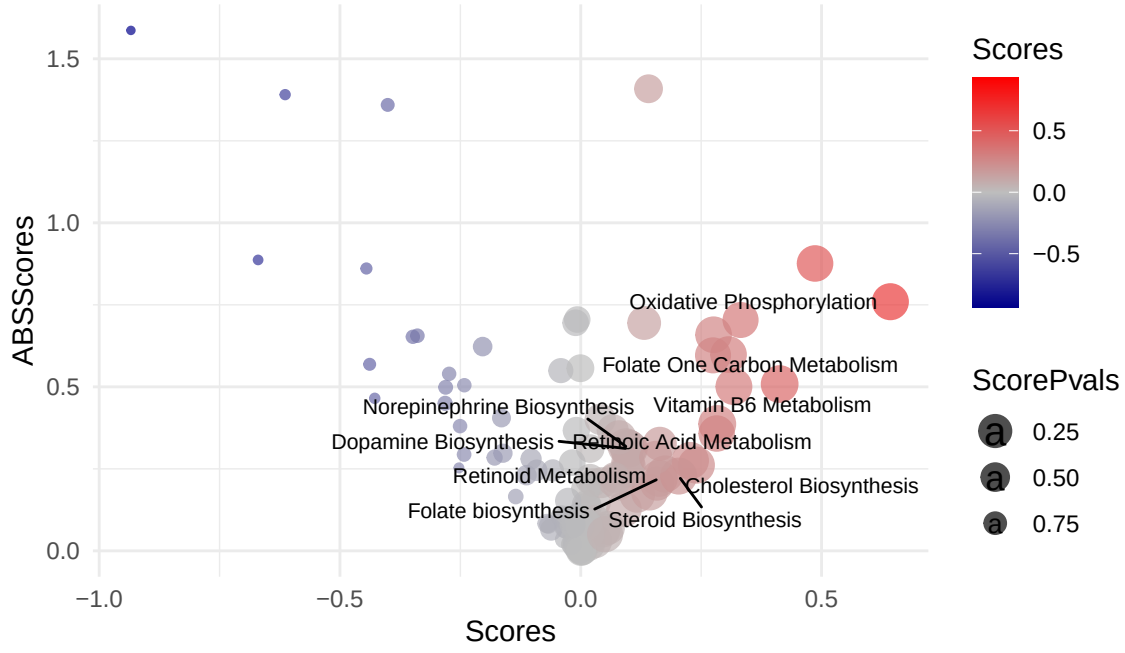
p value or greatest dysregulation by score, are labeled with their pathway names. The plot demonstrates which pathways are the most statistically and biologically relevant.

bubblePlot() requires:

- The output of pathwayAnalysis(), as a data frame
- An indication which values to use, in order to label points: either “Pval” or “LogFC”
- Optional: Numeric value for point label text size (default = .25)

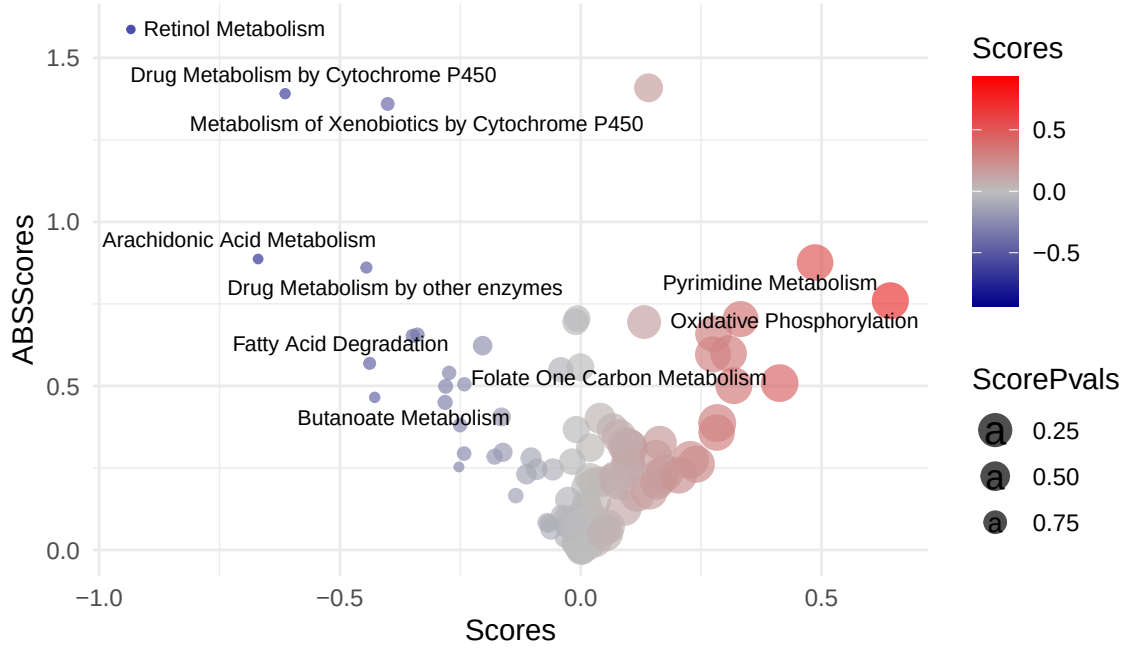
Bubble Plot Labeled by P Value

```
pval <- bubblePlot(brca, "Pval", .85)
plot(pval)
```



Bubble Plot Labeled by LogFC

```
logfc <- bubblePlot(brca, "LogFC", .85)
plot(logfc)
```



metaHeatmap

"metaHeatmap" provides a useful visualization for comparing metabolic profiles between groups, including only significantly dysregulated pathways, and highlighting those which are most highly changed. This function should be used when you have multiple groups/DEGs being compared, e.g. if you have 4 conditions all being compared to each other. This will not be useful if you have a single DEG list. This function can be used only when multiple DEG comparisons are scored by "pathwayAnalysis." The absolute pathway scores are scaled across outputs and plotted via pheatmap, selecting only those which have absolute score p values below the level of significance.

Note: A heatmap cannot be produced if there are no pathways significantly dysregulated below the p value cut off.

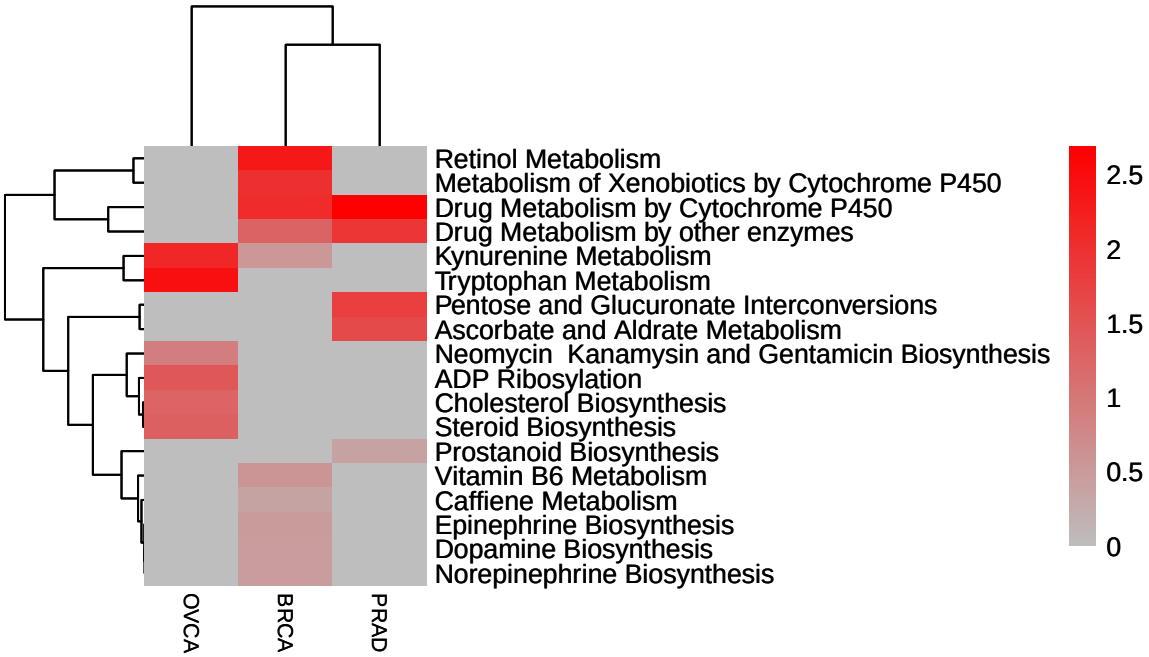
metaHeatmap() requires:

- A list of outputs from pathway analysis, as data frames
- A character vector of names for labeling each output
- Optional: The p value cut off to be used (default = 0.05)

```
##read in two additional sets of scores,
##run in the same manner as brca for comparison

ovca <- read.csv(system.file("extdata/OVCA_Scores.csv",
                             package = "MetaPhOR"), header = TRUE, row.names = 1)
prad <- read.csv(system.file("extdata/PRAD_Scores.csv",
                             package = "MetaPhOR"), header = TRUE, row.names = 1)

all.scores <- list(brca, ovca, prad)
names <- c("BRCA", "OVCA", "PRAD")
metaHeatmap(all.scores, names, 0.05)
```



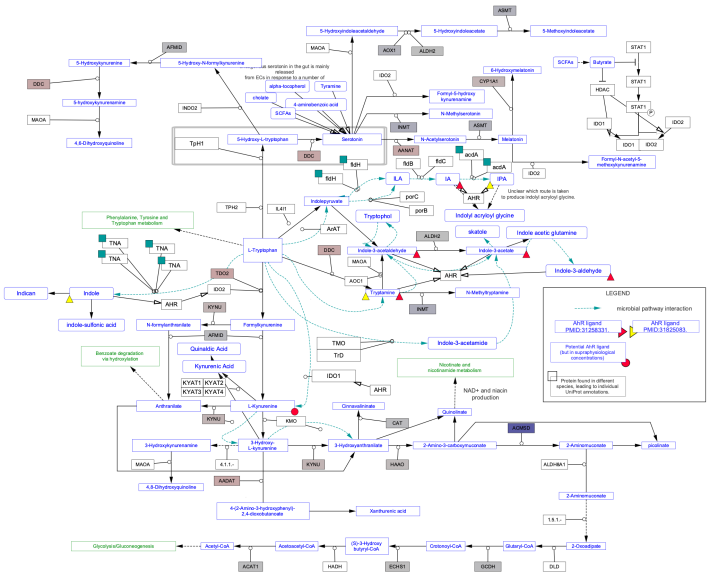
cytoPath

"cytoPath" models metabolic pathways, sourced from WikiPathway (Martens et al. (2021)), using the transcriptional change of individual genes in the pathway. The resulting figure can be used to identify candidate genes for further study by providing a detailed look at transcriptional dysregulation within and between "pathwayAnalysis" outputs. Pathways of interest can be readily identified from "bubblePlot" or "metaHeatmap." This function utilizes Cytoscape (Gustavsen et al. (2019)); it is required that Cytoscape be open locally with R to render this plot. For details regarding legends, see the Cytoscape Legend Creator Tutorial (link).

cytoPath() requires:

- The name of the pathway to be plotted
- The file path to the DEG list of interest
- The file path to which the figure will be saved
- The name of the column containing HUGE gene names

```
cytoPath("Tryptophan Metabolism", "BRCA_DEGS.csv", paste(getwd(),
"BRCA_Tryptophan_Pathway", sep = "/"), "X", headers = c("logFC", "adj.P.Val"))
```



pathwayList

The WikiPathways available within MetaPhOR have been restricted to metabolic pathways. The pathwayList function will provide a complete list of WikiPathways for mapping with "cytoPath."

```
pathwayList()
```

Conclusion

MetaPhOR contains five functions for the analysis of metabolic dysregulation using transcriptomic-level data and the production of publication-quality figures. This is version of MetaPhOR, 0.99.0, is available at https://github.com/emilyise/MetaPhOR.

wpid2name$name
Acetylcholine synthesis
Activation of vitamin K-dependent proteins
Adipogenesis
Aerobic glycolysis
Aflatoxin B1 metabolism
AGE/RAGE pathway

SessionInfo

```
sessionInfo()
## R version 4.2.0 (2022-04-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur/Monterey 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] MetaPhOR_0.99.0  kableExtra_1.3.4
##
## loaded via a namespace (and not attached):
##   [1] shadowtext_0.1.2       uuid_1.1-0             backports_1.4.1
##   [4] fastmatch_1.1-3        systemfonts_1.0.4      plyr_1.8.7
##   [7] igraph_1.3.4           RecordLinkage_0.4-12.3 lazyeval_0.2.2
##  [10] repr_1.1.4             RCy3_2.16.0            splines_4.2.0
##  [13] BiocParallel_1.30.3    listenv_0.8.0          GenomeInfoDb_1.32.4
##  [16] ggplot2_3.3.6          digest_0.6.29          yulab.utils_0.0.5
##  [19] htmltools_0.5.3        GOSemSim_2.22.0        viridis_0.6.2
##  [22] GO.db_3.15.0           fansi_1.0.3            magrittr_2.0.3
##  [25] memoise_2.0.1          base64url_1.4          graphlayouts_0.8.1
##  [28] globals_0.16.1         Biostrings_2.64.1      svglite_2.1.0
##  [31] enrichplot_1.16.2      colorspace_2.0-3       ggrepel_0.9.1
##  [34] blob_1.2.3             rvest_1.0.3            xfun_0.32
##  [37] dplyr_1.0.10           crayon_1.5.1           RCurl_1.98-1.8
##  [40] jsonlite_1.8.0         scatterpie_0.1.8       graph_1.74.0
##  [43] ape_5.6-2              survival_3.4-0         glue_1.6.2
##  [46] polyclip_1.10-0        gtable_0.3.1           ipred_0.9-13
##  [49] zlibbioc_1.42.0        XVector_0.36.0         webshot_0.5.3
##  [52] evd_2.3-6.1            future.apply_1.9.1     BiocGenerics_0.42.0
##  [55] scales_1.2.1           DOSE_3.22.1            pheatmap_1.0.12
##  [58] DBI_1.1.3              Rcpp_1.0.9             viridisLite_0.4.1
##  [61] xtable_1.8-4           tidytree_0.4.0         gridGraphics_0.5-1
##  [64] bit_4.0.4              proxy_0.4-27           stats4_4.2.0
```

```
##  [67] lava_1.6.10           prodlim_2019.11.13    httr_1.4.4
##  [70] fgsea_1.22.0          RColorBrewer_1.1-3    ff_4.0.7
##  [73] pkgconfig_2.0.3       XML_3.99-0.10         farver_2.1.1
##  [76] nnet_7.3-17           utf8_1.2.2            RJSONIO_1.3-1.6
##  [79] labeling_0.4.2        ggplotify_0.1.0       tidyselect_1.1.2
##  [82] rlang_1.0.5           reshape2_1.4.4        AnnotationDbi_1.58.0
##  [85] munsell_0.5.0         tools_4.2.0           cachem_1.0.6
##  [88] downloader_0.4        cli_3.3.0             generics_0.1.3
##  [91] RSQLite_2.2.16        evaluate_0.16         stringr_1.4.1
##  [94] fastmap_1.1.0         yaml_2.3.5            ggtree_3.4.2
##  [97] knitr_1.40            bit64_4.0.5           fs_1.5.2
## [100] tidygraph_1.2.2       purrr_0.3.4           KEGGREST_1.36.3
## [103] ggraph_2.0.6          nlme_3.1-159          future_1.28.0
## [106] aplot_0.1.7           DO.db_2.9             xml2_1.3.3
## [109] compiler_4.2.0        rstudioapi_0.14       png_0.1-7
## [112] e1071_1.7-11          treeio_1.20.2         tibble_3.1.8
## [115] tweenr_2.0.2          stringi_1.7.8         highr_0.9
## [118] lattice_0.20-45       IRdisplay_1.1         Matrix_1.4-1
## [121] vctrs_0.4.1           pillar_1.8.1          lifecycle_1.0.1
## [124] data.table_1.14.2     bitops_1.0-7          patchwork_1.1.2
## [127] qvalue_2.28.0         R6_2.5.1              gridExtra_2.3
## [130] IRanges_2.30.1        parallelly_1.32.1     ada_2.0-5
## [133] codetools_0.2-18      MASS_7.3-58.1         withr_2.5.0
## [136] uchardet_1.1.0        S4Vectors_0.34.0      GenomeInfoDbData_1.2.8
## [139] parallel_4.2.0        clusterProfiler_4.4.4 ggfun_0.0.7
## [142] grid_4.2.0            rpart_4.1.16          IRkernel_1.3
## [145] tidyr_1.2.0           class_7.3-20          rmarkdown_2.16
## [148] pbdZMQ_0.3-7          ggforce_0.3.4         Biobase_2.56.0
## [151] base64enc_0.1-3
```

References

Gustavsen, Julia A, Shraddha Pai, Ruth Isserlin, Barry Demchak, and Alexander R Pico. 2019. "RCy3: Network Biology Using Cytoscape from Within r." *f1000Research* 8 (1774). https://doi.org/10.12688/f1000research.20887.3.

Kanehisa, Minoru, and Susumu Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28: 27–30. https://doi.org/10.1093/nar/28.1.27.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15: 550. https://doi.org/10.1186/s13059-014-0550-8.

Martens, Marvin, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N Slenter, Kristina Hanspers, Ryan A Miller, et al. 2021. "WikiPathways: Connecting Communities." *Nucleic Acids Research* 49: D613–621. https://doi.org/10.1093/nar/gkaa1024.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Rosario, S R, M D Long, H C Affronti, A M Rowsam, K H Eng, and D J Smirglia. 2018. "Pan-Cancer Analysis of Transcriptional Metabolic Dysregulation Using the Cancer Genome Atlas." *Nature Communications* 9 (5330). https://doi.org/10.1038/s41467-018-07232-8.