Automated, Reproducible Investigation of gene set Differential Enrichment via the AUTO-go framework

Eleonora Sperandio[1], Isabella Grassucci[1], Lorenzo D'Ambrosio[2] and Matteo Pallocca[1*]

[1] UOSD Biostatistics, Bioinformatics and Clinical Trial Center, IRCSS Regina Elena National Cancer Institute, Rome

[2] UOSD Immunology and Tumor Immunotherapy, IRCSS Regina Elena National Cancer Institute, Rome

*matteo.pallocca@ifo.gov.it

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Reproducibility in Life Sciences is challenged in the analysis of large multi-omics datasets. One of the final steps of said processes is Gene Set enrichment, where web tools represent a valuable resource but not a reliable surrogate for standardized, high-quality visualizations. The AUTO-go framework proposes standardization of the Gene Functional Enrichment process along with an R framework able to produce high-quality visualization in an automated manner, improving the reproducibility of the whole analytical process. We present three use cases in Cancer Transcriptomics and Epigenomics datasets as a proof-of-concept to visualize Multiple Differential Expression and Single Sample Gene Set Enrichment Analysis.

Author Summary

Bioinformatics and Data Science are routinely challenged to distill intelligible results from huge amounts of data. These results, in turn, are conveyed through plots and visualizations that should be easily reproducible for scientific soundness and ethical reasons. A specific area in which these analyses are of critical importance is Genomics, where Genes functions need to be enriched when comparing pathological states or treatments. Here we present a software framework that aims at standardizing said differential analyses and visualizations when dealing with genomics data. Finally, we show how it can be employed to shear light on publicly available datasets, even in small casuistry of Rare Cancers.

Introduction

Gene Ontology (GO) and Pathway Enrichment Analysis are pivotal aspects of Life Science research – but the level of standardization and reproducibility is worryingly low for such popular techniques [1].

Additionally, most of the enrichment analyses currently published rely on web applications that, on the one hand, enable non-bioinformaticians to conduct exploratory analyses; on another, open concern for result reproducibility, being a *manual* step of data processing strongly contrasting the rules for reproducible bioinformatics [2-3].

Virtualization techniques such as Docker and Singularity helped to encapsulate software enabling total reproducibility, while additional workflow management layers such as Nextflow and Snakemake [4-5] enabled to build of complex virtualized pipelines and run them in High-Performance Computing Clusters. Unfortunately, what is presented on a life science paper is not primary output matrices, but functional enrichments that currently do not benefit from such advancements.

Among the R packages available to the community, the *clusterProfiler* is a notable exception, with a development that has focused many features on genomics coordinates enrichment and specific high-throughput experiments, while our focus lies on the high-level conceptualization and visualization of differential analysis [6].

Here we present AUTO-go, a logical and bioinformatics framework that enables (1) reproducible GO analyses; (2) high quality automated visualizations; (3) proposes a high-level visualization for complex experimental designs with multiple comparisons.

Design and implementation

48 According to the logical framework (Fig. 1), a Differential Expression is the most frequent starting
49 input from which one or several gene lists are extracted according to fold change and statistical sig-
50 nificance filters (e.g., *strongly upregulated, log2FC > 1 and padj < 0.05).* The protocol core is an
51 atomic function that enriches a gene list over a list of selected databases, from which several visuali-
52 zations are produced (Fig 1). The gene list can derive from several Genomics applications as de-
53 scribed in the Use Cases section.
54
55 **Gene List Enrichment Visualization**
56 Every <*gene list, database*> combination produces a high-quality bar plot with the top *N*
57 terms enriched, with dynamic resizing to accommodate long terms naming in the final plot.
58 The current implementation of the core module relies on the Enrichr API [7], but it is engi-
59 neered to be generalized with other enrichment functions, with the only constrain of having a gene
60 list as input and a tuple matrix with <*Term, Enrichment Score*> as output.
61
62 **Multiple Comparison Visualization**
63 A classical need in -omics analysis is the representation of functional terms enriched in sever-
64 al conditions or comparisons. The HeatmapGO module is built to provide a high-level visualization
65 of multiple comparisons enrichment, with rows representing terms, such as GO components and
66 Transcription Factors, and columns being experimental comparisons.
67
68 **ssGSEA**
69 In many fields, the scarcity of sample availability does not allow classical statistical modeling. The
70 challenge in obtaining robust results is exacerbated by the employment of -omics profiling, collecting
71 thousands of features per observation. To this purpose, we expanded the AUTO-go package with the

single-sample implementation of the Gene Set Enrichment Analysis Algorithm [8-9], allowing researchers to compare discrete cohorts of samples over known gene signatures.

For all the visualization depicting a subset of the enriched terms, a ranking choice must be made to represent a human-readable number of terms and clusters. In the ssGSEA and HeatmapGO, the top 20 terms are selected by ascending -log10(p-adjusted) score. Other developers and data scientists would pick a different ranking employing a mixture of significance and variance among samples and comparisons to show the functions having a strong modulation.

Results

To provide a proof-of-concept application of our package, we envisioned three analytical settings to test it, namely 1. Large RNA-seq casuistry with multiple comparisons Differential Expression and Enrichment (Tumor Cancer Genome Atlas, TCGA) 2. Discrete in-vitro enrichment of gene lists representing epigenetic signals (Encyclopedia of DNA elements, ENCODE) 3. Discrete in-vivo rare tumor samples profiled via total RNA-seq. (Fig. 2). In the first use case, the TCGA dataset of Skin Cutaneous Melanoma Adenocarcinoma (TCGA-SKCM) was partitioned according to a specific immunotherapy biomarker, the Tumor Mutational Burden (TMB). Differential expression was carried out by comparing all TMB quartiles (Fig. 2,3) [10]. The KEGG 2021 Heatmap shows a stronger enrichment in Ras signaling pathway in the higher group comparison (Q3-Q4), suggesting a switch in the higher mutational load group (Fig. 2,3). This enrichment can be further investigated at the LolliGO level showing that most Ras-related genes are upregulated except for FGF5, while down-regulated genes are more enriched in Cortisol synthesis and secretion, less evident from the Heatmap.

Next, the epigenetic unit test was fetched from the ENCODE database, fetching all the RNA Immunoprecipitation sequencing (RIP-seq) available in the K562 cell line. Gene lists were obtained by annotating with Homer [11] the enriched peaks and extracting only the promoter-TSS records. In

this scenario, the Cellular Component database coupled with the *lolliGO* modules shows stronger enrichment hydrogen peroxide metabolic and catabolic process in mRNA targets of ELAVL1 in (Fig. 4).

Finally, the third unit test was carried out on the on the GSE168493 record, containing total RNA-seq profiles from a small casuistry of Epithelioid hemangioendothelioma, a rare tumor with an incidence of 1 out of million people [12]. In this instance, the ssGSEA package enables to shear light into the pathway activation peculiarities of said tumors, with a stronger enrichment of PSMB5 target genes in samples hEHE.6 and hEHE5 (Fig. 5).

Taken together, all these examples point out many analytical scenarios in which the Auto-GO package can provide a solid foundation and a valuable engineering tool for -omics-focused Bioinformaticians.

Availability and Future Directions

The package is available at https://gitlab.com/bioinfo-ire-release/auto-go. The repository contains a step-by-step tutorial for the whole framework usage and the data input to reproduce the first use case presented in the results section, along with a Dockerfile. All the generated outputs, folders, and figures are available in the tutorial and in Fig. S1.

Figure Captions

Fig 1**. Logical framework and implementation workflow.**

Fig 2**. Use case schema.** Workflow of the three use cases with different casuistry and comparison sizes.

Fig 3**. Results on DE genes enriched on the TCGA-SKCM multiple TMB comparisons.** Rows: enriched terms over KEGG_2021 Enrichr library. Columns: genes regulated in multiple comparisons. Cell content: *-log10(padj + 1)* reported only for significant clusters.

Fig 4**. LolliGO plot from RIP-seq:** Ontology enrichment over a list of ELAVL targets derived from RIP-seq. Color: percentage of the cluster given as input with respect to the total functional cluster. Dot size: gene count for cluster.

Fig 5**. Single-Sample Gene Set Enrichment Analysis heatmap:** Heatmap showing ssGSEA enrichment over the *Hallmark* term for the 6 RNA-seq samples (eEHE1-6). Z-score of the Enrichment Score in cell content.

Fig. S1. **Folder tree of the AUTO-go output**

Acknowledgements

We thank Francesca Nardozza for editorial assistance.

Author Contributions

Conceptualization: Matteo Pallocca, Eleonora Sperandio

Formal analysis: Matteo Pallocca, Eleonora Sperandio

Methodology: Matteo Pallocca, Eleonora Sperandio, Isabella Grassucci, Lorenzo D'Ambrosio

Software: Eleonora Sperandio

Supervision: Matteo Pallocca

Validation: Lorenzo D'Ambrosio, Isabella Grassucci

Writing – original draft: Matteo Pallocca

Writing – review & editing: Matteo Pallocca, Isabella Grassucci

Funding

This work has been supported by the Italian Ministry of Health (Ricerca Corrente 2020).

Conflict of Interest: none declared.

References

1. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics 2000 25:1*, *25*(1), 25–29. https://doi.org/10.1038/75556
2. Kulkarni, N., *et. al* (2018). Reproducible bioinformatics project: A community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics*, *19*(10), 5–13.
3. Dolinski, K., & Botstein, D. (2013). Automating the construction of gene ontologies. *Nature Biotechnology 2013 31:1*, *31*(1), 34–35. https://doi.org/10.1038/nbt.2476
4. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/NBT.3820
5. Köster, J., & Rahmann, S. (2018). Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England)*, *34*(20), 3600. https://doi.org/10.1093/BIOINFORMATICS/BTY350
6. Wu, T., *et. al* (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (New York, N.Y.)*, *2*(3).
7. Kuleshov, M. *et. al* (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90–W97.

8. Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., … Hahn, W. C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature 2009 462:7269*, *462*(7269), 108–112. https://doi.org/10.1038/nature08460
9. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/PNAS.0506580102
10. Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., Arachchi, H., Arora, A., Auman, J. T., Ayala, B., Baboud, J., Balasundaram, M., Balu, S., Barnabas, N., Bartlett, J., Bartlett, P., Bastian, B. C., Baylin, S. B., Behera, M., … Zou, L. (2015). Genomic Classification of Cutaneous Melanoma. *Cell*, *161*(7), 1681. https://doi.org/10.1016/J.CELL.2015.05.044
11. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/J.MOLCEL.2010.05.004
12. Seavey, C. N., Pobbati, A. v., Hallett, A., Ma, S., Reynolds, J. P., Kanai, R., Lamar, J. M., & Rubin, B. P. (2021). WWTR1(TAZ)-CAMTA1 gene fusion is sufficient to dysregulate YAP/TAZ signaling and drive epithelioid hemangioendothelioma tumorigenesis. *Genes and Development*, *35*(7), 512–527. https://doi.org/10.1101/GAD.348220.120/-/DC1

implementation

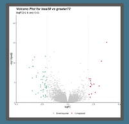
logical framework

visualization

DeSeq2

Differential Expression

volcano



DE filters: p-adj + Fold Change

Parse Gene List

<Gene List, Library>

EnrichR API

<Cluster, Enrichment Score>

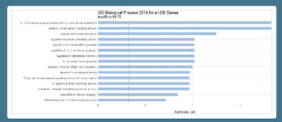
Parse Enrichment Tables

visualizations

Auto-GO

Gene Lists

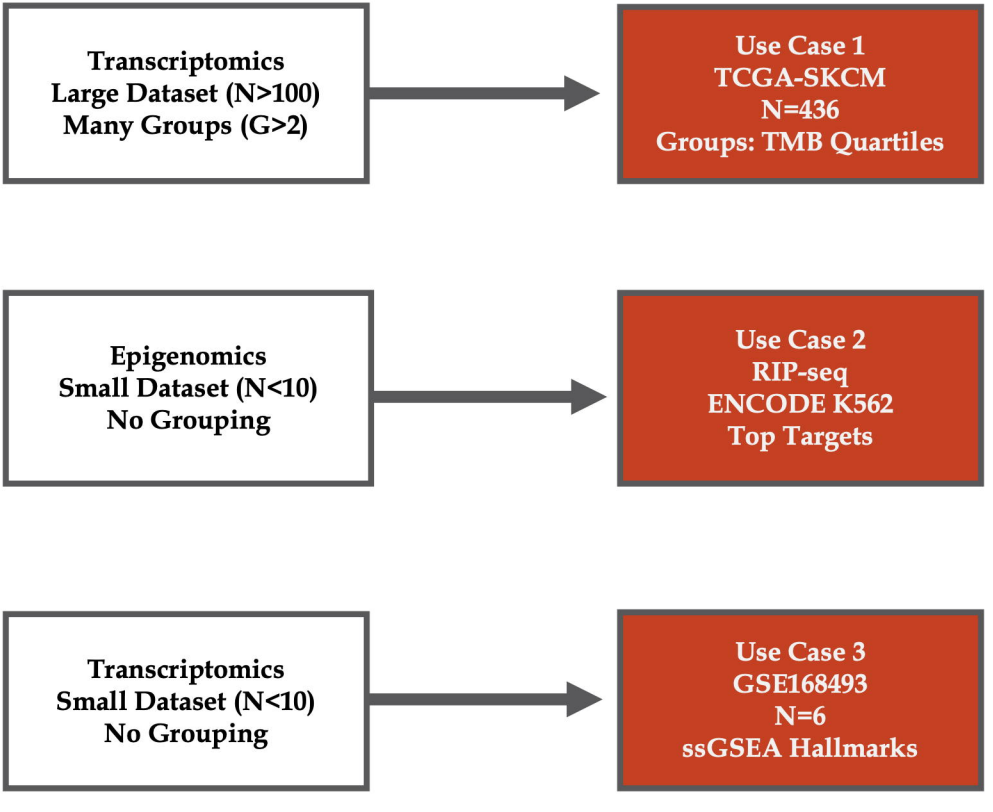
barplotGO



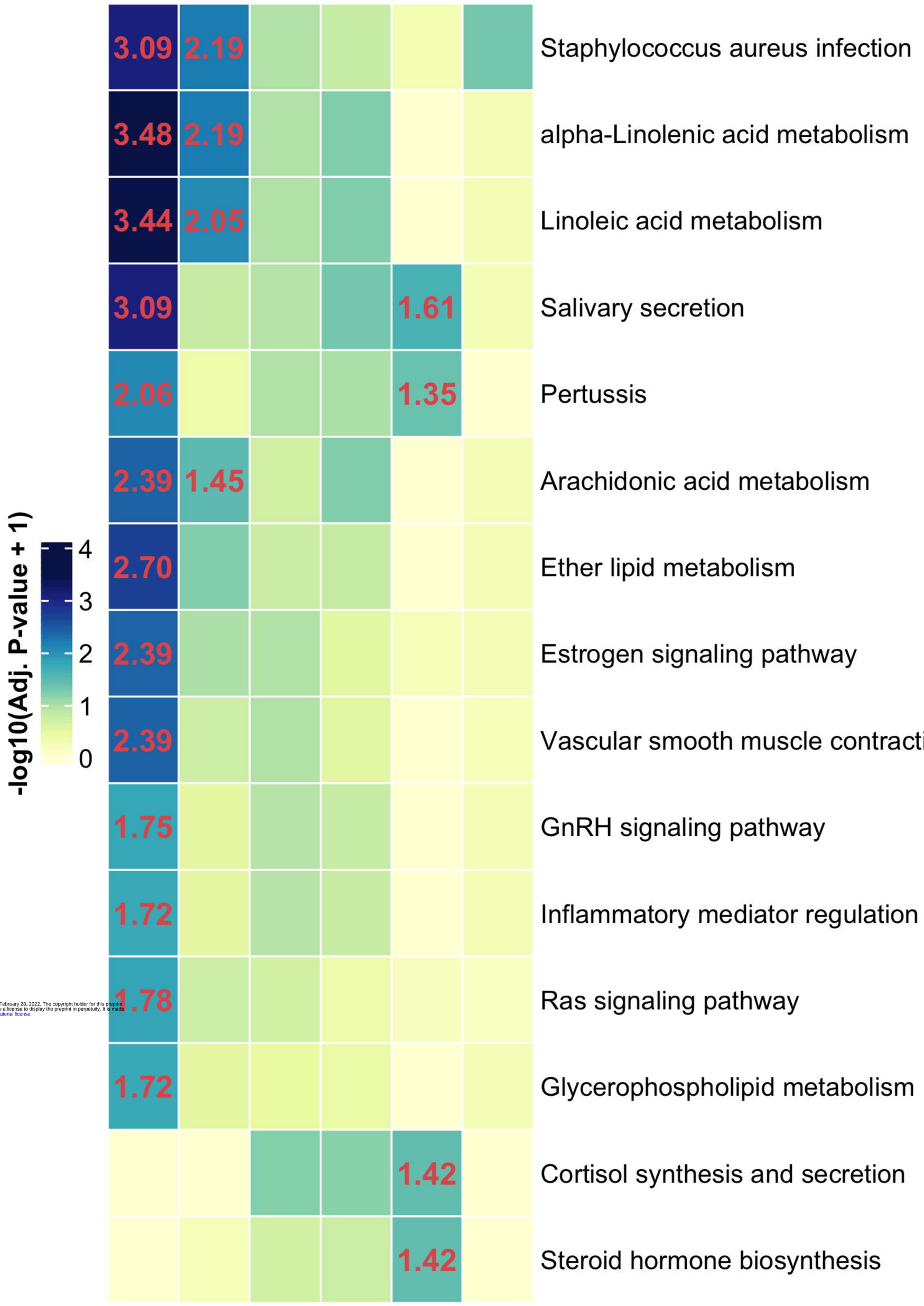
lolliGO

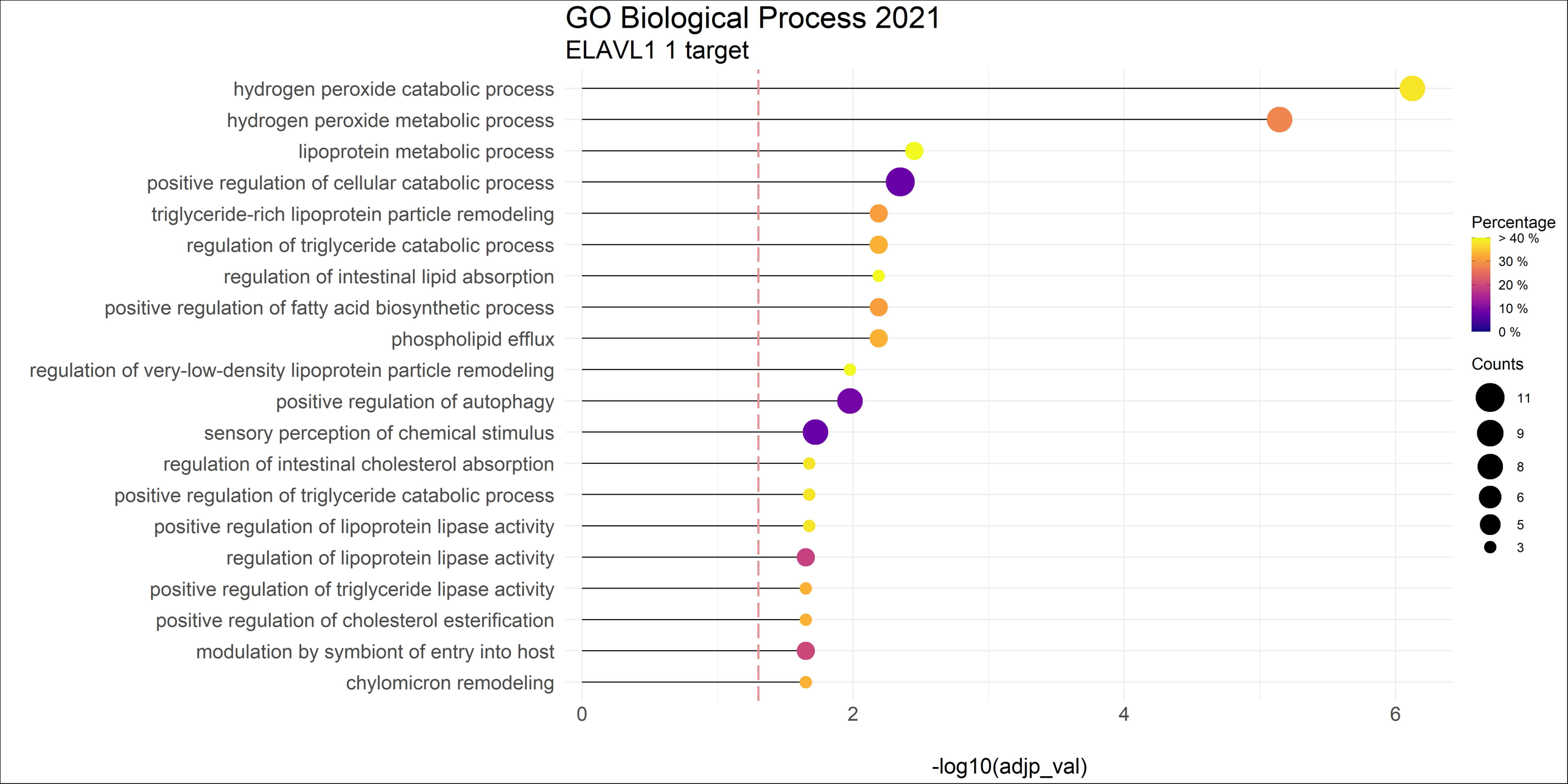
Enrichment

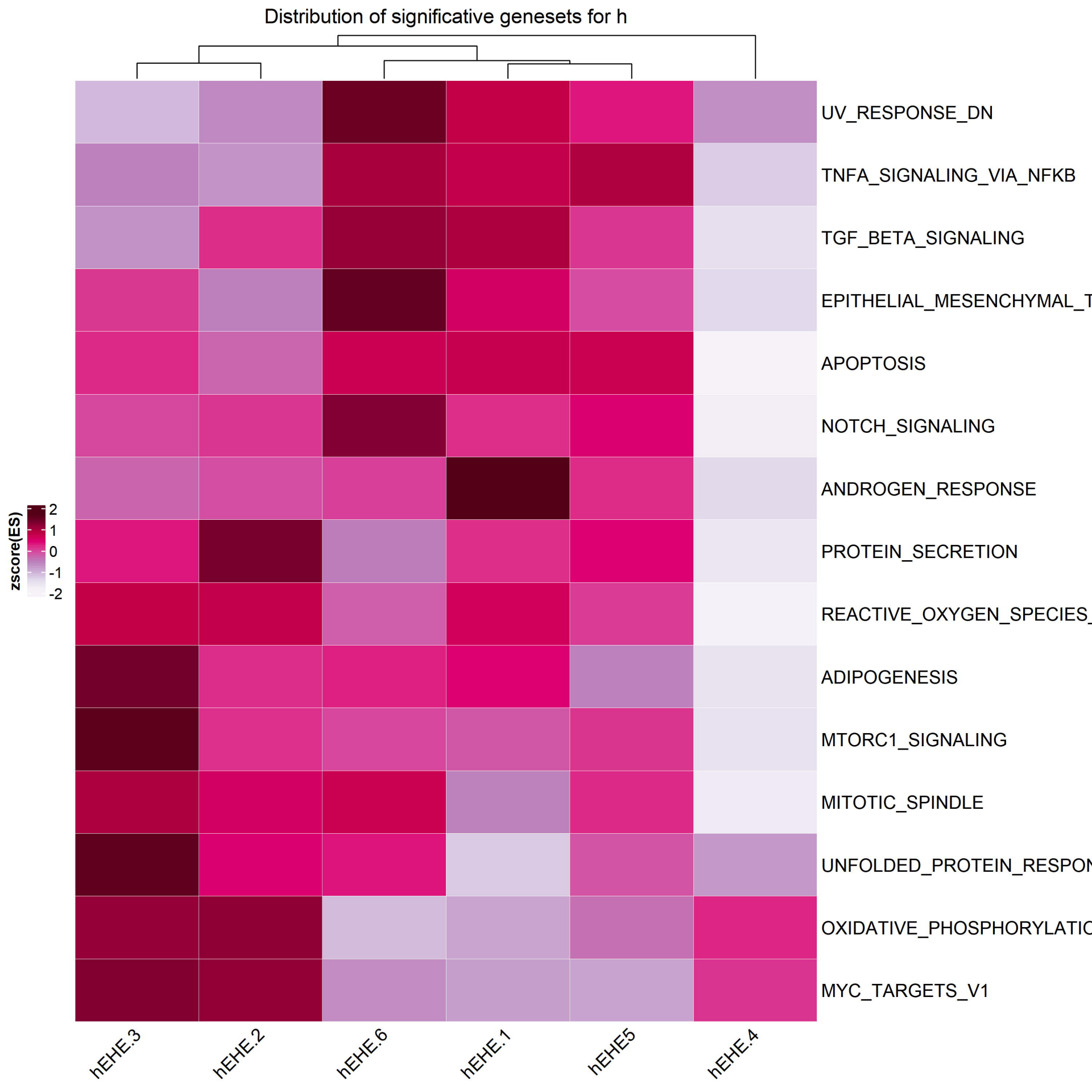
heatmapGO

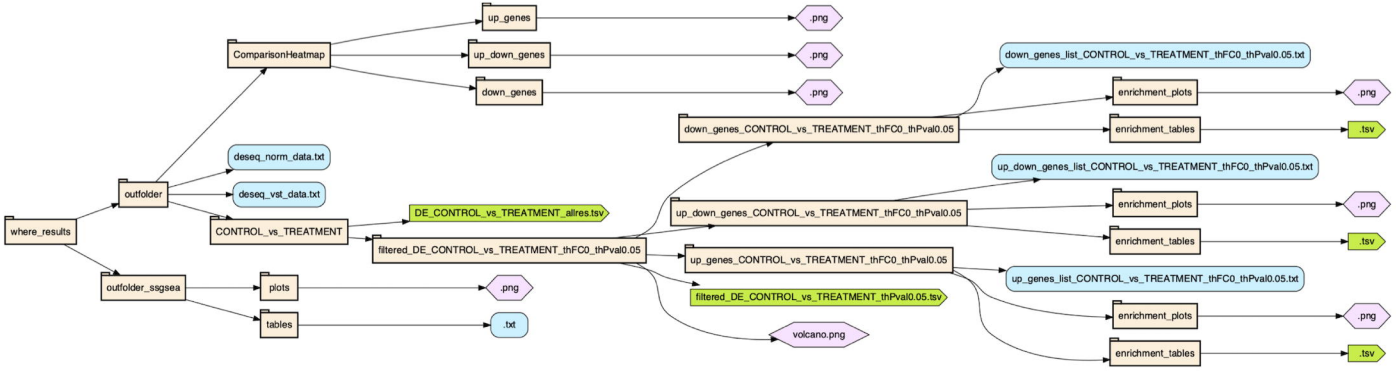

Transcriptomics
Large Dataset (N>100)
Many Groups (G>2)
Use Case 1
TCGA-SKCM
N=436
Groups: TMB Quartiles
Epigenomics
Small Dataset (N<10)
No Grouping
Use Case 2
RIP-seq
ENCODE K562
Top Targets
Transcriptomics
Small Dataset (N<10)
No Grouping
Use Case 3
GSE168493
N=6
ssGSEA Hallmarks

Multi-DE Ontology Heatmap for KEGG 2021 Human for all DE Genes




GO Biological Process 2021
ELAVL1 1 target
hydrogen peroxide catabolic process
hydrogen peroxide metabolic process
lipoprotein metabolic process
positive regulation of cellular catabolic process
triglyceride-rich lipoprotein particle remodeling
regulation of triglyceride catabolic process
regulation of intestinal lipid absorption
positive regulation of fatty acid biosynthetic process
phospholipid efflux
regulation of very-low-density lipoprotein particle remodeling
positive regulation of autophagy
sensory perception of chemical stimulus
regulation of intestinal cholesterol absorption
positive regulation of triglyceride catabolic process
positive regulation of lipoprotein lipase activity
regulation of lipoprotein lipase activity
positive regulation of triglyceride lipase activity
positive regulation of cholesterol esterification
modulation by symbiont of entry into host
chylomicron remodeling
0
2
4
6
-log10(adjp_val)
Percentage
> 40 %
30 %
20 %
10 %
0 %
Counts
11
9
8
6
5
3


Distribution of significative genesets for h
UV_RESPONSE_DN
TNFA_SIGNALING_VIA_NFKB
TGF_BETA_SIGNALING
EPITHELIAL_MESENCHYMAL_T
APOPTOSIS
NOTCH_SIGNALING
ANDROGEN_RESPONSE
PROTEIN_SECRETION
REACTIVE_OXYGEN_SPECIES_
ADIPOGENESIS
MTORC1_SIGNALING
MITOTIC_SPINDLE
UNFOLDED_PROTEIN_RESPON
OXIDATIVE_PHOSPHORYLATIC
MYC_TARGETS_V1
zscore(ES)
2
1
0
-1
-2
hEHE.3
hEHE.2
hEHE.6
hEHE.1
hEHE5
hEHE.4


where_results
outfolder
deseq_norm_data.txt
deseq_vst_data.txt
ComparisonHeatmap
up_genes
.png
up_down_genes
.png
down_genes
.png
CONTROL_vs_TREATMENT
DE_CONTROL_vs_TREATMENT_allres.tsv
filtered_DE_CONTROL_vs_TREATMENT_thFC0_thPval0.05
down_genes_CONTROL_vs_TREATMENT_thFC0_thPval0.05
down_genes_list_CONTROL_vs_TREATMENT_thFC0_thPval0.05.txt
enrichment_plots
.png
enrichment_tables
.tsv
up_down_genes_CONTROL_vs_TREATMENT_thFC0_thPval0.05
up_down_genes_list_CONTROL_vs_TREATMENT_thFC0_thPval0.05.txt
enrichment_plots
.png
enrichment_tables
.tsv
up_genes_CONTROL_vs_TREATMENT_thFC0_thPval0.05
up_genes_list_CONTROL_vs_TREATMENT_thFC0_thPval0.05.txt
enrichment_plots
.png
enrichment_tables
.tsv
filtered_DE_CONTROL_vs_TREATMENT_thPval0.05.tsv
volcano.png
outfolder_ssgsea
plots
.png
tables
.txt
A