Systems Biology

HiMaLAYAS: enrichment-based annotation of hierarchically clustered matrices

Ira Horecka[1,2] and Hannes Röst[1,2,3,*]

[1]Terrence Donnelly Centre for Cellular & Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada

[2]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada

[3]Department of Computer Science, University of Toronto, Toronto, Ontario M5S 2E4, Canada

*Corresponding author. Terrence Donnelly Centre for Cellular & Biomolecular Research, University of Toronto, 160 College Street, Room 604, Toronto, Ontario M5S 3E1, Canada. Tel: +1 (416) 978-7224. E-mail: hannes.rost@utoronto.ca (H.R.)

Abstract

Summary: Hierarchical clustering organizes high-dimensional biological matrices and is commonly used for visualization rather than statistical inference. Most enrichment-based analyses of hierarchically clustered matrices are confined to gene expression data and fixed workflows. We introduce Hierarchical Matrix Layout and Annotation Software (HiMaLAYAS), a framework for post hoc enrichment-based annotation of such matrices. HiMaLAYAS treats clusters as statistical units, tests annotation enrichment, and renders significant annotations alongside the matrix. HiMaLAYAS supports biological and non-biological domains.

Availability and Implementation: HiMaLAYAS is a Python package available via pip, distributed under the BSD 3-Clause License at https://github.com/himalayas-base/himalayas, and archived on Zenodo at https://doi.org/10.5281/zenodo.18610373.

1 Introduction

Hierarchical clustering organizes high-dimensional biological matrices by ordering rows and columns into contiguous regions of related observations while color encodes data. This enables visual interpretation across thousands of measurements and is widely adopted for exploratory analysis in functional genomics and systems biology (Eisen *et al*., 1998; Spellman *et al*., 1998).

Clustering creates a dendrogram that is typically cut to define clusters (Langfelder and Horvath, 2008). These clusters are often annotated as pathway membership tracks alongside the matrix (Gu *et al*., 2016). Although methods exist to assess cluster stability (Suzuki and Shimodaira, 2006), dendrogram-defined clusters are rarely used for statistical inference.

Previous methods that combine clustering with enrichment-based analysis were largely developed for gene expression data and use enrichment to guide cluster pruning in a fixed workflow (Krushevskaya *et al*., 2009; Kolberg *et al*., 2018). No framework exists for post hoc enrichment-based annotation of hierarchically clustered matrices.

We introduce Hierarchical Matrix Layout and Annotation Software (HiMaLAYAS), a framework for post hoc enrichment-based annotation of hierarchically clustered matrices across biological and non-biological domains. HiMaLAYAS treats dendrogram-defined clusters as statistical units, supports enrichment across dendrogram depths, and renders significant annotations alongside the matrix.

We apply HiMaLAYAS to *Saccharomyces cerevisiae* genetic interaction matrices from Costanzo *et al*. (2016), revealing hierarchical organization of biological processes. We show that enrichment depends on dendrogram depth and that HiMaLAYAS generalizes beyond biological data.

2 Methods

2.1 HiMaLAYAS and cluster definition

HiMaLAYAS hierarchically clusters a real-valued matrix capturing relationships among observations (Fig. 1A). Cluster size and number are defined by cutting the dendrogram at a user-defined depth, and undersized clusters are merged with their parent clusters, preserving hierarchy. These clusters are carried forward for enrichment.

2.2 Enrichment, multiple testing correction, and visualization

HiMaLAYAS systematically tests enrichment for categorical annotations across clusters (e.g., biological process terms associated with genes). It restricts annotations to observations in the matrix and applies a one-sided hypergeometric test to determine if a term's members are overrepresented among a cluster's observations. HiMaLAYAS controls multiple testing across cluster–term tests using the Benjamini–Hochberg false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995) and renders significant annotations alongside the matrix. In this study, we consider cluster–term enrichments significant at FDR ≤ 0.05.

3 Results

3.1 Enrichment-based annotation reveals hierarchical organization

We applied HiMaLAYAS to a hierarchically clustered *Saccharomyces cerevisiae* genetic interaction profile similarity matrix (Costanzo *et al*., 2016), focusing on ~1,100 genes with high profile variance (Fig. 1B). Dendrogram-defined clusters significantly enriched for Gene Ontology Biological Process (GO BP) terms (Ashburner *et al*., 2000) were annotated alongside the matrix, revealing a hierarchical partition between vesicle trafficking–related processes and DNA-associated functions.

3.2 Enrichment depends on dendrogram depth

We applied HiMaLAYAS to a hierarchically clustered yeast genetic interaction score matrix (Costanzo *et al*., 2016), focusing on ~1,100 genes with high score variance. Cutting the dendrogram at different depths revealed hierarchical GO BP enrichment (Supplementary Fig. S1). At a higher depth, clusters are enriched for broad processes such as DNA replication (P=2.99 × 10^{-38}), whereas at a lower depth they are enriched for subprocesses such as pre-replicative complex assembly (P=2.26 × 10^{-18}). Restricting the analysis to a single depth obscures this hierarchical organization.

3.3 Generalization beyond biological matrices

To demonstrate generality, we applied HiMaLAYAS to a non-biological matrix: a worldwide recipe dataset (Magomere *et al*., 2025), in which recipes are clustered by ingredient similarity and annotated

by country of origin (Supplementary Fig. S2). HiMaLAYAS reveals regional similarities among cuisines; for example, recipes enriched for Algeria (P=1.45 × 10^{-4}) cluster near those enriched for Egypt (P=1.42 × 10^{-3}), sharing staple ingredients such as flour and sugar.

4 Conclusion

HiMaLAYAS is a framework for post hoc enrichment-based annotation of hierarchically clustered matrices across biological and non-biological domains.

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

The authors declare no conflict of interest.

Funding

This work was supported by the Canada Research Chairs (CRC-2022-00215) and by the Canadian Institutes of Health Research (CIHR; grant 497277).

Data availability

The HiMaLAYAS documentation repository (https://github.com/himalayas-base/himalayas-docs) provides software documentation and Jupyter notebook tutorials. The HiMaLAYAS publication repository (https://github.com/himalayas-base/himalayas-publication) contains the datasets and workflows used to generate the figures in this study.

References

Ashburner M, Ball CA, Blake JA, *et al*. Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–29.

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

Costanzo M, VanderSluis B, Koch EN, *et al*. A global genetic interaction network maps a wiring diagram of cellular function. *Science* 2016;**353**:aaf1420.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci* 1998;**95**:14863–14868.

Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 2016;**32**:2847–2849.

Kolberg L, Kuzmin I, Adler P, Vilo J, Peterson H. funcExplorer: a tool for fast data-driven functional characterisation of high-throughput expression data. *BMC* Genomics 2018;**19**:817.

Krushevskaya D, Peterson H, Reimand J, Kull M, Vilo J. VisHiC—hierarchical functional enrichment analysis of microarray data. *Nucleic Acids Res* 2009;**37**:W587–W592.

Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559.
Magomere J, Ishida S, Afonja T, Salama A, Kochin D, Foutse Y, *et al*. The world wide recipe: a community-centred framework for fine-grained data collection and regional bias operationalisation. *Proc ACM Conf Fairness Accountability Transparency* 2025:246–282.
Spellman PT, Sherlock G, Zhang MQ, *et al*. Comprehensive identification of cell cycle–regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 1998;**9**:3273–3297.
Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 2006;**22**:1540–1542.

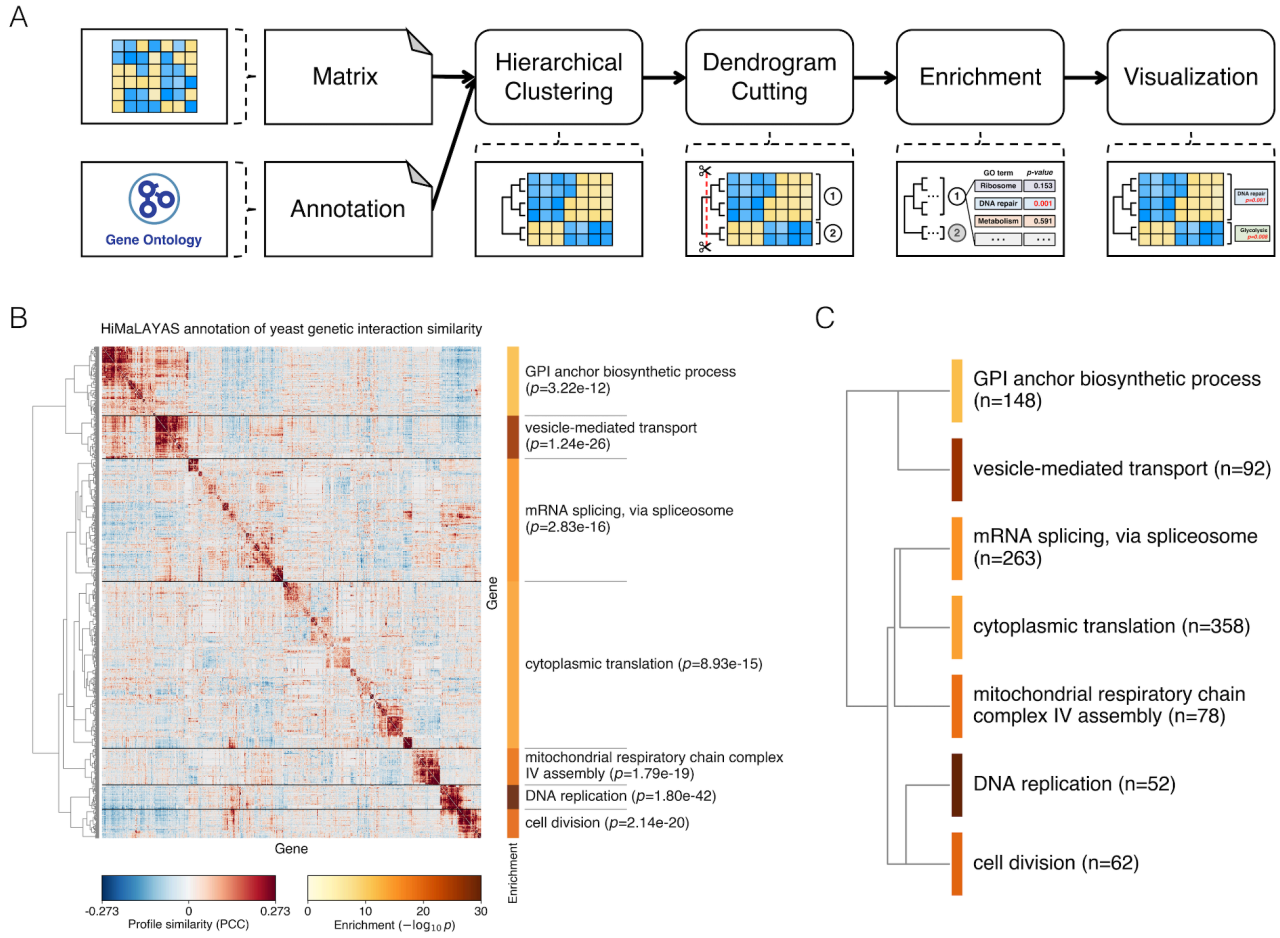
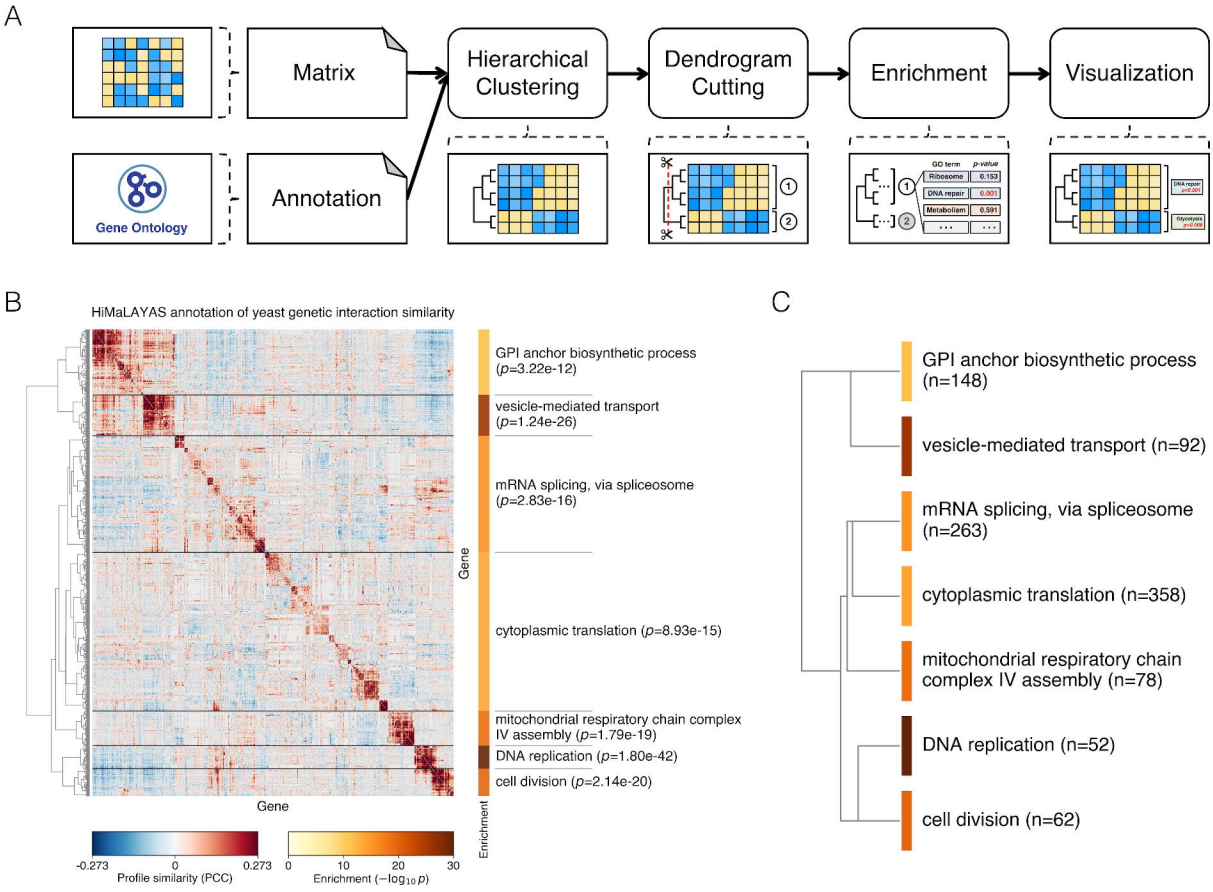


Figure 1. HiMaLAYAS workflow and its application to a hierarchically clustered *Saccharomyces cerevisiae* genetic interaction profile similarity matrix (Costanzo *et al*., 2016). (**A**) HiMaLAYAS workflow. A real-valued matrix and categorical annotations serve as inputs. HiMaLAYAS clusters the matrix to create a dendrogram, which is cut at a user-defined depth to identify clusters. Enrichment is tested for categorical annotations across dendrogram-defined clusters using an overrepresentation test, and significant annotations are rendered alongside the matrix. (**B**) Application of HiMaLAYAS to the clustered yeast genetic interaction profile similarity matrix, focusing on ~1,100 genes with high profile variance. Each entry represents two genes' similarity in their genetic interaction profiles measured by Pearson correlation. Dendrogram-defined clusters were tested for Gene Ontology Biological Process (GO BP; Ashburner *et al*., 2000) enrichment, with significant annotations rendered alongside the matrix. (**C**) Condensed dendrogram showing the same hierarchy.


A
Matrix
Hierarchical Clustering
Dendrogram Cutting
Enrichment
Visualization
Gene Ontology
Annotation
GO term
p-value
Ribosome
0.153
DNA repair
0.001
Metabolism
0.591
DNA repair p=0.001
Glycolysis p=0.008
B
HiMaLAYAS annotation of yeast genetic interaction similarity
GPI anchor biosynthetic process (p=3.22e-12)
vesicle-mediated transport (p=1.24e-26)
mRNA splicing, via spliceosome (p=2.83e-16)
cytoplasmic translation (p=8.93e-15)
mitochondrial respiratory chain complex IV assembly (p=1.79e-19)
DNA replication (p=1.80e-42)
cell division (p=2.14e-20)
Gene
Gene
Enrichment
-0.273
0
0.273
Profile similarity (PCC)
0
10
20
30
Enrichment (−log10 p)
C
GPI anchor biosynthetic process (n=148)
vesicle-mediated transport (n=92)
mRNA splicing, via spliceosome (n=263)
cytoplasmic translation (n=358)
mitochondrial respiratory chain complex IV assembly (n=78)
DNA replication (n=52)
cell division (n=62)