Integrated visualization of metabolomics and transcriptomics with Galaxy

Michele Ferrari[1]₊, Francesco Lapi[1]₊, Lucrezia Penati[2], Marco Vanoni[1], Bruno Galuzzi[3], Chiara Damiani[1,3]*

[1] Dept. of Biotechnology and Biosciences, Universitá degli Studi di Milano-Bicocca, Milan, Italy.
[2] Dept. of Informatics, Systems and Communication, Universitá degli Studi di Milano-Bicocca, Milan, Italy.
[3] Institute of bioimaging and complex biological system (IBSBC), Milan, Italy.

*corresponding author
chiara.damiani@unimib.it, ORCID code 0000-0002-5742-8302

₊ equal contributors

Keywords: RNA-seq, metabolomics, Galaxy, enrichment.

Abstract. The regulation of cell metabolism is complex and multifold. Hence the metabolic alterations that have been reported in many physio-pathological conditions can be fully characterized only by using model-based multi-omics data integration frameworks. We present here version 2 of the Marea4Galaxy tool, integrated into the Galaxy platform. The previous version of Marea4Galaxy allowed users to visualize deregulated reactions at the transcriptomic level. The new version extends these capabilities by enabling the simultaneous visualization of deregulated reactions at the metabolic level using metabolomics data. Significant improvements have been made, including a more comprehensive metabolic network model, a module for extracting necessary inputs from any metabolic model in XML or JSON format, better compatibility with alternative gene nomenclatures, and faster reaction activity scores (RASs) calculation. We demonstrate the utility of this tool by comparing different groups of cancer cell lines using paired datasets from the Cancer Cell Line Encyclopedia.

1 Introduction

Metabolism is regulated by the complex interaction between the availability of reaction substrates and the activity of enzymes. Enzyme activity can be in turn controlled either at the transcriptional, post-transcriptional, or post-translational level. Enzymatic regulation can indirectly affect the availability of substrates of neighbor reactions in the metabolic network. Hence metabolism cannot be fully understood by analyzing -omics data alone, but only by using innovative multi-omics data integration frameworks rooted in data science and computational systems biology.

The increasing interest in the characterization of metabolic alterations in physio-pathological conditions has driven the demand for user-friendly tools that provide life scientists with an effective overview of the multi-level deregulation of metabolic pathways. Researchers and clinicians seek an intuitive yet detailed global visualization of metabolic alterations, avoiding overly complex statistical methods in favor of data-centric and organized insights. Given the growing trend among research and clinical groups to collect paired transcriptomics and metabolomics data[1, 2], an integrated view has become increasingly valuable.

Galaxy[3] is a user-friendly, web-based workflow system designed to enable biomedical researchers to utilize computational biology tools without needing advanced computer science skills. It provides an accessible platform for users with varying levels of technical expertise to create, execute, and share complex bioinformatics analyses through an intuitive interface. Galaxy integrates a wide range of tools and supports external resources, promoting the reproducibility and sharing of scientific results.

Current pathway enrichment frameworks leveraging gene expression data, such as GSEA ([4, 5]) exhibit limitations, lacking expressiveness in flux direction and proliferation rate indication. They might also exhibit biases towards pathways with numerous isoforms or subunits.

The Metabolic Reaction Enrichment and Analysis tool for Galaxy (Marea4Galaxy) [6], which focuses on transcriptomics data, exploiting the Gene Reaction Protein associations rules to aggregate genes at the reaction-level has achieved notable success, as evidenced by the number of users on our server.

To account that the metabolic flux through a reaction is influenced both by the enzyme (transcript) and the substrate, Marea4Galaxy 2.0 allows the user to compute both transcriptomics-based Reaction Activity Scores (RASs) and metabolomics-based Reaction Propensity Scores (RPSs) as proposed in [7, 8].

Given that RPSs are based on substrate abundance, reversible reactions will display a distinct score for the forward and backward directions. This information well complements the information provided by RAS that, being based on enzyme expression, are non-directional.

We here define a net RPS score that expresses the favorite direction usage. When comparing different conditions, a reaction-based statistical test is performed for the net RPS score and the RAS score. To offer an integrated view, MaREA 2.0 features a visualization module that maps statistically significant deregulations on an SVG map of the human metabolic network. Each metabolic reaction is represented as an arrow connecting substrate and products. The color and thickness of the arrow shaft are set according to the RAS variation, whereas the arrow tip is colored according to the RPS variation. When a reaction is reversible only the most meaningful direction is colored.

To capture more metabolic pathways we have created and included a more comprehensive SVG map based on the ENGRO2 model presented in [8]. To sustain interoperability, we included a dictionary of the main used synonyms for metabolite and gene names.

2 Release information

This paper marks the latest stable release of the MaREA toolset (version 2.0).
Review source code and documentation in the Galaxy toolshed:
https://bimib@toolshed.g2.bx.psu.edu/repos/bimib/marea_2_0. A “MaREA4Galaxy” demo is available at http://marea4galaxy.cloud.ba.infn.it/galaxy/.

3 New tools & functionalities

3.1 *Reaction Propensity Scores generator*

A Reaction Propensity Scores (RPS) consists in a reaction score computed as the product of the concentrations of the reacting substances, with each concentration raised to a power equal to its stoichiometric coefficient. According to the mass action law, the rate of any chemical reaction is indeed proportional to this product. This assumption holds as long as the substrate is in significant excess over the enzyme constant KM. If the reaction is reversible, we defined the net RPS as difference between the forward and the backward reaction.

The tool loads the intracellular metabolomics provided by the user, and data information about reactions in the selected model. Then it computes one RPS for each of them.

Each reaction is identified by a name, unique in the selected model, and contains data about all the metabolites acting as substrates in the reaction. It is worth mentioning that reversible reactions are split into two distinct irreversible reactions, corresponding to backward and forward, and treated as different reactions.

The metabolomics dataset input by the user is split by cell-line (or sample) and queried by each reaction in order to retrieve the abundance measured for each metabolite. In order to maximize the tool's ability to recognize metabolite names the dataset undergoes a number of pre-processing steps, most notably information about common synonyms for each metabolite in the selected model is loaded from a local dictionary created using the "Human Metabolome DataBase" [9].

Some common metabolites that tend to appear in a large number of reactions have been manually blacklisted, that is they won't be considered in the RPS computation of a reaction that contains them. We believe this helps making the computed scores statistically more significant, as a high concentration of a metabolite as pervasive as ATP would otherwise increase the scores of most reactions without being informative.

3.2 *Custom data generator*

This tool allows users to quickly obtain all the auxiliary input files needed by the RAS and RPS tools to work with custom models. Starting from the model itself (any .xml or .json that can be interpreted as a metabolic model by the cobrapy (https://opencobra.github.io/cobrapy/ package) this tool will extract custom gene reaction rules and reaction formulas from it, with the option of getting them in a ready-to-go pre-parsed and optimized form.

4 Updates to existing tools

A considerable amount of time and effort was put into refactoring, documenting and organizing the project as a whole, with the addition of internal testing and utility packages. The process also brought about changes in the "RAS generator" tool, which can now work faster due to a new, more robust and versatile rule parser. Details about the parser implementation and the new rules themselves can be provided as supplementary material. The parser is located in a shared utility package, which means that the same rules will be processed in the exact same way by the Custom Data Generator tool: consistency is guaranteed independently from the user's preferred workflow.

The MaREA enrichment and visualization tool now supports both RAS and RPS datasets, allowing for a more nuanced multi-omics comparison between samples or conditions. While the baseline enrichment algorithm remains the same for RAS datasets, the interaction between the two scores necessarily introduced some changes that apply when RPS datasets are involved. Moreover, a new enrichment algorithm was implemented specifically for RPS datasets and works alongside the first one. The specifics of how exactly a metabolic map is enriched by the MaREA tool with different types of datasets are described in section 5.1.

We also updated the default metabolic model to the recently published metabolic network model "ENGRO2" [8] and provided a graphical map of it. "ENGRO2" is a constraint-based generic core model of human central carbon and essential amino acids metabolism. It contains 484 reactions, 403 metabolites, and 494 genes and represents a follow-up of the HMR_{core} model. 337 model reactions are associated with a gene-protein-reaction (GPR) rule. More in detail, there are 202 single-gene GPRs, 122 OR-expression, 36 AND-expression, and 23 complex rules (i.e., logical expression with both AND or OR operator). The new map highlights broad pathways and reaction groups to improve readability.

5 Technical implementation

5.1 *Net RPS-based enrichment*

The Metabolic Reaction Enrichment Analysis (MaREA) tool can edit various visual properties of the arrows representing reactions in the selected/provided metabolic map which means that a customized map can be created rather freely, as long as the arrows can be recognized by the program. Notably, the “body” or “shaft” of each arrow must be distinct from its “head”(s) or “tip”(s) and possess an “id” parameter equal to the reaction’s name in the provided datasets but prefixed with “R_” (as in “reaction”). The same applies for the arrow’s heads, which need to share the “F_” prefix (as in “forward”) if they point towards the products or the “B_” prefix (as in “backward”) if they point towards the substrates.

While the arrow heads did not matter in the previous enrichment implementation they now offer an interesting opportunity to showcase both RAS and RPS data in the same map: this is because RAS data is not directional and thus only affects arrow bodies, instead the propensity scores treat reversible reactions as distinct halves and encode the direction of each half in the reaction ID. As such, whenever the user decides to also provide RPS datasets these scores will be compared on a per-dataset and per-reaction basis across all samples in the usual way and will affect arrow tips of the corresponding direction.

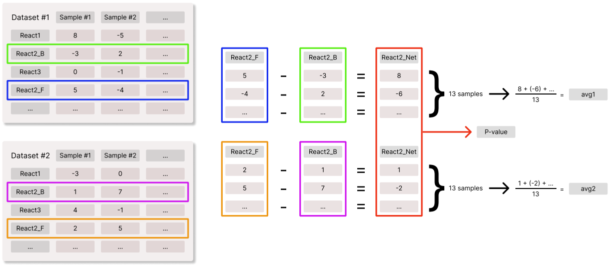
If the user decides to only provide RPS datasets as input the arrow bodies of reversible reactions will be styled based on a ”net” RPS comparison, as shown in Figure 1:

- Each dataset maps reaction IDs to their corresponding list of scores, ordered by sample/cell-line. Since the RPS module treats the two sides of a reversible reaction as distinct reactions, we first need to aggregate the two separate RPSs into a net score. The list of net scores for each dataset is obtained from element-wise subtraction between the two separate lists;

- Datasets are always compared in pairs on a per-reaction basis. Therefore, for each pairwise comparison, we will have two net lists, one for each of the two datasets. From these, we obtain the P-value for that pair and an average value for each of the datasets under comparison, respectively named avg1 and avg2;

- Lastly, the following equation computes the final comparison score from the two averages avg1 and avg2, as follows:

$$\frac{\text{avg1} - \text{avg2}}{|\text{avg2}|}. \tag{1}$$



Dataset #1	Sample #1	Sample #2	...
React1	8	-5	...
React2_B	-3	2	...
React3	0	-1	...
React2_F	5	-4	...
...	...	...	...

Dataset #2	Sample #1	Sample #2	...
React1	-3	0	...
React2_B	1	7	...
React3	4	-1	...
React2_F	2	5	...
...	...	...	...

Figure 1: An example of the net enrichment algorithm.

This same algorithm can also be employed for arrow tips if the user wishes it so. In both cases the averages that created each final comparison score are also saved and contribute to appropriately styling the metabolic map based on the relationship between their *signs*: if the signs of the two averages are opposite the applied color will be a different shade of red/blue as a note to this sign disparity. When the net comparison is applied to the arrow tips as well only the "forward" tips will usually be styled, notably however if both averages are negative the sign of the final comparison score is inverted and said score applied to the "backward" tips only.

5.2 *Integration with the Galaxy framework*

It is customary for Galaxy tools to receive their input arguments from the UI as command-line arguments under specific names. In an attempt to enforce stricter type-checking and improve python's ability to statically analyze the code various custom types and enumerators have been implemented, and they work to keep the code cleaner and more maintainable.

6 Intended workflow

6.1 *Marea in action*

To demonstrate the intended workflow of the MaREA4Galaxy toolset, we computed a RAS (with the updated RAS generator tool) and a RPS (with the new RPS generator module) for each cell line within the Cancer Cell Line Encyclopedia[10] for which gene expression and metabolite abundance data were available. To identify groups of cells with distinct metabolic profiles, we first used the Cluster Analysis tool, using the RPS as features. We then used the MaREA tool to perform a differential RPS and RAS analysis and to visualize results on the integrated map. The galaxy workflow and a portion of the generated map are illustrated in Figure 2.

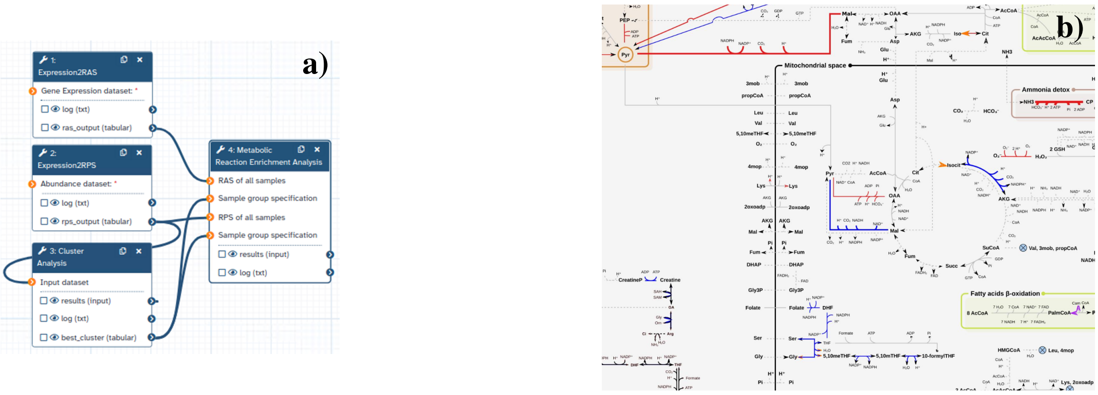


Figure 2: a) example of Galaxy workflow using MaRea4Galaxy b) a snippet of the ENGRO2 map visualizing RPS and/or RAS enriched reactions for two clusters of cancer cell lines. Upregulated reactions are red-colored, downregulated blue-colored. Orange and purple are used to highlight up and downregulated reactions, respectively, that display a net RPS with opposite sign in the two groups. Arrow thickness is proportional to fold change.

7 Advantages and limitations

The introduction of a set of standards to abide to, pushing for extensive documentation and type safety and the creation of testing and utility modules has greatly improved the readability, maintainability and robustness of the code, including everything we will write in the future, which will be able to ship much faster and more reliably. This release already includes two new tools and many new features for the old ones, allowing the user to expand the scope of their analysis to include other omics, potentially also obtaining faster results. That being said, some of these features are sadly left underused or unutilized due to the many integration difficulties with the old code, which could lead to many small bugs. Most of these issues don't concern the users given that every tool is working as intended, but work will be done in future releases of the toolset to ensure best practices moving forward.

Directional enrichment with RPS data opens the door for an even more comprehensive approach including flux balance analysis and gene sets enrichment that could be performed by additional tools: this is where the project is headed next.

8 Conclusion

The metabolic flux through a reaction is influenced both by the enzyme and the substrate. Therefore, instead of a general enrichment analysis Marea4Galaxy 2.0 characterizes metabolic pathways on a reaction-by-reaction basis, generating both a metabolic score and a transcriptomics score. This dual-level analysis provides a more comprehensive understanding of metabolic pathway deregulation, enhancing the utility of the Galaxy platform for multi-omics data analysis.

Funding

This work has been supported by ELIXIRxNextGenerationIT" (Code IR0000010)-CUP B53C22001800006 and PRIN PNRR (CUP H53D2300768 0001).

References

[1] Maria A Wörheide, Jan Krumsiek, Gabi Kastenmüller, and Matthias Arnold. Multi-omics integration in biomedical research–a metabolomics-centric review. *Analytica chimica acta*, 1141:144–162, 2021.

[2] Kiran Maan, Ruchi Baghel, Seema Dhariwal, Apoorva Sharma, Radhika Bakhshi, and Poonam Rana. Metabolomics and transcriptomics based multi-omics integration reveals radiation-induced altered pathway networking and underlying mechanism. *NPJ Systems Biology and Applications*, 9(1):42, 2023.

[3] The galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Research*, page gkae410, 2024.

[4] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[5] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, et al. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267–273, 2003.

[6] Chiara Damiani, Lorenzo Rovida, Davide Maspero, Irene Sala, Luca Rosato, Marzia Di Filippo, Dario Pescini, Alex Graudenzi, Marco Antoniotti, and Giancarlo Mauri. Marea4galaxy: Metabolic reaction enrichment analysis and visualization of rna-seq data within galaxy. *Computational and structural biotechnology journal*, 18:993–999, 2020.

[7] Alex Graudenzi, Davide Maspero, Marzia Di Filippo, Marco Gnugnoli, Claudio Isella, Giancarlo Mauri, Enzo Medico, Marco Antoniotti, and Chiara Damiani. Integration of transcriptomic data and metabolic networks in cancer samples reveals highly significant prognostic power. *Journal of biomedical informatics*, 87:37–49, 2018.

[8] Marzia Di Filippo, Dario Pescini, Bruno Giovanni Galuzzi, Marcella Bonanomi, Daniela Gaglio, Eleonora Mangano, Clarissa Consolandi, Lilia Alberghina, Marco Vanoni, and Chiara Damiani. Integrate: Model-based multi-omics data integration to characterize multi-level metabolic regulation. *PLoS computational biology*, 18(2):e1009337, 2022.

[9] David S Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, Chris Fung, Lisa Nikolai, Mike Lewis, Marie-Aude Coutouly, Ian Forsythe, Peter Tang, Savita Shrivastava, Kevin Jeroncic, Paul Stothard, Godwin Amegbey, David Block, David D Hau, James Wagner, Jessica Miniaci, Melisa Clements, Mulu Gebremedhin, Natalie Guo, Ying Zhang, Gavin E Duggan, Glen D Macinnis, Alim M Weljie, Reza Dowlatabadi, Fiona Bamforth, Derrick Clive, Russ Greiner, Liang Li, Tom Marrie, Brian D Sykes, Hans J Vogel, and Lori Querengesser. HMDB: The human metabolome database. *Nucleic Acids Res.*, 35(Database issue):D521–6, January 2007.

[10] Haoxin Li, Shaoyang Ning, Mahmoud Ghandi, Gregory V Kryukov, Shuba Gopal, Amy Deik, Amanda Souza, Kerry Pierce, Paula Keskula, Desiree Hernandez, et al. The landscape of cancer cell line metabolism. *Nature medicine*, 25(5):850–860, 2019.