Optimized Python library for reconstruction of ensemble-based gene co-expression networks using multi-GPU: Supplementary Information

1. APPLICATION TO THE STUDY OF PATIENTS AFTER ALLOGENEIC STEM CELL TRANSPLANTATION SUFFERING FROM INVASIVE ASPERGILLOSIS

The two real datasets described in Section 3.3.3 of the manuscript have been used to demonstrate the biological utility of pyEnGNet. These datasets, named control and invasive, are associated with patients undergoing allogeneic stem cell transplantation and patients with invasive aspergillosis, respectively [1]. *Aspergillus* is a ubiquitous, saprophytic filamentous fungus that forms and releases conidia that are airbone and inhale daily. After inhalation of the conidia their germination is prevented by the mucosal barrier and alveolar macrophagues. If the fungus is not effectively killed, the conidia germinate and form invasive hyphae that can penetrate lung tissue and spread by vascular invasion. Physiological protection against *Aspergillus* depends on a highly coordinated interaction between innate and adaptative inmunity. Recognition of fungi through pathogen-associated molecular pattern receptors leads to complement activation, phagocytosis and fungal destruction. T-helper or CD4+ cells, specific in the immune response against fungi, produce cytokines such as interferon and interleukin 17 that increase macrophages activation and neutrophil recruitment [2]. Genetic or drug-induced defects in components of antifugical immunity increase the risk of invasive aspergillosis following chemotherapy or hematopoietic stem cell transplantation [1].

The primary purpose of this experiment is to compare the immune system responses of the two patients groups. Several research have shown that invasion by *Aspergillus* leads to changes in the expression of important genes for host defense related to innate immune system [3, 4]. PyEnGNet generated three gene co-expression networks for each of these groups by applying distinct threshold values. The threshold for hub nodes has been maintained at 3 across all gene co-expression networks. A spectrum from a moderate correlation (0.7) to a very strong correlation (0.9) has been applied to the other threshold values. Therefore, the chosen threshold values are 0.7, 0.8, and 0.9. As shown in Table A9 of the manuscript, information on the total number of nodes and edges for each of the generated networks has been obtained. In addition, the number of nodes and edges that comprise the network's central cluster have been determined. These final two values will assist us in determining which network is best suited for biological analysis. The dimensionality of networks generated with thresholds equal to 0.9 is insufficient to conduct an exhaustive ontological analysis. In networks with thresholds of 0.8, 46.5% of the genes are in the central cluster, while 82% of the total number of nodes are in the central cluster for networks with thresholds of 0.7. Therefore, networks generated by pyEnGNet with thresholds equal to 0.7 are excellent candidates for further biological analysis.

Our approach is rooted in the understanding that biological networks, including those representing gene coexpression, exhibit a scale-free topology, a principle extensively documented by Barabási [5]. This characteristic implies a non-uniform distribution of connectivity, with a few nodes (hubs) being highly connected and most nodes having fewer connections. In aiming to replicate this intrinsic property of biological networks, our heuristic pruning strategically simplifies the network. This methodological choice precedes clustering to ensure that the resulting subnetworks are reflective of the most crucial biological processes in the context of invasive aspergillosis. By applying clustering to a pruned network that approximates a scale-free topology, we aim to identify dense interconnections that are likely to represent meaningful biological mechanisms. This approach aligns with our objective to elucidate the underlying immune system responses in patients suffering from invasive aspergillosis post-allogeneic stem cell transplantation, leveraging the biologically pertinent structure of the pruned network for detailed analysis.

Then, a clustering evaluation was conducted on the two candidate gene networks. This

evaluation permits the identification of genes that are frequently involved in the same biological process and are densely interconnected [6]. Community clustering (GLay) [7] implemented through Cytoscape's ClusterMaker application [8] was selected as the clustering method. This evaluation of clustering produced 20 clusters for the control group and 15 clusters for the invasive group. Each cluster contains a minimum of ten genes. Once the clusters have been generated, the aim is to keep those clusters that contain immune-related information in their genes. Therefore, as shown in Figure S1, for the control group we obtain a total of 4 clusters, while for the invasive group only one cluster contains information regarding the objective of this experiment.

(a) Inferred clusters in network for control.

(b) Inferred clusters in network invasive aspergillosis.

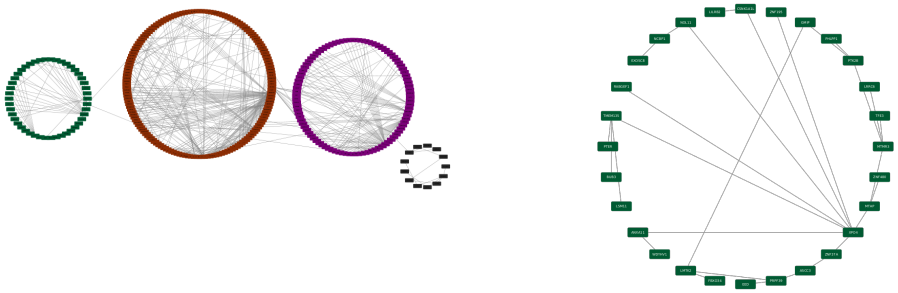

Fig. S1. This figure includes two inferred gene networks with GO information on the immune system. The network on the left shows the clusters of the inferred network for the control group, while those related to invasive aspergillosis are shown on the right.

Gene ontology specifies the relationship between genes by annotating and classifying the molecular function of a gene product and the corresponding biological process [9]. This method enables an analysis of gene enrichment for immune system information clusters extracted from the gene networks of the control and invasive groups. This enrichment analysis was conducted using ClusterProfiler [10], which uses the Gene Ontology database [11] as a gene annotation resource. This database sorts and describes genes and their products according to their molecular functions, biological processes, and cellular components. In the ClusterProfiler application, species (*Homo sapiens*) and biological process were selected as the specific ontology for this experiment. Additionally, Benjamini-Hochberg was used as the adjusted method, and a threshold was established so that only GO terms with a p-value less than 0.05 were generated. In this study, a GO term is a descriptor for a particular biological process associated with a gene. For the control group, a total of 215 GO terms were detected for all clusters involved, while 6 GO terms were detected for the invasive group. Gene information on the most representative GO terms for these two groups can be found in Table A10 in the Appendix manuscript.

The enrichment results of the control clusters show that there are clusters closely related to immune system activation, specifically B lymphocyte differentiation and defense against fungi. Conversely, a single cluster is observed in the invasive group related to the immune system response, where the associated GO terms are related to increased T-lymphocyte tolerance.

Several mechanisms of immunity are often impaired in oncohaematological patients, which justifies the high risk of fungal infection observed in these patients. As can be seen from the different GO terms found for control patients and patients with invasive aspergillosis, there is a clear difference in the mobilization of immune system cells, with a greater role in controls than in infected patients, where there is virtually no evidence of a normal immune response. This clear difference can also be seen in Figure S1, where there are more clusters related to the immune system, 4 in control group versus 1 in invasive group, and how the 4 control clusters have genes that allow the connection between the different study clusters.

The ontological analysis derived from the representative graph of the samples showing invasive aspergillosis is clearly consistent with a loss of efficiency of the immune system in these cases, including depressed T-lymphocyte action, which, as explained above, has an important action in cases of fungal infection. The ontological analysis derived from the representative graph of samples without invasive aspergillosis is consistent with the expected functioning of the immune system in normal cases. It can be detected in the activation of B lymphocytes, which is directly related to the activation of T helper lymphocytes, the fungal-specific response, and the

immune response-regulating cell surface receptor signalling pathway [12]. With these results, it is possible to argue that pyEnGNet is an useful tool in order to study biological processes using gene networks.

REFERENCES

1. D. P. Kontoyiannis, K. A. Marr, B. J. Park, *et al.*, "Prospective surveillance for invasive fungal infections in hematopoietic stem cell transplant recipients, 2001-2006: overview of the transplant-associated infection surveillance network (transnet) database," Clin. Infect. Dis. **50**, 1091–1100 (2010).
2. V. Balloy and M. Chignard, "The innate immune response to aspergillus fumigatus," Microbes Infect. **11**, 919–927 (2009).
3. K. J. Cortez, C. A. Lyman, S. Kottilil, *et al.*, "Functional genomics of innate host defense molecules in normal human monocytes in response to aspergillus fumigatus," Infect. Immun. **74**, 2353–2365 (2006).
4. S. Heldt, J. Prattes, S. Eigl, *et al.*, "Diagnosis of invasive aspergillosis in hematological malignancy patients: Performance of cytokines, asp lfd, and aspergillus pcr in same day blood and bronchoalveolar lavage samples," The J. Infect. **77**, 235–241 (2018).
5. A.-L. Barabási, "Scale-free networks: a decade and beyond," science **325**, 412–413 (2009).
6. W. Li, M. Wang, J. Sun, *et al.*, "Gene co-opening network deciphers gene functional relationships," Mol. Biosyst. **13**, 2428–2439 (2017).
7. G. Su, A. Kuchinsky, J. H. Morris, *et al.*, "Glay: community structure analysis of biological networks," Bioinformatics **26**, 3135–3137 (2010).
8. J. H. Morris, L. Apeltsin, A. M. Newman, *et al.*, "clustermaker: a multi-algorithm clustering plugin for cytoscape," BMC bioinformatics **12**, 1–14 (2011).
9. M. Ashburner, C. A. Ball, J. A. Blake, *et al.*, "Gene ontology: tool for the unification of biology," Nat. genetics **25**, 25–29 (2000).
10. G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterprofiler: an r package for comparing biological themes among gene clusters," OMICS: A J. Integr. Biol. **16**, 284–287 (2012).
11. Gene Ontology Consortium, "Gene ontology consortium: going forward," Nucleic Acids Res. **43**, D1049–D1056 (2015).
12. H. Hebart, C. Bollinger, P. Fisch, *et al.*, "Analysis of t-cell responses to aspergillus fumigatus antigens in healthy individuals and patients with hematologic malignancies," Blood, The J. Am. Soc. Hematol. **100**, 4521–4528 (2002).