Ensemble and Greedy Approach for the inference of Large Gene Co-Expression Networks

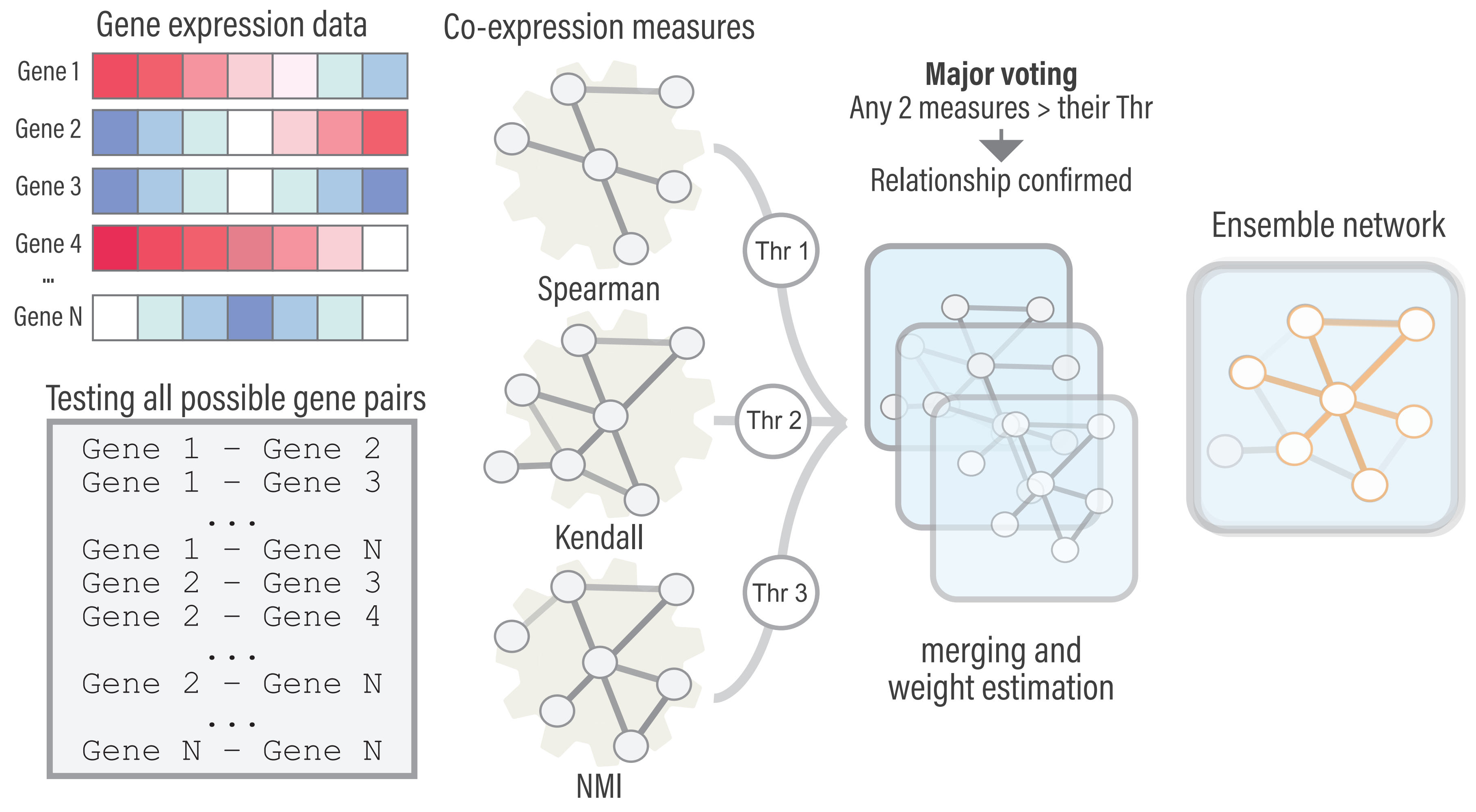
Francisco Gómez-Vela[1*], Fernando M. Delgado-Chaves[2], Domingo S. Rodríguez-Baena[1], Miguel García-Torres[1] and Federico Divina[1]
[1] Computer Science Division [2] Faculty of Experimental Sciences, Pablo de Olavide University, ES-41013 Seville, Spain

Co-expression gene networks have become a powerful tool in the comprehensive analysis of gene expression. Due to the diversity and the increasing amount of the available data, computational methods for networks generation must seek for the reliability of the obtained results. We present Ensemble and Greedy networks (EnGNet), a novel two-step method for gene networks inference. First, EnGNet uses an ensemble strategy for co-expression networks generation. Second, a greedy algorithm optimizes both the size and the topological features of the network. Achieved results show the method's ability to obtain reliable networks, also improving topological features.

EnGNet
Ensemble & Greedy Networks

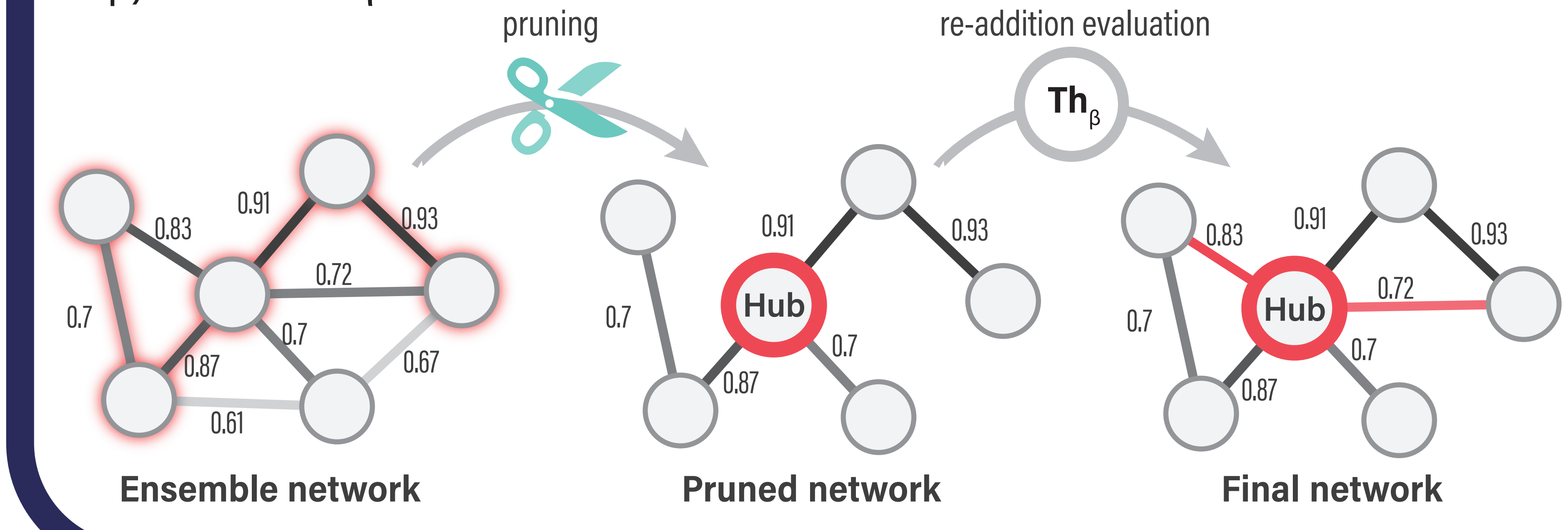
1 Estimation of an ensemble network

EnGNet infers individual co-expression networks, using three different measures, in this case, **Spearman** and **Kendall** coefficients and **Normalized Mutual Information (NMI)**, which evaluate every gene pair relationship. The measures provide a value vi, 0 ≤ vi ≤ 1, where 0 represents no dependency and 1 a total dependency between the genes. For each measure, a significance threshold (Thi, 1 ≤ i ≤ 3) is used to determine whether the relationship is considered valid by that specific measure. A relationship is considered relevant if supported by at least two measures. Its final weight is the average value of the three measures.



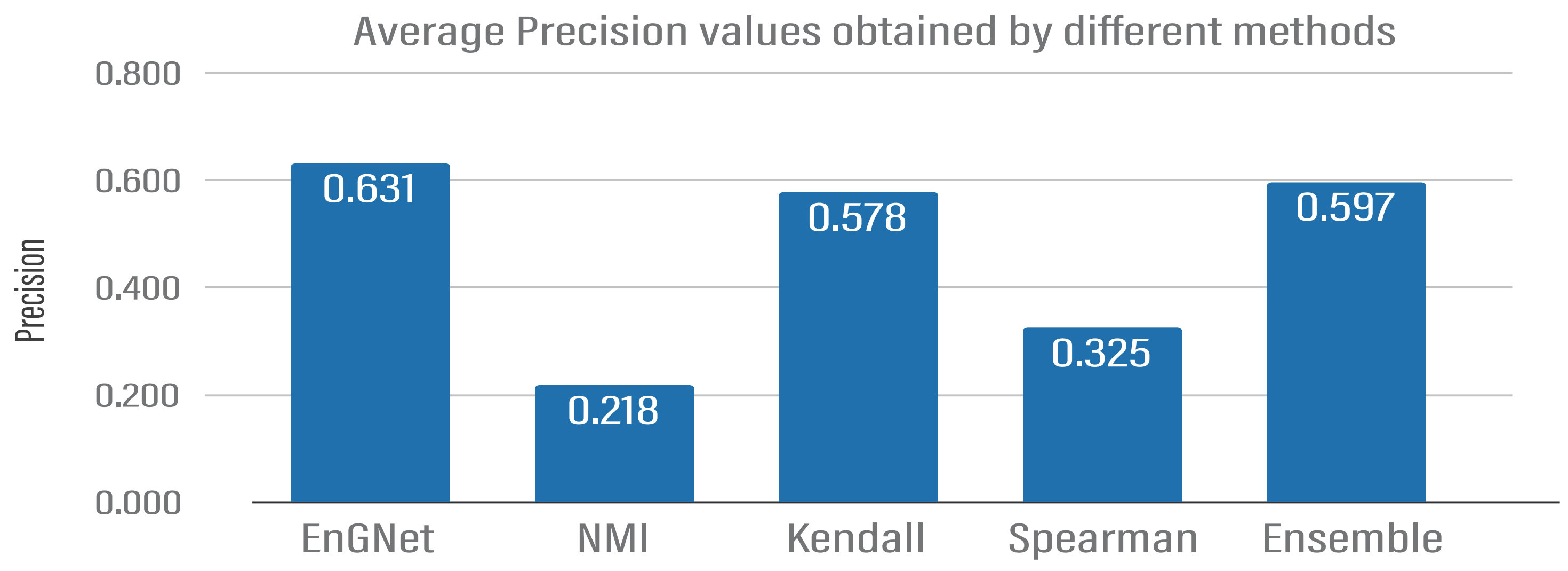
2 Greedy MST topology optimization

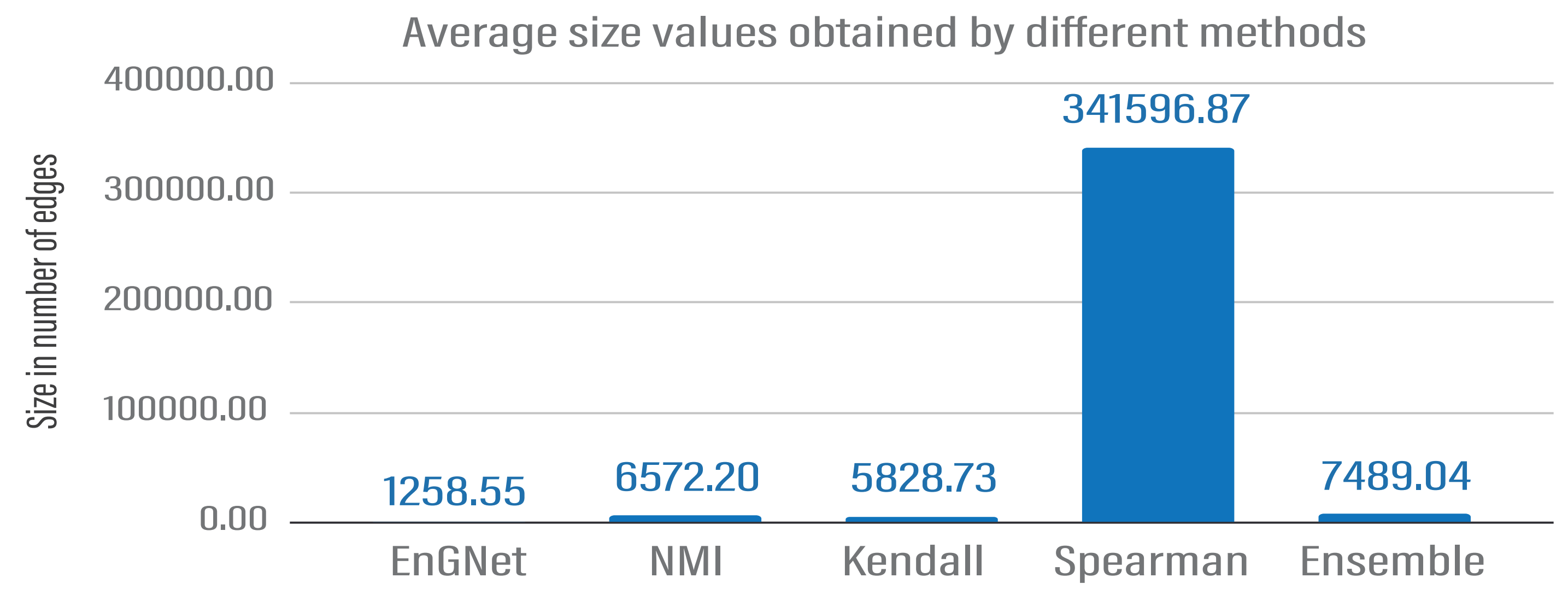
The ensemble network is pruned using a greedy-based heuristic algorithm, which is a modification of Kruskal's **minimum spanning tree (MST)** algorithm to obtain the longest path between each pair of genes, which selects most significant edges until all nodes are connected with no cycles. Then, a topological analysis of the pruned network is performed in order to identify hubs, which are selected as those nodes whose connection degree exceeds the average network connectivity. For each hub, its linking edges that were removed in the pruned network are again evaluated using a threshold (**Th**β), set by the user. Each individual edge will be added to the network if its weight (calculated in the ensemble step) exceeds Thβ.



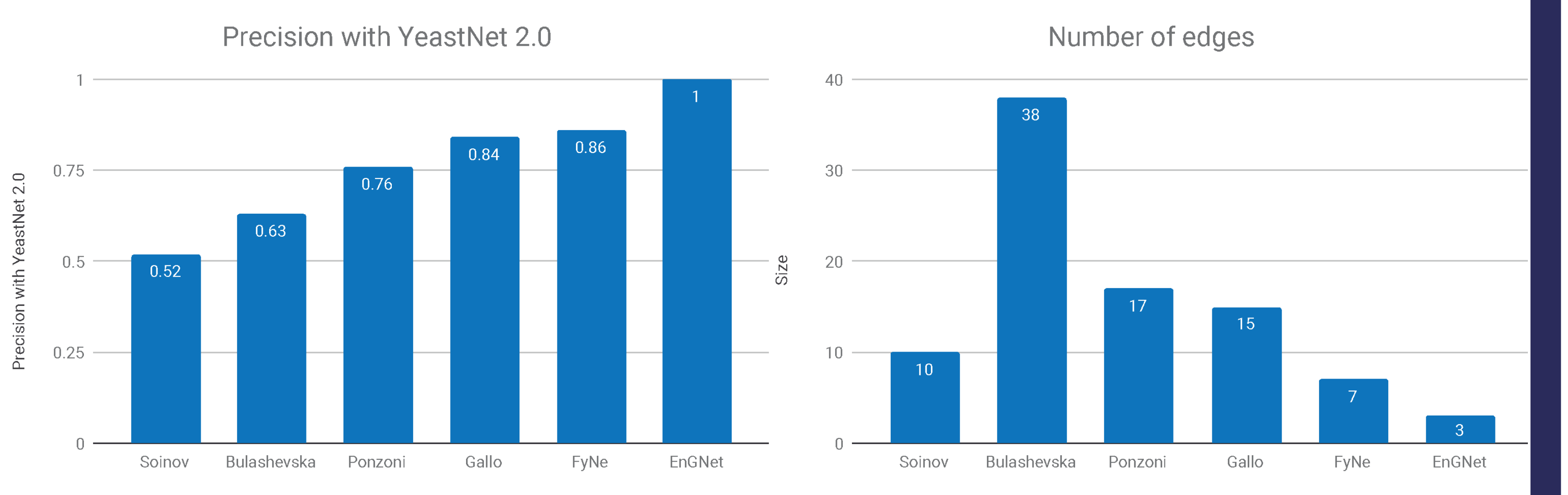
3 Comparative Analysis for Large Gene Networks

The biological significance of the obtained networks was tested by in a direct comparison with GeneMANIA database.





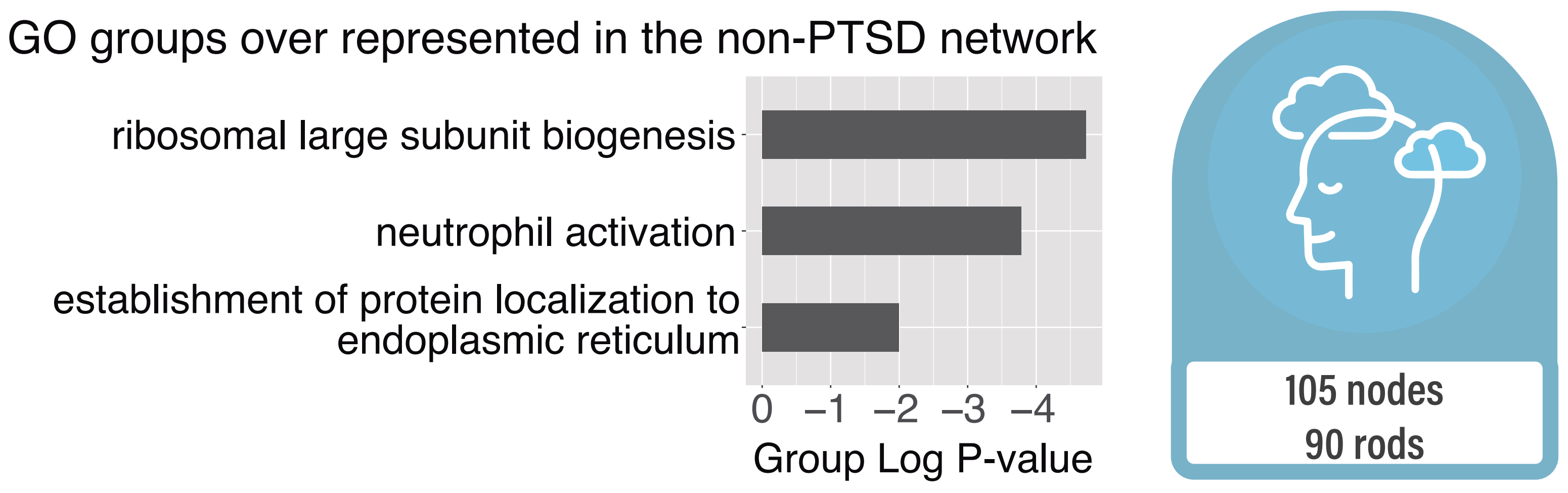
4 Comparative Analysis for Small Networks

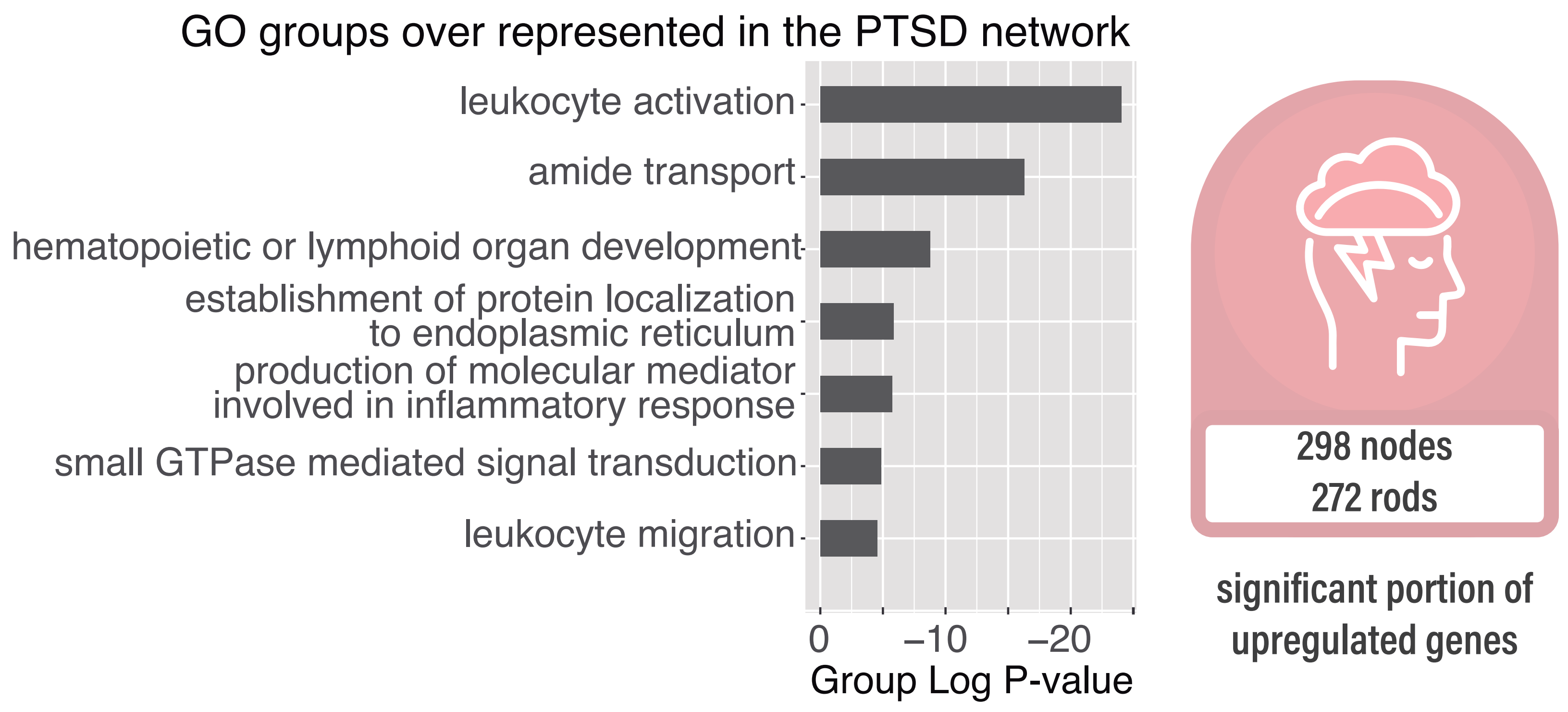


Different methods were applied to a dataset from the Yeast Cell Cycle. Networks quality was assessed regarding the precision values obtained against the data stored in YeastNet, as a GN gold standard.

4 Comparative Analysis for Small Networks

The biological relevance of EnGNet was successfully tested in the application to human **Post Traumatic Stress Disorder (PTSD)** dataset. EnGNet inferred gene association networks from the gene expression dataset, revealing an innate immunity-mediated response in PTSD cases, which was accompanied by considerable gene upregulation.





SCAN ME for further reading

Gómez-Vela, F., Delgado-Chaves, F. M., Rodríguez-Baena, D. S., García-Torres, M., & Divina, F. (2019). Ensemble and Greedy Approach for the Reconstruction of Large Gene Co-Expression Networks. *Entropy*, 21(12), 1139.