



This file explains the analysis output that includes some additional analysis and visualization features. Feedback on these results will help direct future design of website and standard analysis features.

Note: Your visualizations may not be as beautiful as hoped. This is a new feature and we need your feedback to optimize the dynamic scaling of layouts etc to customize the visualizations for different gene set sizes and network architectures. Your results are also available online for exploration and interaction so you can customize the placement of genes and customize the network image using our online interface.

Outline

1. Illustration of our pathway learning algorithm
2. The definition of a community
3. The definition of a candidate
4. The two output folders
 - a. With Candidates
 - b. No Candidates
5. Community summaries
6. Community figures
7. Community cross-talk figure and stability/significance
8. Gene set enrichment analyses
9. Supplementary figures

The underlying pathway algorithm that is used to score your gene set and predict new candidates in GeNets is called “Quack” – because if it walks like a pathway, and talks like a pathway, it’s probably a pathway! This classifier is already pre-trained before you enter your gene set, and scoring your gene set is very fast (< 1min). A high-level illustration of the key steps of this process is provided in Figure 1 below.

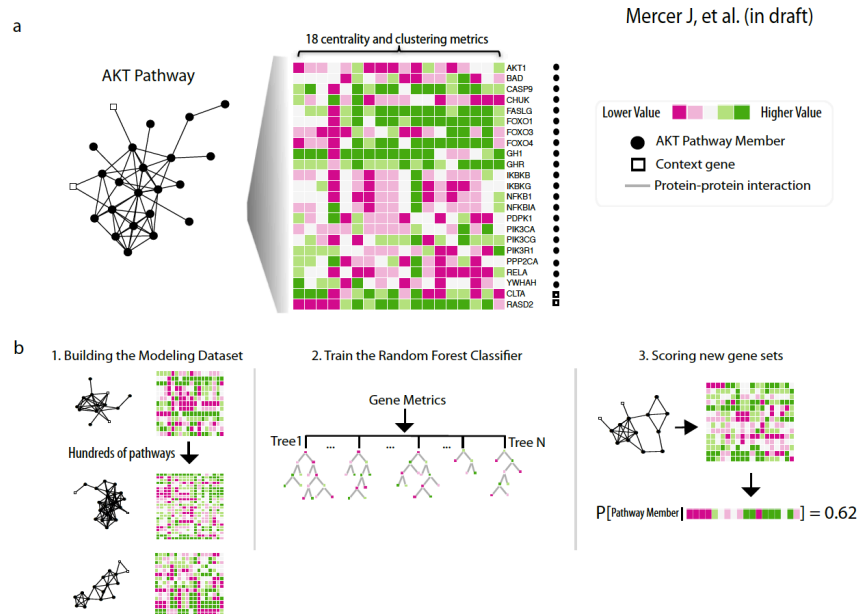


Figure 1 | Building a general classifier to predict pathway membership from networks. a) For a given pathway, we measure its architectural properties exemplified here with the 21 genes of the AKT pathway in the InWeb protein-protein interaction network. In the matrix, the 18 architectural metrics are shown as columns and the corresponding values for each of the 21 genes in the AKT pathway (black circles) as rows (metric values correspond to colors as indicated in the figure legend). One row in this matrix corresponds to one row in the final modeling dataset. We make the same measurements for genes in the context of the AKT pathway (white squares); only 2 of 2,449 context genes shown in the illustration. **b)** This procedure is repeated for 536 pathways from which the modeling dataset used to train the classifier is derived. For any candidate gene in a network, the classifier can assign a probability that it belongs to a pathway (e.g., the AKT pathway) as defined by the candidates architectural properties in the overall network and in relation to a specific set of genes (e.g., the 21 AKT genes).

The Definition of a Community

An important idea in network analyses is that of a *community*. We use this concept to help identify functional modules in gene sets and simplify visualizations. A community is a set of nodes (e.g. genes) that are more connected to one another than they are to other groups of genes. We use the method presented in “Finding Community Structure in Very Large Networks” by Clauset, et al. An illustration is provided in Figure 2 below.

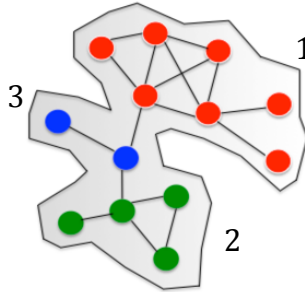


Figure 2: An illustration of three communities in one gene set.

Communities are annotated with numbers, i.e. 1, 2, 3, etc.


The Definition of a Candidate


Candidates are predicted genes that, based on their connectivity to your gene set, exhibit the same features as those found in known pathways.

The Definition of Network P-value

For significance, in v1.5 we transitioned from using the *clustering coefficient* of the network to using the *density* of the network as defined by $\text{density} = (\# \text{ of edges} / \# \text{ possible edges})$ and then compute the density for randomly sampled gene sets. If the density of your gene set is greater than 95% of the randomly sampled gene sets, we deem it “significantly more connected than random.” So it’s an empirically determined p-value. We provide p-values for your network overall (both with and without candidates) and also we compute p-values by community as you may find that the whole network isn’t compelling, but components of it are so we want to test connectivity of these sub-modules separately.



Two Output Folders

 *[Analysis name]_NoCandidates*: this folder includes visualizations and analyses performed on your original gene set.

 *[Analysis name]_WithCandidates*: this folder includes visualizations and analyses performed after adding in top tier candidates – predicted by Quack- into your gene set.

Community Summaries

These two files contain community membership and summary information. The community membership file also contains a candidate indicator (1= is candidate; 0=your test gene).

 CommunityMembership.csv
 CommunityStatistics.csv



Community Figures

In the `_WithCandidates` folder, there are 5 versions of the network visualizations.

Figure1: The original “hairball” with the community encoded as node color. This is to provide contrast to the more appealing community layouts.

**Figures 2-5 we have separated the communities and highlighted the connectivity within community.*

Figure2: This is a “community layout” that highlights candidates.

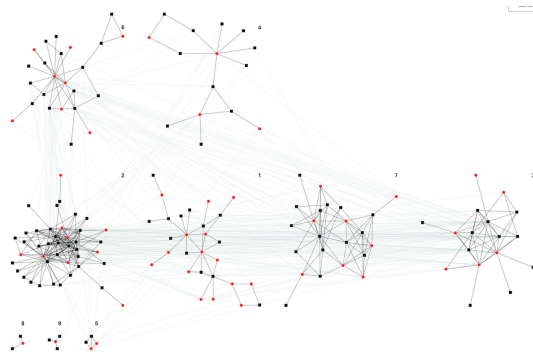


Figure3: Same as Figure2, but *without* node labels (sometimes this helps to identify patterns).

Figure4: This is a “community layout” that highlights community membership (node color is based on community).

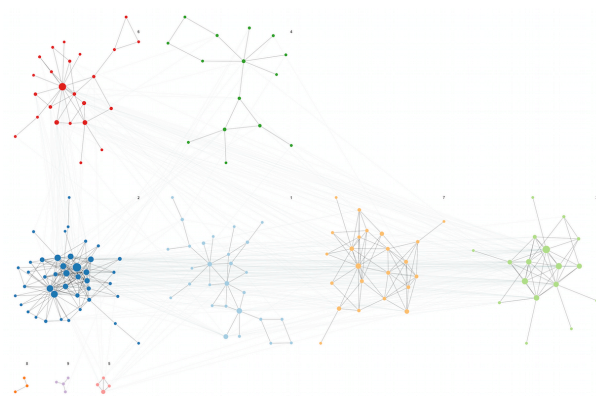


Figure5: Same as Figure4, but *without* node labels (sometimes this helps to identify patterns).


In the `_NoCandidates` folder, there are 3 versions of the network visualizations (1,4, and 5 from above).






By-Community Analyses

There are two “by-community” sub-folders: enrichment and visualization.

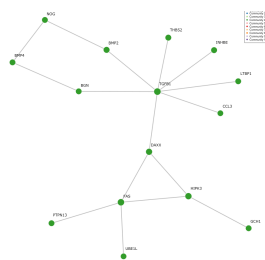
Within each of these folders, there are three subfolders:

 **_EnrichmentByCommunity**: this folder contains enrichment results for each community found in your gene set. Gene set enrichment is conducted using a Bonferroni-adjusted hypergeometric test. Only the C2 set (n=4,722 gene sets from Biocarta, KEGG, Reactome, PID, etc) from MSigDB is considered during this analysis.

 **_VisualizationByCommunity**: This folder contains three supplement figures *per community* that was found. For example, Community 1 would have:

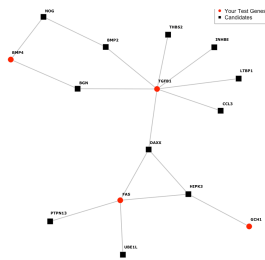
 Supplement_1A.jpg
 Supplement_1B.tif
 Supplement_1C.png

Supplement 1A



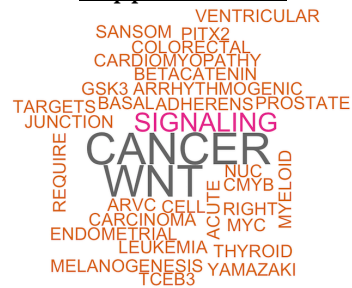
Node color is based on community

Supplement 1B



Node color/shape encoded “candidate” or “your test genes”

Supplement 1C




Word cloud of significantly enriched gene sets

The **_NoCandidates/_VisualizationByCommunity** folder will only contain Supplements A and C, because this folder pertains to only your test genes.


 **_VisualizationByCommunity**: This folder contains three supplement figures *per community* that was found. For example, Community 1 would have:


Drug-target interactions


 **_DrugTargetInteraction**: This folder contains all the visualization and summary files you see in the parent folders, but with drug-target interactions integrated with the original network. These drug-target interactions were procured from the paper below, which includes sources such as DrugBank and Drugs in Clinical Trials Database:





Rask-Andersen, M., Masuram, S., & Schiöth, H. B. (2014). The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annual Review of Pharmacology and Toxicology*, 54, 9–26.
<http://doi.org/10.1146/annurev-pharmtox-011613-135943>

 Drug-Target-Community-Membership-Summary.csv

 Drug-Target-Community-Membership.csv

 Network-Community-Layout.pdf

 Network-Force-Layout.pdf

 NetworkEdges-With-Drug-Target-Interactions.csv