Supplementary Materials for

A gene ontology inferred from molecular networks

Janusz Dutkowski, Michael Kramer, Michal A Surma, Rama Balakrishnan, J Michael Cherry, Nevan J Krogan & Trey Ideker

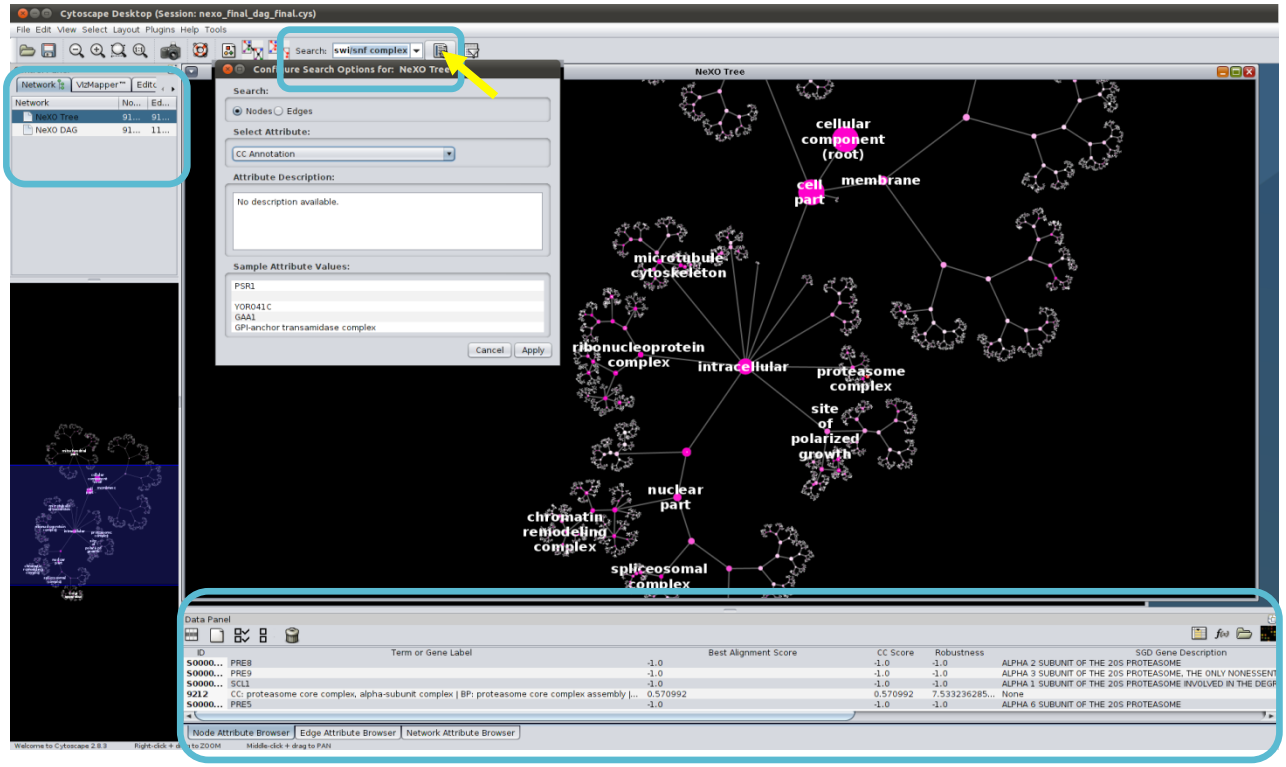
NeXO USER MANUAL

Navigation with Cytoscape

The NeXO.cys (Supplementary File 1) is a Cytoscape session file that can be loaded directly into Cytoscape version 2.8.3 (available from http://www.cytoscape.org/) to allow for visualization and interactive analysis of the entire ontology. The session provides three versions of NeXO: the tree, displaying the backbone of the ontology, the DAG where supplementary term-term relations are added, and also the DAG with additional gene-term assignments. The following steps illustrate an exemplary analysis using NeXO.cys and Cytoscape.

1. **Launch Cytoscape** and use the **File→Open menu** to locate and select the file NeXO.cys.

2. Select the **NeXO Tree** from the Network tab of the Control Panel at left (below screenshot).



3. Select the menu item **View→Show Graphics Details** (if the menu reads **Hide Graphics Details**, this command is already active). NeXO is visualized with high-level terms labeled via alignment to the GO Cellular Component ontology. Node size scales with the number of genes assigned at or below that term. Node color intensity scales with the Best Alignment Score to GO. See above screenshot and Fig. 2 in the main text.

4. **Select parts of the tree using the mouse** (e.g. the proteasome subtree). Information on selected nodes (terms or genes) are shown in the table in the Data Panel. The main attributes (columns) are as follows. Further details about the derivation of each attribute are provided in Supplementary Methods.

ID The NeXO identifier (terms) or SGD identifier (genes) of the node.

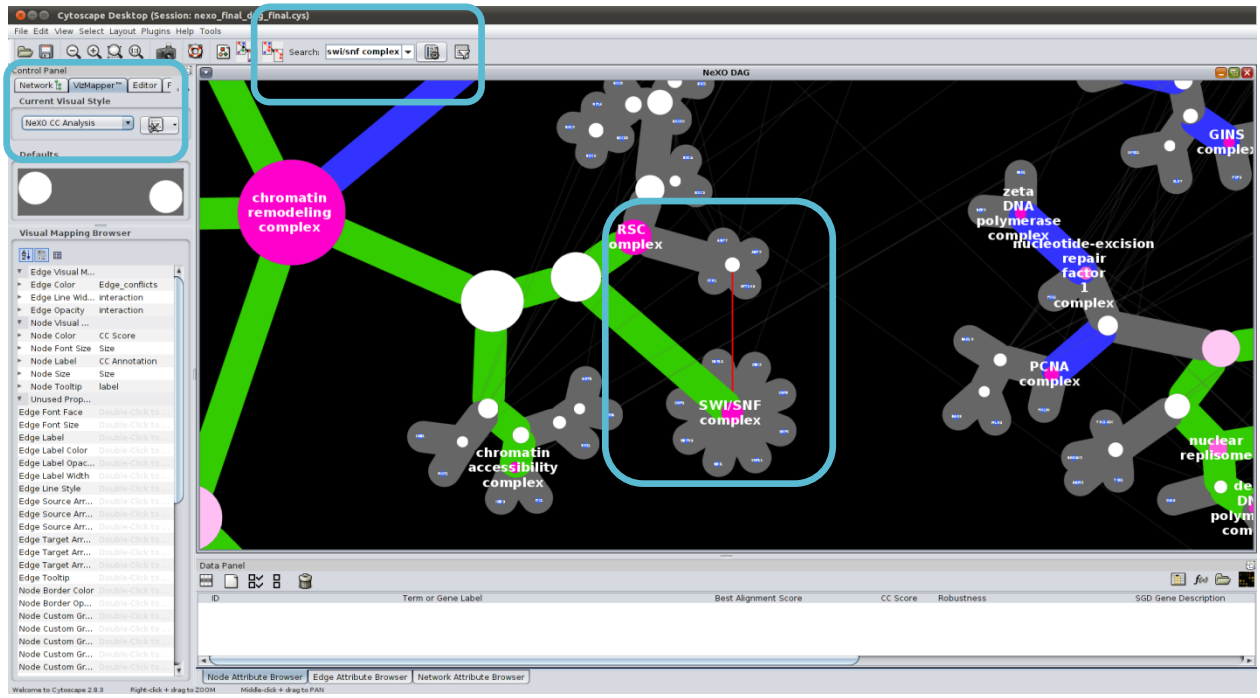
Term or Gene Label	For terms, names are transferred from GO alignment (scores > 0.1, CC = Cellular Component, BP = Biological Process, MF = Molecular Function). For genes, common names are shown or, if unavailable, the ORF is shown.
CC Score	The alignment score vs. GO Cellular Component (defaults to -1.0 for genes).
Best Alignment Score	Maximum alignment score against the CC, BP, and MF ontologies of GO.
Robustness	The term robustness score (Supplementary Methods, defaults to -1.0 for genes).
SGD Gene Description	Provides the free-form Gene Description field from SGD (genes only).

Additional attributes such as term definitions (CC, BF, MF), term size (number of assigned genes) and interaction density can also be loaded using the **Data Panel**.

5. Select the **VizMapper** tab of the Control Panel (below screenshot) and switch from the **NeXO Overview** to the **NeXO CC Analysis** visual style. This display will color the nodes according to their alignment score against the GO Cellular Component ontology. Green paths indicate consistent term-term relations between the two ontologies, i.e. ancestor-descendant pairs that are aligned to GO terms that are also in an ancestor-descendant relation. In contrast, a blue edge (parent to child) indicates a relation present in NeXO but absent from GO, such that the GO terms aligned to the child and the nearest aligned NeXO ancestor are not in a descendant-ancestor relation in GO. Only NeXO terms with high alignment score (CC score ≥ 0.2) are considered in this relation analysis – other terms are neutral for determining agreements and conflicts (e.g., a green path may pass through them).

6. Search NeXO for cellular components or genes of interest. Use the mouse to push the button to the right of the **Search field** (above screenshot) to configure search options. Select the **CC Annotation** attribute and push **Apply**. Type "swi/snf" into the search field and press **Enter**. Gene names are also searchable using this attribute. The SWI/SNF complex will be shown.

7. Select the **NeXO DAG** or **NeXO DAG (filtered)** from the Network tab of the Control Panel and repeat steps 3-6. In the full DAG version of the ontology, edges are added which connect child terms to additional parents (see below example, in which a subunit of RSC is connected to the SWI/SNF complex). Similar analysis can also be performed using NeXO DAG with additional gene-term assignments.



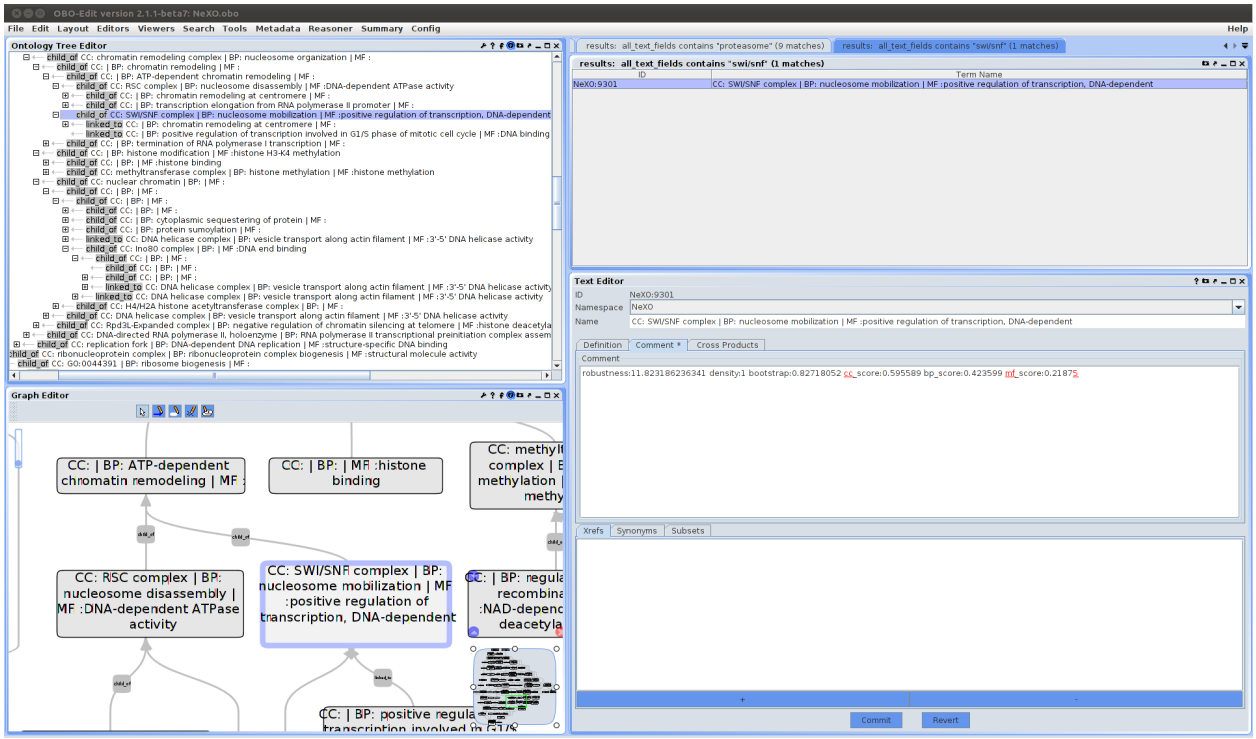
Navigation with OBO-Edit

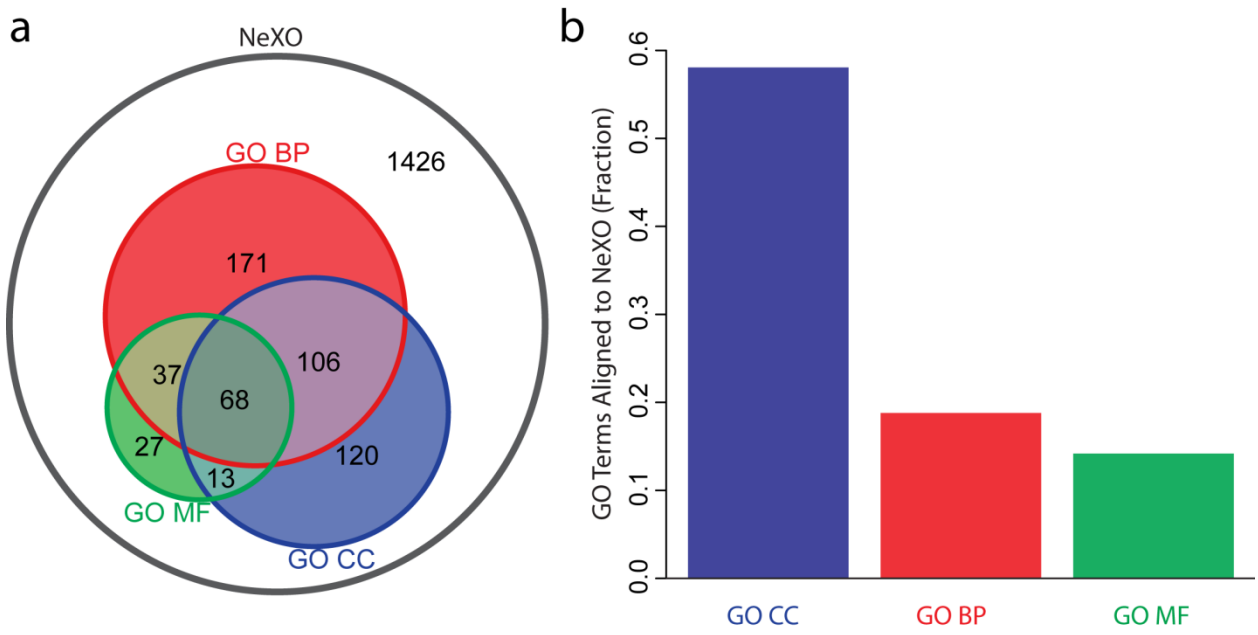
The NeXO ontology can also be explored using designated ontology editors and browsers such as OBO-Edit (available from http://www.oboedit.org/) as shown in the below screenshot. For this purpose we have provided the file NeXO.obo (Supplementary File 2) which conforms to the Open Biological and Biomedical Ontology (OBO) format. Using OBO-Edit the NeXO ontology can be explored in three basic ways:

- Using the Ontology Tree Editor
- Using the Graph Editor
- By searching for keywords appearing in NeXO term annotations.

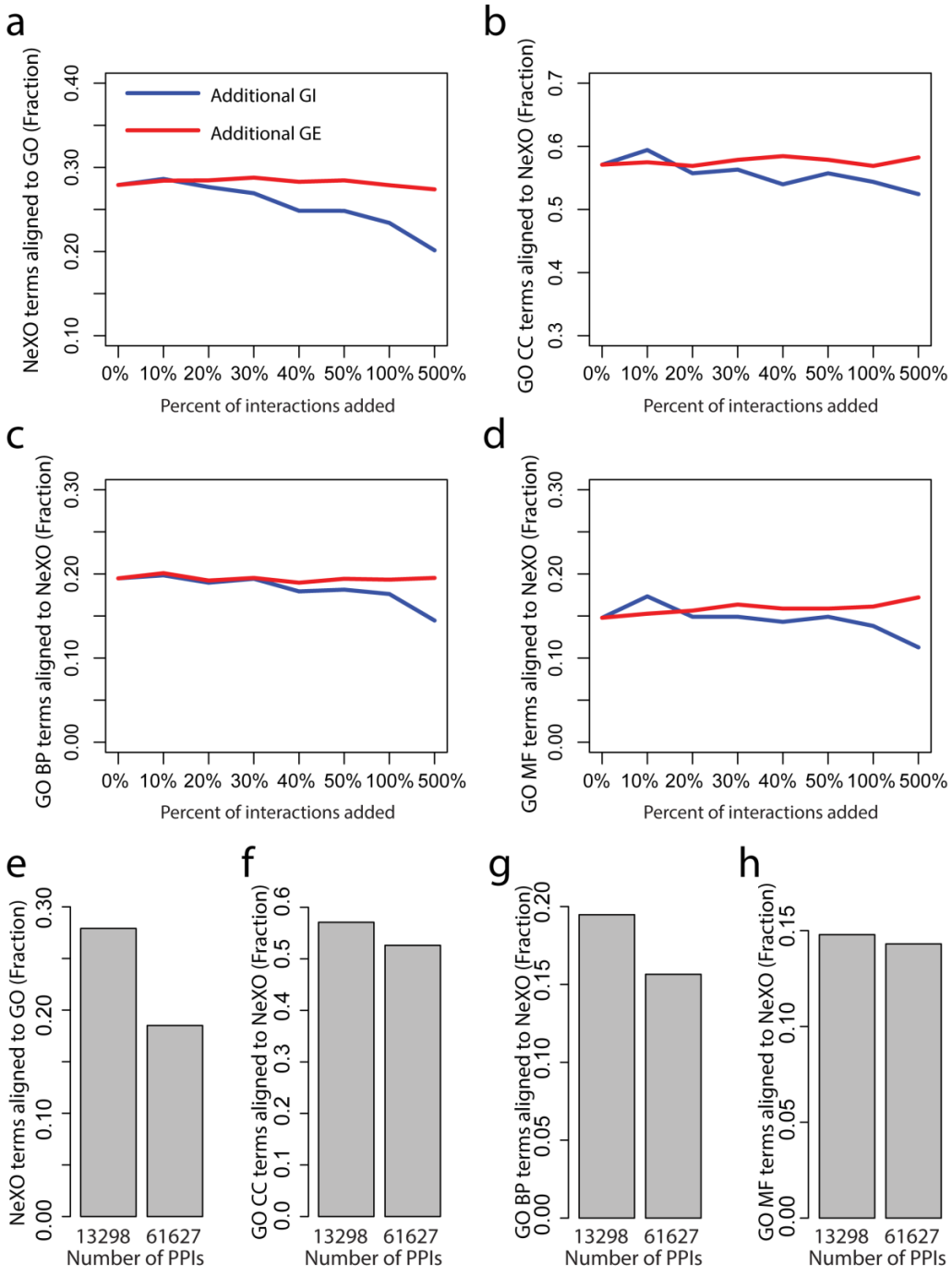
Terms in the NeXO.obo file are annotated with text labels transferred through alignment to GO. The label indicates matchings with score ≥ 0.1 to any one of the three GO ontologies: Cellular Component (CC), Biological Process (BP), and Molecular Function (MF). Additional information about the term such as its robustness score, interaction density, and alignment scores are provided in the Comment field.

Each term relationship is annotated with one of two types: “child_of” or “linked_to”. The “X child_of Y” relationship indicates that X is a child of Y in the tree forming the backbone of the NeXO ontology. “X linked to Y” indicates that X was linked to Y by an additional edge in the NeXO DAG.

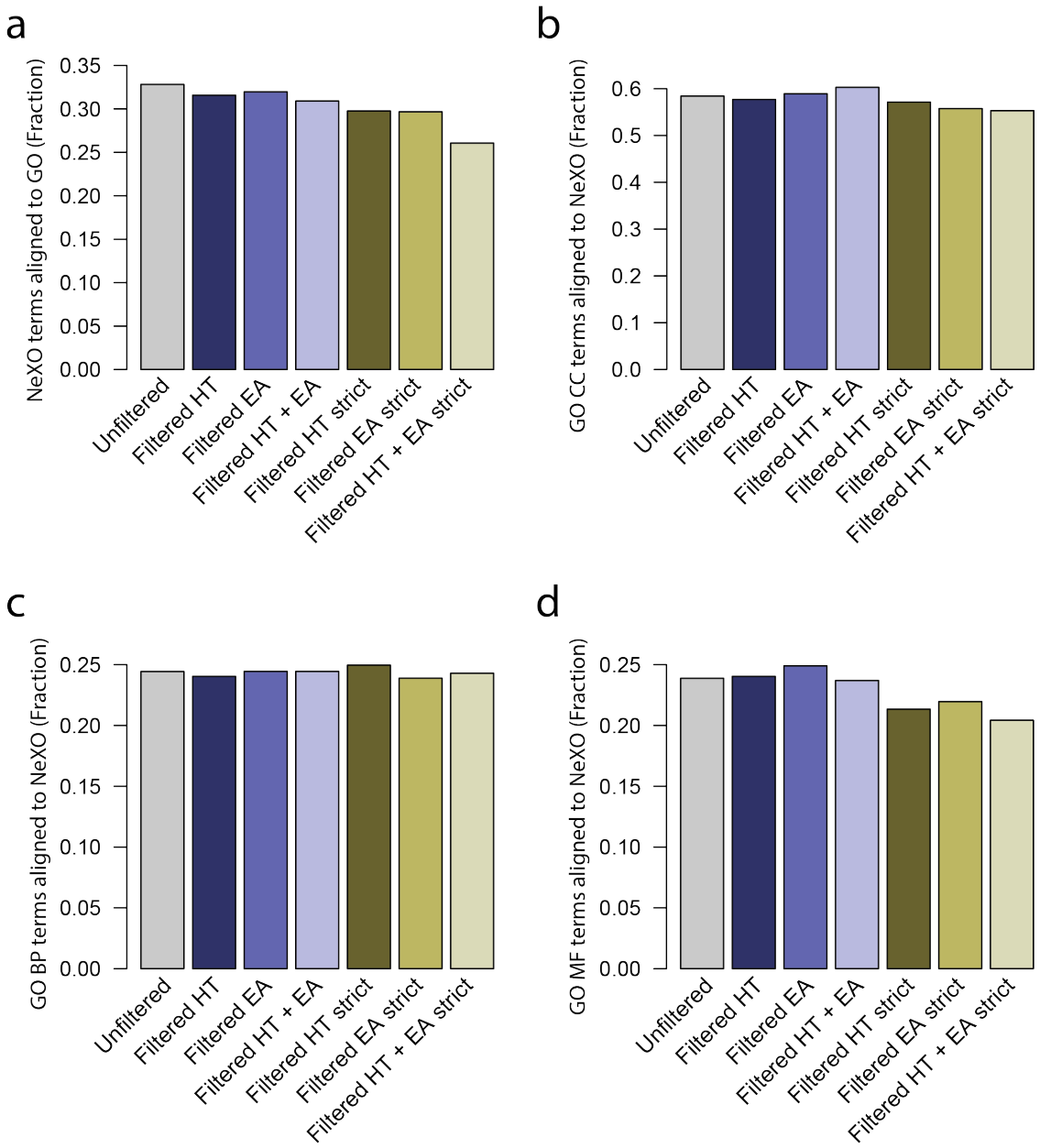




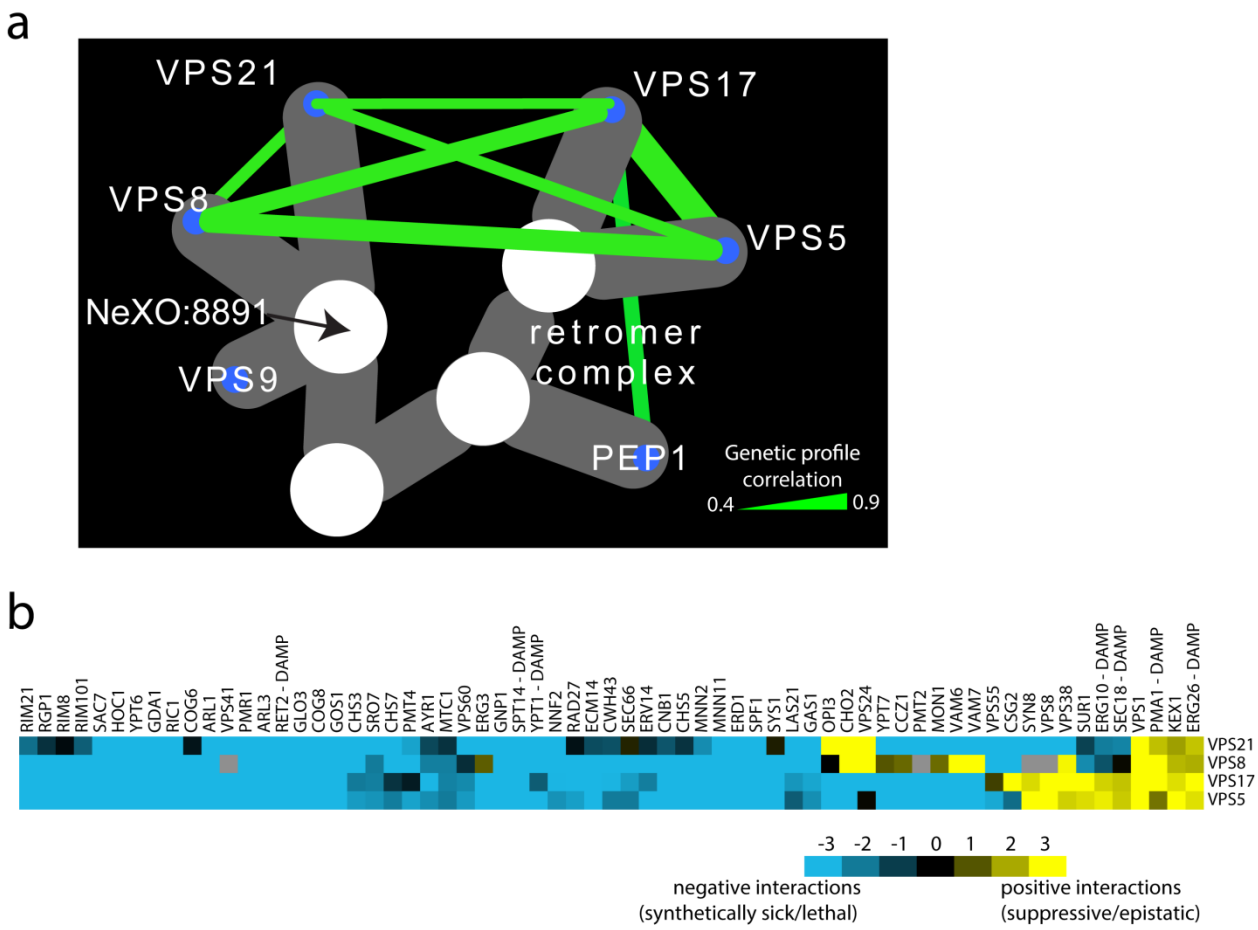
Supplementary Figure 1. Alignment results for NeXO computed without the YeastNet and co-expression networks. (a) The Venn diagram shows the number of NeXO terms of size ≥ 4 that align to one or more GO ontologies (colored circles) or do not align (white). (**b**) The fraction of terms in each GO ontology that are covered by alignment to NeXO. The overlap with GO (in particular GO Cellular Components) closely matches the results obtained with the original NeXO that included information from all four input networks.



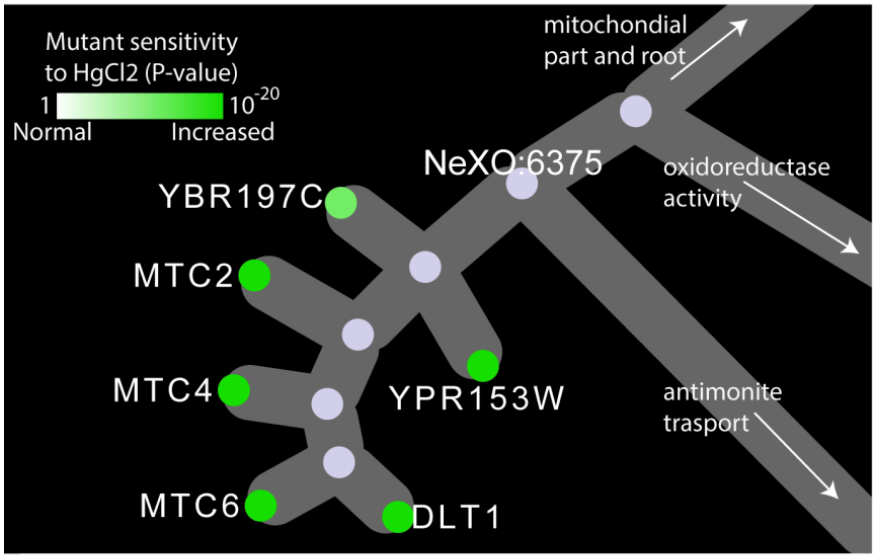
Supplementary Figure 2. Alignment results for NeXO using unfiltered networks. The percent of NeXO terms that aligned to GO (a), and GO terms that aligned to NeXO (b-d) when NeXO is built using more permissive interaction networks with 0-500% more genetic (GI) or co-expression (GE) interactions. The percent of NeXO terms that aligned to GO (e), and GO terms that aligned to NeXO (f-h) when NeXO is built using the original network and a more permissive interaction network which includes all 61627 physical protein-protein interactions (PPIs) in BioGRID. We observe that alignment results are robust across a wide range of more permissive input networks. Note that in all cases the input networks contain PPI, GI, and GE data, but exclude YeastNet to make sure that GO is not used for network training.



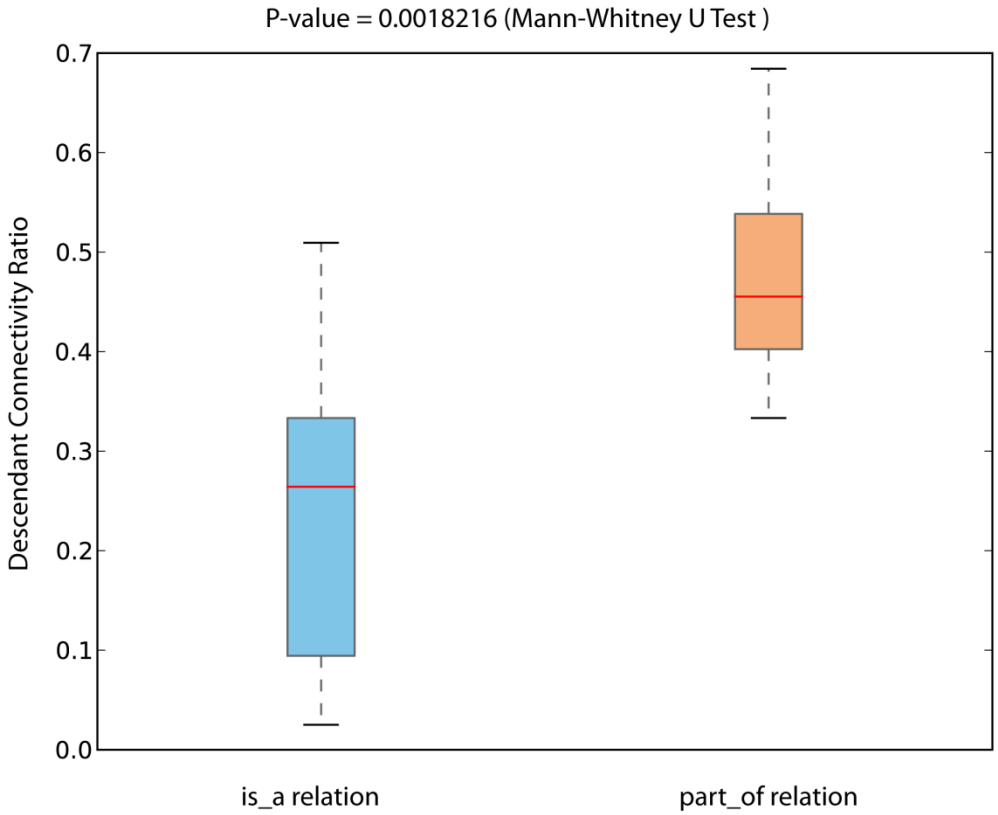
Supplementary Figure 3. Alignment results using GO with excluded high-throughput annotations. The percent of NeXO terms that aligned to GO (a), and GO terms that aligned to NeXO (b-d) when gene-to-term association based on high-throughput interaction data (HT) and/or those inferred from electronic annotations (EA) are removed from GO. The specific evidence codes related to high-throughput interaction data (HT) are: IPI - Inferred from Physical Interaction; IEP - Inferred from Expression Pattern; and IGI - Inferred from Genetic Interaction. In the 'strict' version we removed all associations of gene X to term Y if at least one of these associations met the condition for removal. In the non-strict version gene X could still be associated with term Y if it had at least one evidence code that did not meet the condition for removal. NeXO alignment results appear robust to removal of HT and EA annotations from GO.



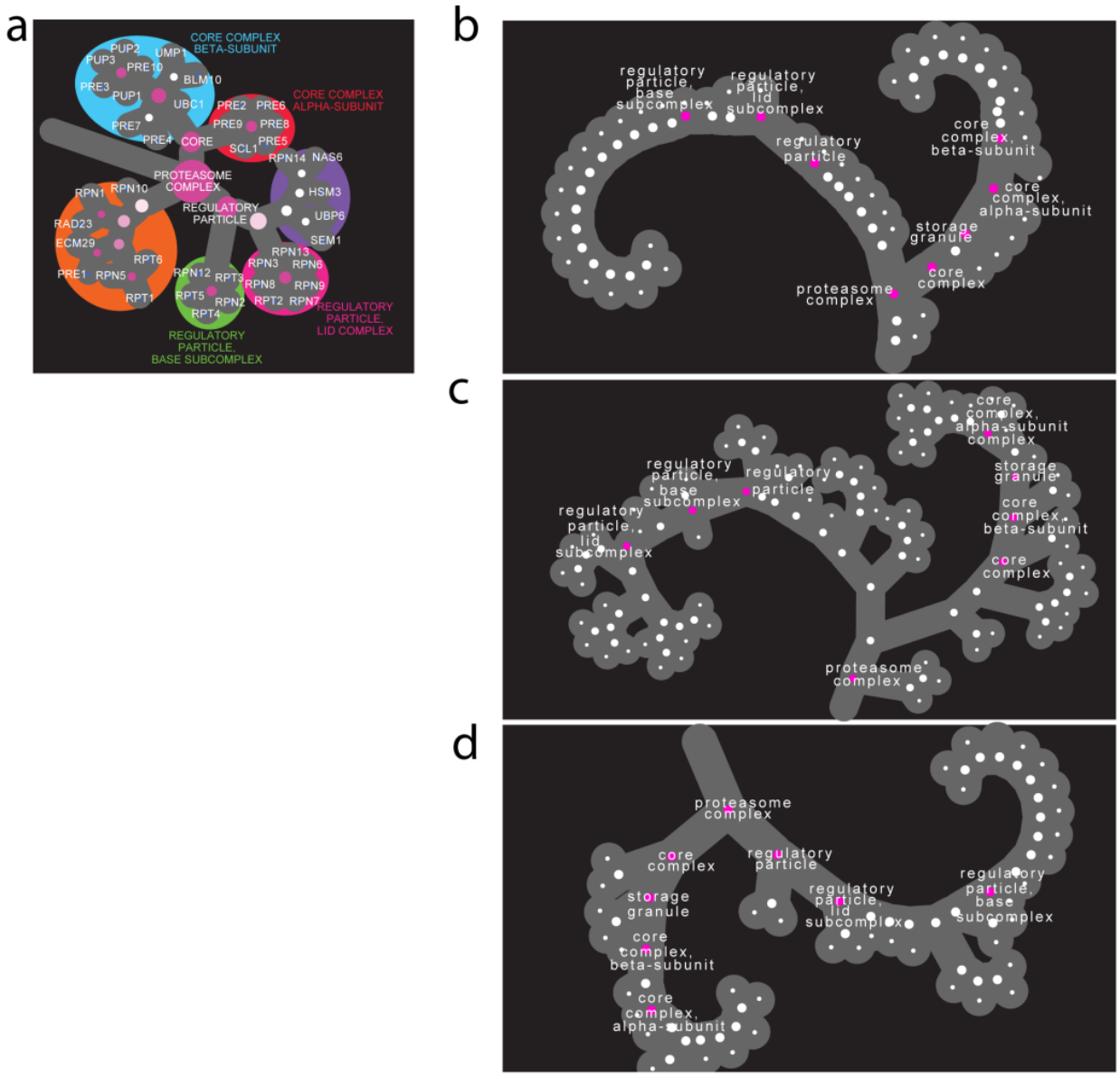
Supplementary Figure 4. New NeXO term associated with the retromer complex. New term NeXO:8891, composed of *VPS8*, *VPS21* and *VPS9,* and placed directly next to the retromer complex in NeXO (a). The new genetic interaction screen identified very high profile correlation between genes in the new term (VPS8, VPS21) and the retromer subunits VPS17 and VPS5 (a,b), providing further support for the position of the new term next to the retromer complex in the NeXO ontology.



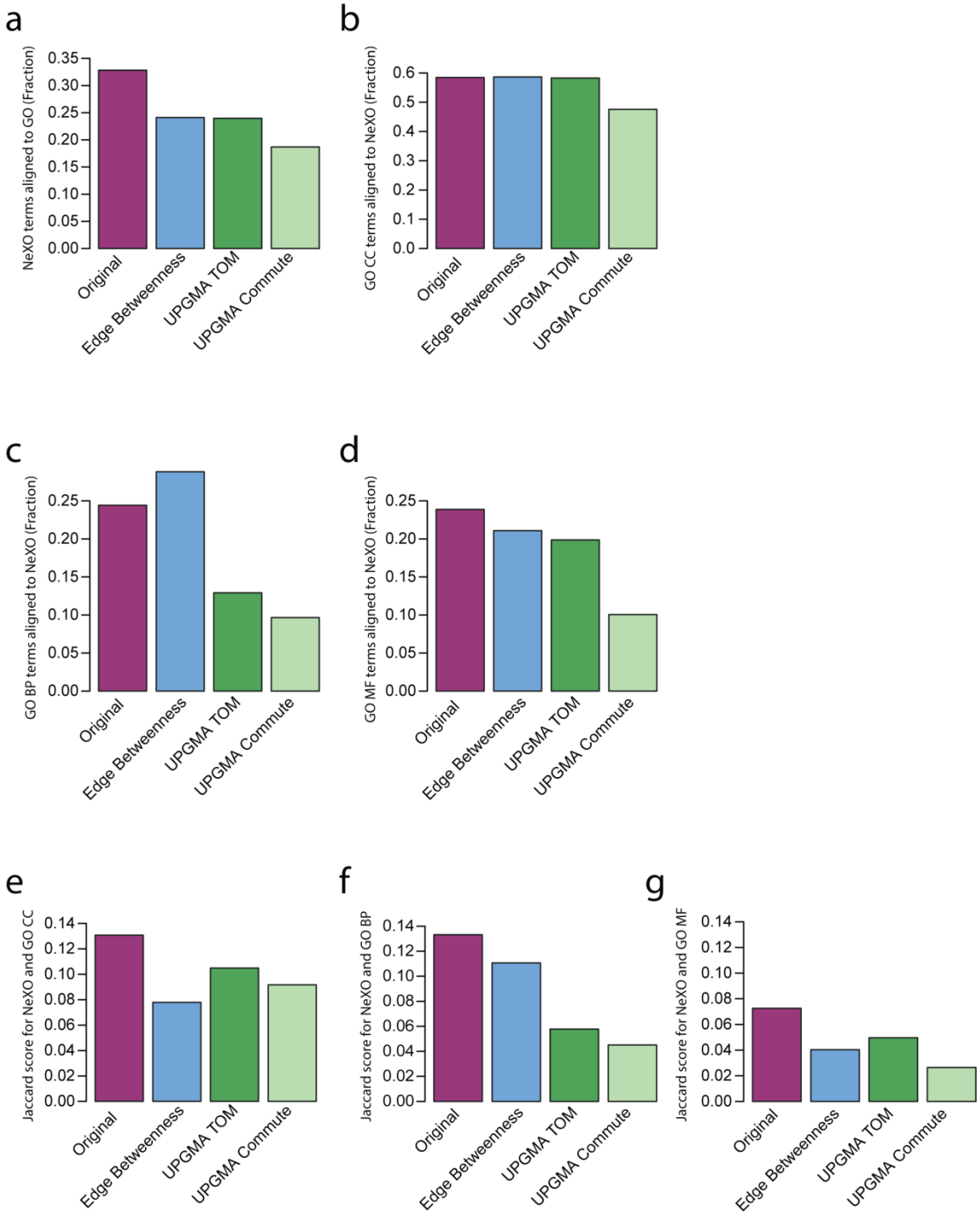
Supplementary Figure 5. New term enriched for mutants sensitive to mercury chloride ($HgCl_2$). The term NeXO:6375 is composed of six poorly uncharacterized ORFs: *MTC2*, *MTC4*, *MTC6, DLT1*, YBR197C and YPR153W. NeXO places this terms under the mitochondrial component next to genes annotated with oxoreductase activity and antimonite transport. Deletions of each of 6 genes in this term make yeast sensitive to mercury chloride ($HgCl_2$).



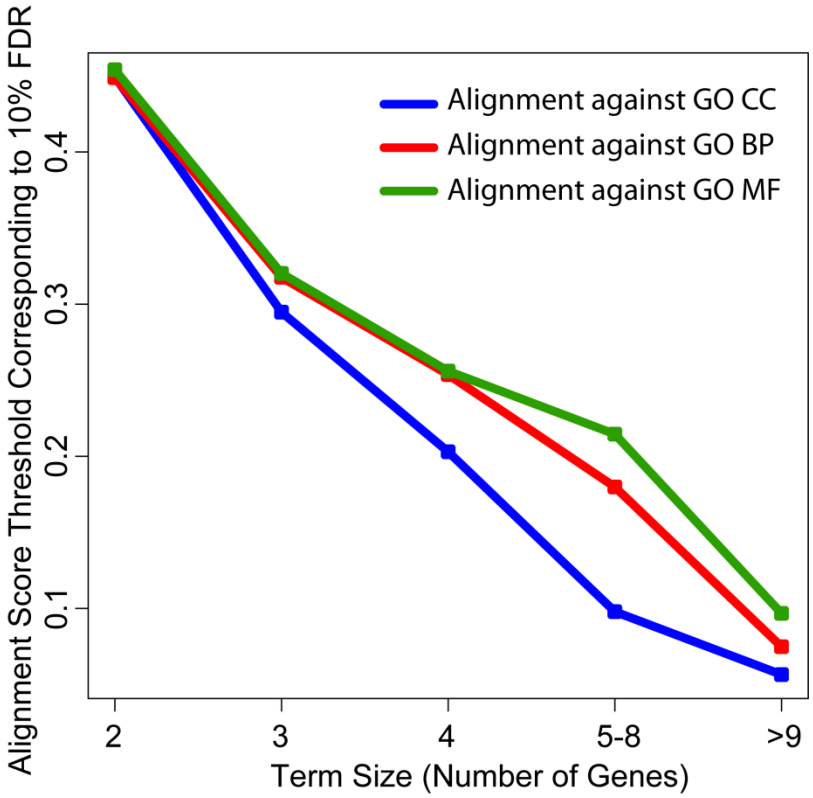
Supplementary Figure 6. Distinguishing "is_a" and "part_of" relations in NeXO. For each parent-child relation in NeXO that aligned to a parent-child relation in GO, the parent was examined to compute its Child Connectivity Ratio (CCR). This was the ratio of the number of molecular interactions spanning different children of the parent to the total number of interactions falling within the parent. CCR values for parent-child relations in NeXO mapping to "is_a" relations in GO are significantly lower than for NeXO relations mapping to "part_of" relations.



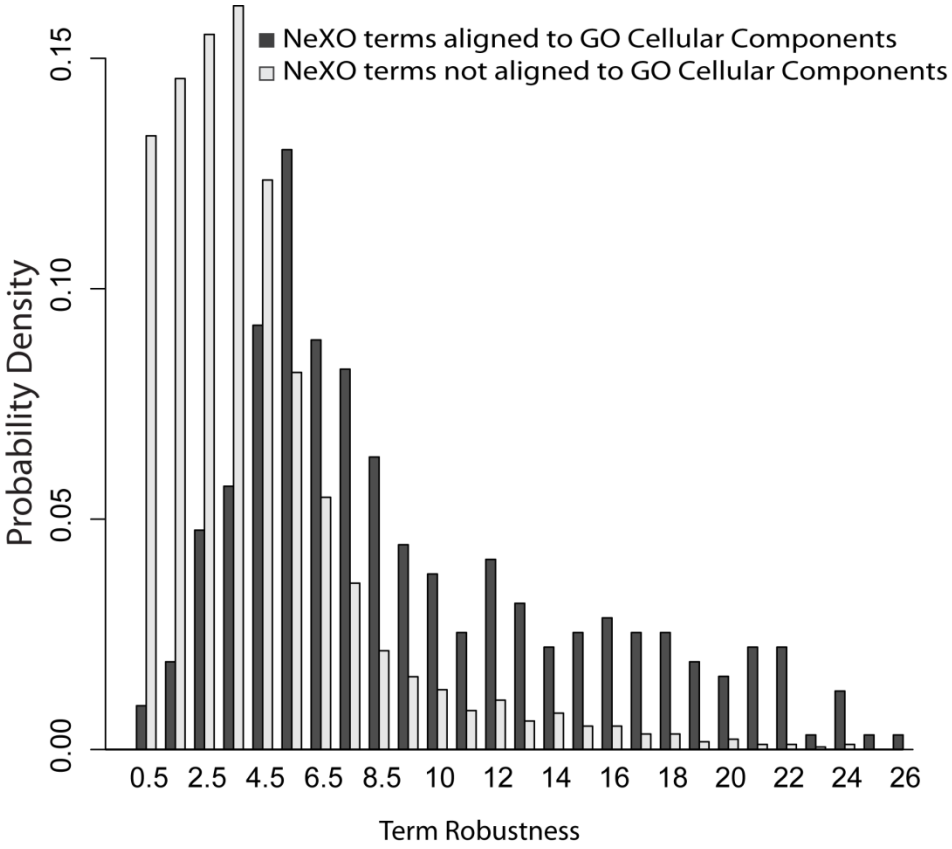
Supplementary Figure 7. **Representation of the proteasome when using other clustering methods in the NeXO approach**: original (a), edge betweenness (b), UPGMA based on topological overlap distance (c), and UPGMA based on commute distance (d). NeXO alignment is performed to map the resulting trees to the GO Cellular Component ontology. The alternative clustering methods (b-d) do not recover the hierarchy to the same extent as the original method used by NeXO (a). While core and regulatory particles are correctly identified, organization of more specific components is incorrect (b-d).



Supplementary Figure 8. Alignment results for NeXO using other hierarchical clustering methods. Percent of NeXO terms that aligned significantly (FDR = 10%) to any of the three GO ontologies (a), percent of terms in each of the GO ontologies that aligned to NeXO (b-d). Jaccard index comparing the set of terms in NeXO to the set of GO Cellular Components (e), GO Biological Processes (f), and GO Molecular Functions (g). The original method achieves better agreement with GO than other standard methods.



Supplementary Figure 9. Alignment score threshold at 10% FDR. In the ontology alignment of NeXO to GO, different alignment score thresholds were used for different NeXO term sizes in order to maintain a false discovery rate of 10%. CC = Cellular Component, BP = Biological Process, MF = Molecular Function.



Supplementary Figure 10. Distribution of robustness for aligned and unaligned terms. The distributions are well separated for robustness > 5.

Supplementary Table 1. Input networks investigated in this study

Network Type	Used in Figs. 1B, C		HQ used in all other analyses	
	Number of Genes	Number of Interactions	Number of Genes	Number of Interactions
Physical Protein-Protein	5,658	61,627	3,401	13,298
Co-Expression	3,915	61,627	228	1,705
Genetic	4,396	61,627	3,090	11,240
YeastNet	5,268	61,627	3,351	10,573
Integrated High-Quality (HQ)	--	--	5,051	29,789