# Deleteome-Tools: Utilizing a compendium of yeast deletion strain transcriptomes to identify co-functional genes

Maxwell L. Neal[1], Sanjeev K. Choudhry[1,†], John D. Aitchison[1,*]

[1]Center for Global Infectious Disease Research, Seattle Children's Research Institute, Seattle, WA, USA

* To whom correspondence should be addressed.

† Current affiliation: Center for Life Sciences, Mahindra University, Hyderabad, Telangana, India

Correspondence to:

John D. Aitchison, Ph.D.
Professor and Co-Director
Center for Global Infectious Disease Research
Seattle Children's Research Institute
307 Westlake Avenue North, Suite 500
Seattle Washington, 98109-5219

206-884-3125 OFFICE
john.aitchison@seattlechildrens.org

**ABSTRACT**

We developed an R codebase that uses a publicly-available compendium of transcriptomes from yeast single-gene deletion strains - the Deleteome - to predict gene function. Primarily, the codebase provides functions for identifying similarities between the transcriptomic signatures of deletion strains, thereby associating genes of interest with others that may be functionally related. We describe how our tool predicted a novel relationship between the yeast nucleoporin Nup170 and the Ctf18-RFC complex, which was confirmed experimentally, revealing a previously unknown link between nuclear pore complexes and the DNA replication machinery. We also discuss how our strategy for quantifying similarity between deletion strains differs from other approaches and why it has the potential to identify functional relationships that similar approaches may not. Deleteome-Tools is implemented in R and is freely available at https://github.com/AitchisonLab/Deleteome-Tools.

## INTRODUCTION

A central motivation for systems biology research is to understand the function of complex biological systems from a holistic perspective. Within the context of a single cell, such research can illuminate how the interplay of thousands of gene products produce a cellular phenotype of interest. The expanding number and availability of public, omics-level data resources can be leveraged to address this challenge, since they offer holistic readouts from biological systems. One such resource applicable for systems cell biology is the Deleteome compendium (Kemmeren *et al.*, 2014) which contains 1,484 transcriptomic profiles obtained from *Saccharomyces cerevisiae* single-gene deletion strains. For each of these strains, the Deleteome's creators used microarrays to measure transcript abundances. The compendium consists of the expression changes and associated *P*-values derived from comparisons between these abundances and wild-type controls. Thus, each single-gene deletion strain in the Deleteome has an associated transcriptomic "signature" consisting of its set of differentially expressed genes. The Deleteome has been used, for example, to generate and analyze systems-level gene regulatory networks, determine the proportion of gene-specific transcriptional repressors, and predict novel protein-protein interactions in yeast (Kemmeren *et al.*, 2014; Dorrity *et al.*, 2020).

We have used the Deleteome to functionally characterize the nucleoporin Nup170, which is a structural constituent of the nuclear pore complex (NPC) and also plays a role in gene regulation, including sub-telomeric gene silencing. To help reveal the mechanism whereby Nup170 mediates silencing, we designed and implemented R code for collecting the differential expression signature of the Nup170 deletion strain (*nup170*Δ) in the Deleteome, iteratively comparing it to signatures from all other strains in the compendium, and then identifying and ranking deletion strains with similar signatures. We designed our codebase so that this task of identifying transcriptionally similar deletion strains could be performed for any of the strains in the compendium, not just for *nup170*Δ. These similarity analyses were critical to our discovery of Nup170's functional association with the chromosome transmission fidelity 18-replication factor C (Ctf18-RFC) complex (Kumar *et al.*, 2023). Among the deletion strains similar to *nup170*Δ were the chromosome transmission fidelity 18 (*ctf18*Δ), chromosome transmission fidelity 8 (*ctf8*Δ), and defective in sister chromatid cohesion 1 (*dcc1*Δ) strains. These represent the three non-essential components of the Ctf18-RFC complex, which helps establish chromosome cohesion during mitosis and is involved in loading and unloading the proliferating cell nuclear antigen (PCNA) clamp. (Deletion strains for the other members of the Ctf18-RFC

complex, due to their essentiality, are not present in the Deleteome and could not have been identified as co-functional with Nup170 using our analysis.) Our subsequent experimental investigations showed that the Ctf18-RFC complex is recruited to a subset of distinctly-structured NPCs, revealing a novel association between NPCs, the Ctf18-RFC complex, and the PCNA replication clamp, which is required for DNA replication and repair (Kumar *et al.*, 2023). Given that our computational analyses predicted the co-functional relationship between Nup170 and the Ctf18-RFC complex, we have made our code publicly available in the hopes that the research community may find it useful for identifying novel functional relationships among other genes. In the following sections, we detail the computational methods implemented in Deleteome-Tools for quantifying similarity between deletion strain transcriptomes as well as additional visualization and analysis features that it provides.

## METHODS

Primarily, Deleteome-Tools was developed to quantify the similarity between transcriptomic abundances in a deletion strain of interest (a “query” strain) and all other Deleteome strains. In a broad sense, the software performs phenotype matching: the set of gene expression values present in the query strain constitutes a phenotype, which is compared against expression values in other strains. Then, functionally profiling the genes deleted in matching strains can predict the functionality of the gene deleted in the query strain. This approach for predicting gene function from a compendium of deletion strains has been applied previously using a smaller set of strains (Hughes *et al.*, 2000), and we aimed to apply a similar approach using the more comprehensive set provided by the Deleteome.

After Deleteome-Tools computes similarity scores between a query strain and all other Deleteome strains, scores are ranked and thresholds applied to remove lower-confidence scores, thereby generating a list of deletion strains similar to a query strain. The software provides two complementary methods for quantifying similarity, which we detail next. Following that, we detail additional visualization and analysis features also provided by the software.

### Similarity based on reciprocal correlation

The first similarity metric provided by Deleteome-Tools is based on correlation tests. In the following example, we detail how this metric is computed given a deletion strain of interest (“strain A”) and another strain from the Deleteome to which it is being compared (“strain B”).

First, the set of differentially expressed genes in strain A (those significantly different from wild-type based on *P*-values provided in the Deleteome) are identified. These genes constitute the strain A's transcriptomic "signature". We then compare the expression values of signature genes from strain A to the corresponding values in strain B. If the correlation is statistically significant and positive, we perform a reciprocal test where expression values of strain B's signature genes are correlated against those in strain A. This step helps ensure that the entirety of strain A's signature is similar to the entirety of strain B's. Consider the case where the function of the gene deleted in strain A is limited to a specific gene regulatory pathway and the gene deleted in strain B also functions in that pathway, but also several others. A correlation test between strains A and B limited to the differentially-expressed genes in strain A's deletion signature might show significance but would ignore the larger changes present in strain B's signature. In this example, the reciprocal correlation test may not prove significant since deletion of the gene in strain B would introduce a larger number of transcriptional changes which might not be present in strain A. Thus, if a reciprocal correlation test is insignificant or negative, the two strains are not classified as similar. In other words, because the gene deleted in strain B has additional functionality compared to the gene deleted in strain A, our tools may not necessarily classify the genes as similar, even though they may share some overlapping functions.

Deleteome-Tools performs this reciprocal correlation test between a query strain and all other deletion strains in the Deleteome and then collects the strains that were significantly, positively correlated in both directions. It then ranks similarity based on FDR-corrected *P*-values from the first correlation test (expression values of strain A's signature compared to corresponding values in strain B). In initial tests using this approach, we found that some deletion mutants, including *nup170*Δ, had several hundred similar deletion strains. We therefore added a feature to further down-select this list by including only those strains whose correlation *P*-values are within a certain percentile of all *P*-values. In our work with *nup170*Δ, we used the 95$^{th}$ percentile as our cutoff for similarity, although users can readily adjust this threshold to decrease stringency. Thresholds for gene expression values and FDR-corrected correlation test *P*-values used in these analyses can be easily adjusted by users.

**Similarity based on signature enrichment**

The second similarity metric implemented in Deleteome-Tools uses a more qualitative approach compared to reciprocal correlation. We developed this complementary alternative because reciprocal correlation tests may not always be significant for co-functional genes, given that they rely on quantitative gene expression values. Consider an example involving two Deleteome

strains A and B where the gene deleted in A co-functions in a transcriptional regulatory pathway with the gene deleted in B. The specific expression changes introduced in strain A may not show a significant correlation with those in B, even though the expression changes are qualitatively similar: expression of all downstream genes in the pathway might increase in both strains, but the specific expression values might not necessary correlate significantly. Therefore, we developed an “enrichment-for-signature” similarity test that could account for such cases.

In this approach, the genes in the signature of a deletion strain of interest are identified using the same approach described above and the expression of each gene is categorized as increased or decreased based on its value. Then, when comparing that strain’s signature to that of another strain, we collect the signature of the second strain and label its specific genes as increased or decreased, and then identify the genes that are present in both signatures and have the same directional expression change (i.e., genes that increased significantly in both strains or decreased significantly in both). We then perform a hypergeometric test to quantify the probability of observing that number of overlapping genes between the two signatures by chance. After performing hypergeometric tests comparing a strain of interest to all other Deleteome strains, we FDR-correct the associated *P*-values, select those that are significant, and apply a percentile cutoff as described above to determine which strains qualify as similar.

In our experience, the enrichment-based approach tends to be less conservative than the reciprocal correlation option as it often identifies a comparatively higher number of similar strains. However, we encourage the complementary use of both approaches, as they are non-redundant. In the case of *nup170*Δ, the enrichment-based approach identified 74 similar strains and the correlation-based approach identified 40. Twenty-eight were common to both (including strains for all three components of the Ctf18-RFC complex), indicating that while there is likely to be substantial overlap in the results from each method, both identify strains the other does not.

**Visualization features**

Deleteome-Tools also provides functions for visualizing transcriptome data across multiple deletion strains and the relative genomic position of a deletion strain’s differentially expressed genes. For the former, users can easily generate heatmaps of gene expression values for a given deletion strain alongside any other strains from the compendium. An example is shown in Figure 1 where the expression values for genes in the *nup170*Δ strain are shown in addition to

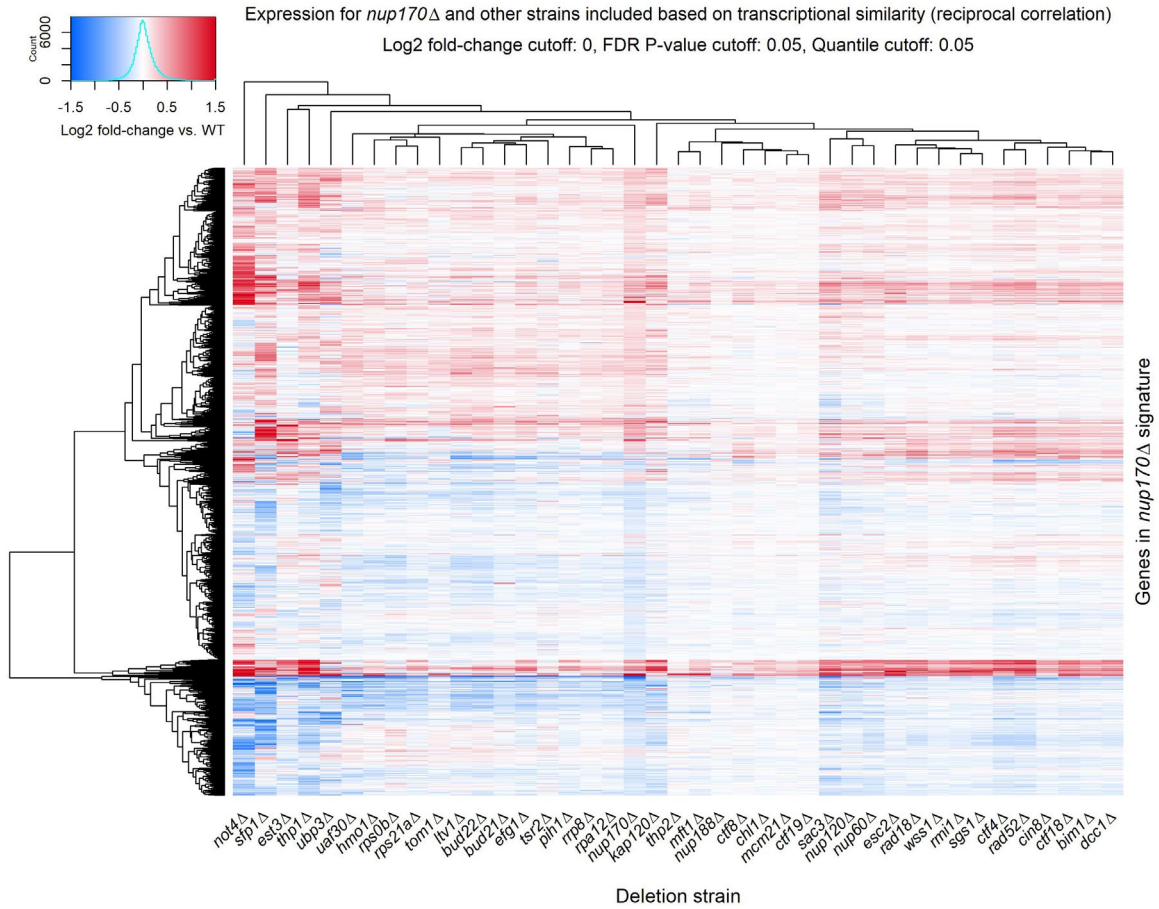


**Fig. 1.** Example visualization provided by Deleteome-Tools: a heatmap of $\log_2$ fold-changes in gene expression for the *nup170*Δ Deleteome strain and 40 other strains with similar transcriptional signatures.

those from the 40 similar deletion strains identified using the reciprocal correlation method described above.

For investigating the genomic position of a transcriptional signature, we provide functions for visualizing the distance of a deletion strain's signature genes from telomeric ends or from the centromere. We have used these visualizations, which were originally devised during early investigations into Nup170's subtelomeric silencing role (Van de Vosse *et al.*, 2013) and which we refer to as "mountain lake" plots, to identify and compare subtelomeric silencing defects among various Deleteome strains. Supplementary Figure S1 shows an example of this plot for the *nup170*Δ deletion strain.

### Other analysis features

Our codebase also provides a convenience function for performing Gene Ontology (GO) (Ashburner *et al.*, 2000; Aleksander *et al.*, 2023) enrichment analyses on user-defined sets of genes that have matching deletion strains in the Deleteome. We have used this feature primarily to functionally profile the deleted genes in Deleteome strains that are similar to a query strain. For example, we used this function to test GO enrichment among the complement of deleted genes in the strains similar to *nup170*Δ that are shown in Figure 1. The function relies on GO enrichment tests provided by the *clusterProfiler* R package (Wu *et al.*, 2021) and uses the full set of genes deleted in Deleteome strains as the global background for the tests.

Deleteome-Tools also provides convenience functions that quantify the over-representation of genes within subtelomeric or centromeric regions. Given an input set of genes, these functions apply hypergeometric tests to detect enrichment of genes within the specified region. We have used these functions to identify and quantify subtelomeric silencing defects in deletion strains. For example, in the *nup170*Δ strain these tests show a highly significant enrichment of upregulated genes within 25 kilobases (kb) of the telomere (Hypergeometric test *P*-value = 3.54e-11). This silencing defect is reflected by the peak in Figure S1 among upregulated genes located near the telomere.

### Availability and dependencies

Deleteome-Tools is implemented as a small set of R scripts and data objects and is publicly available at https://github.com/AitchisonLab/Deleteome-Tools. After downloading from the repository, users can run example R scripts that quantify deletion strain similarity, visualize expression data, and perform GO enrichment tests. Users can then adapt the contents of these scripts into their own analysis workflows.

Identifying similar deletion strains with Deleteome-Tools does not require the installation of additional R packages. However, generating visualizations requires installing the *gplots* and *ggplot2* packages, and performing GO enrichment tests requires the *org.Sc.sgd.db* and *clusterProfiler* R packages, all of which are free and publicly available.

## DISCUSSION

A distinguishing feature of our approach for quantifying similarity between deletion strains is that we limit our strain-to-strain comparisons to gene expression values that changed significantly

versus wild-type. This is in contrast to other approaches that cluster deletion strains based on the full complement of measured gene expression values. This may have important implications for predicting gene function. For example, whereas our similarity analysis predicted a functional relationship between Nup170 and members of the Ctf18-RFC complex with high confidence, a UMAP-based clustering of the Deleteome placed Nup170 and the Ctf18-RFC members in distinct clusters that are relatively distant from each other in UMAP space (Dorrity *et al.*, 2020). We suspect this may be due to the difference in the complement of gene expression values used to quantify strain similarity.

Within the UMAP cluster mentioned above, Nup170 is grouped with 18 other members, and together they circumscribe a set of biological functions distinct from those of the Ctf18-RFC complex. Ten of these 18 members are also predicted to be co-functional with Nup170 by our reciprocal correlation analysis. These findings suggest that while there may be substantial agreement between functional predictions made using full transcriptomes and signature-based approaches such as ours, the codebase described here offers a valuable, complementary strategy for prediction that may highlight important functional relationships between genes not revealed by other methods.

**ACKNOWLEDGEMENTS**

We thank Drs. Fred Mast and Fergal Duffy for their help testing the codebase.

**FUNDING**

This work has been supported by grants P41GM109824 and R01GM112108 from the National Institutes of Health to J.D. Aitchison.

*Conflict of Interest:* none declared.

**REFERENCES**

Aleksander,S.A. *et al.* (2023) The Gene Ontology knowledgebase in 2023. *Genetics*, **224**, iyad031.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Dorrity,M.W. *et al.* (2020) Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nat. Commun.*, **11**, 1537.

Hughes,T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Kemmeren,P. *et al.* (2014) Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, **157**, 740–752.

Kumar,S. *et al.* (2023) Nuclear pore complexes mediate subtelomeric gene silencing by regulating PCNA levels on chromatin. *J. Cell Biol.*, **222**, e202207060.

Van de Vosse,D.W. *et al.* (2013) A role for the nucleoporin Nup170p in chromatin structure and gene silencing. *Cell*, **152**, 969–983.

Wu,T. *et al.* (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov.*, **2**, 100141.