In silico evaluation of Gene Ontology Terms in *Saccharomyces cerevisiae* synthetic chromosomes II and III: a rational design for synthetic lethality proposition

Hugo Lins de Albuquerque Vieira[1], Leonardo Ferreira[1], Luis Oliveira[1], Mateus Meira[1], Matheus Leitão[1], Mayna Gomide[1], Samuel da Silva Andrade[1], Fernando Araripe Gonçalves Torres[2], Ildinete Silva Pereira[3], Georgios Joannis Pappas Júnior[4], Cintia Marques Coelho[1]*

[1]Laboratory of Synthetic Biology, Institute of Biological Sciences, Department of Genetics and Morphology, University of Brasília, Brasília, Brazil

[2]Laboratory of Yeast Biotechnology, Institute of Biological Sciences, Department of Cell Biology, University of Brasília, Brasília, Brazil

[3]Laboratory of Molecular Biology of Pathogenic Fungi, Institute of Biological Sciences, Department of Cell Biology, University of Brasília, Brasília, Brazil

[4]Laboratory of Genomics and Bioinformatics, Institute of Biological Sciences, Department of Cell Biology, University of Brasília, Brasília, Brazil

[*]**Correspondence:**
Cíntia Marques Coelho
cintiacoelhom@unb.br

Keywords: Genome reduction, Synthetic yeast chromosome, Gene Ontology, Synthetic lethality, Cytoscape, ClueGO

Introduction

Genome reduction has been an approach used to study gene essentiality, but it is also desired for generating optimized genomes that will be prone to biosynthetic uses. In *Saccharomyces cerevisiae*, an important eukaryotic model organism, with a wide range of use in metabolic engineering, much has been done for synthetic genome assembly, such as Sc 2.0[1]. However, large scale genome reduction is still impaired by the lack of knowledge of which sequences comprise the set of essential and quasi-essential genes[2]. For the yeast's synthetic chromosome three, synIII, Synthetic Chromosome Rearrangement and Modification by loxP-Mediated Evolution (SCRaMbLE) removed genes considered nonessential at once resulting in cell death[3], and genome wide Hi-C analysis provided evidence that SCRaMbLEd strains undergo severe and unexpected structural chromosome alterations[4]. Taking the results together it became clear that SCRaMbLE would not be the appropriate approach to investigate gene essentiality, and a different rational design is needed. In this manner, an Ontology-based approach for the identification of genes that should not be knocked down together would be a relevant advance. Nevertheless, some difficulties are yet to be overcome when proposing such an approach, for example, the amount of Gene Ontology (GO) classes associated with the same gene product, variation in GO classes and level since very high-level GO Terms that are broad do not necessarily impose a lethality issue, how to access enough information from a genome in order to establish gene set with different probabilities of lethality for *in vivo* validation, and development of bioinformatic tools that enable genome data analysis.

A preliminary work that proposed to set ontology criteria for gene clusterization in groups that are likely to cause synthetic lethality revealed that our pick of GO Terms might have been biased and lead to groups that contain repeated genes and/or are not distinguishable in terms of lethality assumption, which is not desired for *in vivo* validation. Thus, the present study exhausted the Biological Process GO Terms from *S. cerevisiae* synthetic chromosomes II and II (syn2 and syn3)[3,5].

Methods

Genes from syn2 and syn3 classified both as individually nonessential (iNEGs) were chosen for GO Slim Term mapping, because this class was removed when SCRaMbLE was previously performed. A Python script was developed in order to retrieve the genes present in the final synthetic chromosome assembly diagrams. Those genes were then compared through the OGEE v.3.09 to filter the iNEGs based on the most recent data available.

The iNEG list was uploaded to SGD's web service Gene Ontology Slim Term Mapper and a search was run with the following parameters: GO Set: Yeast GO-Slim Process; List of GO Slim Terms: Select All.

All of the Slim Terms were chosen for visualization in ClueGO and had their results containing the Granular Terms and their associated genes exported to a .xlsx file. Function Analysis was performed with default parameters, except for Marker List: *S. cerevisiae*; Ontologies/Pathways: GO Biological Process GOA (updated 09-10-2021); Use GO term Fusion = True; GO tree interval: Min Level = 2 or 5, and Max Level = 8; GO Term/Pathway Selection: Min number of genes = 2 and percentage of genes = 1%. Cytoscape version 3.8.2 and ClueGO version v2.5.8.

The files containing the ClueGO results were subject to a python script developed by the synbiolab group (work in progress) that retrieves and formats the associated Granular Terms and genes for each Slim Term.

For both chromosomes a preliminary evaluation was performed in order to know the amount of granular terms, individually non-essential genes (iNEGs), individually non-essential genes associated with at least one Slim Term (aiNEGs), and individually non-essential genes associated with at least one granular term (gaiNEGs). The repetition of granular terms within the same Slim Term and gene repetition at different Slim Terms was also analysed in order to grasp the information needed for a GO-based synthetic lethality proposition.

Results

Syn2 and syn3 were accessed for the amount of resulting genes (figure 1) and Slim Terms (figure 2) resulting from each step of analysis.

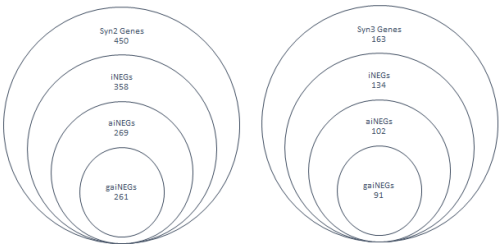

Fig 1. Stacked Venn diagram representing the number of syn2 (left) and syn3 (right) genes at different levels of associated data. iNEGs (individually nonessential); aiNEGs (individually non-essential genes associated with at least one Slim Term); gaiNEGs (individually non-essential genes associated with at least one granular term).

Only the gaiNEGs are eligible for further lethality assumption among GO groups. Of the 358 iNEGs in syn2, 23 were dubious or unannotated ORFs, or had invalid ID, while 66 were not mapped to a GO Term or only mapped to the root GO Term Biological Process according to the SGD's Slim Term Mapper. As for syn3, 7 out of 134 iNEGs were dubious or unannotated ORFs, or had invalid ID, and 25 out of 134 iNEGs were not mapped to a GO Term or only mapped to the root GO Term Biological Process.

Fig 2. Stacked venn diagram with the amount of Slim Terms for syn2 (upper) and syn3 (bottom). Only Slim Terms with 3 or more aiNEGs had those genes subject to ClueGO function Analysis, since the alternative hypothesis for our main question is if our GO-based criteria can distinguish gene groups into three levels of synthetic lethality assumption.

The number of granular terms that were obtained after the ClueGo function analysis was also accessed. For the syn2, the 77 granular terms resulted in 1694 granular terms, but most of them were shared across Slim Terms, with 466 being unique. For syn3, the total amount of Granular Terms was 215, with 134 unique terms.

Discussion

The proposed rationale is currently also being applied for the *S. cerevisiae* synthetic chromosomes II, V, IV and XII, and can be expanded for other organisms that have plenty of data and literature available. This analysis is an effort for a better rational design of synthetic genomes and investigation of gene essentiality. The Granular Terms for the synthetic chromosomes at the ultimate stage are the GO groups that will have their gaiNEGs investigated for putative synthetic lethality aiming at a better genome reduction design proposal.

The presented initial findings suggests that there is a increase in the number of Granular Terms in synthetic chromosomes with more iNEGs, but the overall loss of information for the synthetic lethality assumption might not be biased by the initial number of a chromosome, since the difference between aiNEGs and gaiNEGs is 8 for syn2 and 11 for syn3.

The next steps of the DBT cycle are String visualization of gaiNEG products interactions, manual curation of the pieces of evidence suggesting the interactions, the proposition of a synthetic lethality among the Granular Terms, and the *in vivo* validation of the proposed criteria.

It is important to point out that our goal is to propose a rational design for genome reduction to develop organisms that better suit research and industrial needs, a complete minimal genome will not be achieved by this rationale. Nevertheless, there are important factors when looking for a work that propose a genome reduction such as the evolving accuracy of genome and GO annotations, the enormous probabilistic landscape of genes and gene groups that might be synthetic lethal in a genomic background but not in another, and the philosophical nature of what is a minimal genome in a complex organism that has different habitats, that can grow in different conditions and that can both be manipulated in a laboratory controlled facility and survive in nature.

References

1. Pretorius, I. S. & Boeke, J. D. Yeast 2.0—connecting the dots in the construction of the world's first functional synthetic eukaryotic genome. FEMS Yeast Research 18, 32 (2018).
2. HutchisonIII, C. A. et al. Design and synthesis of a minimal bacterial genome. Science 351, (2016).
3. Annaluru, N. et al. Total synthesis of a functional designer eukaryotic chromosome. Science 344, 55–58 (2014).
4. Mercy, G. et al. 3D organization of synthetic and scrambled chromosomes. Science 355, (2017).
5. Shen, Y. et al. Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome. Science 355, (2017).