Genome annotation of two thousand hemiascomycete yeast strains: towards the development of an information system dedicated to Yeast Comparative Genomics and Pangenomics

Paulo Jorge Dias

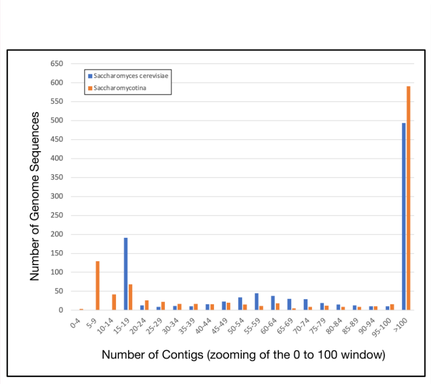
*1 iBB - Institute for Bioengineering and Biosciences, 2 Associate Laboratory i4HB - Institute for Health and Bioeconomy, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal, *Corresponding author: pjdias@tecnico.ulisboa.pt*

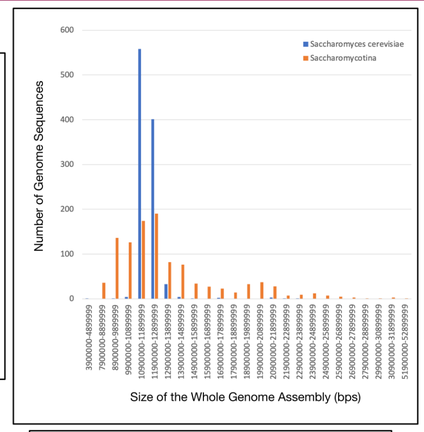
ABSTRACT

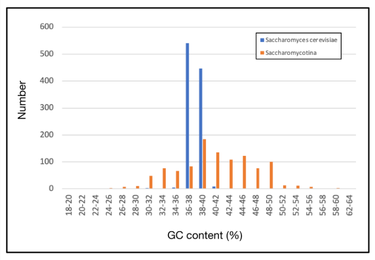
A decade ago, our research group started an initiative dedicated to the field of Comparative Genomics focusing on model-organism *S. cerevisiae* and the Hemiascomycetes. This initiative led to the development of two databases, the GenomeDB and BlastDB. The GenomeDB compiles genomic and biological information on the genes encoded in yeast species with publicly available genome sequences. Presently, the SaccharomycotinaDB spans 2074 hemiascomycete strains, corresponding to 528 yeast species. A SaccharomycesDB has also been compiled, spanning 1024 Saccharomyces cerevisiae strains. Funannotate software was used to annotate these genome sequences, allowing the identification of approximately 13 million genes. A wide range of assembly and annotation metrics on these yeast genomes were obtained, including number of contigs, total nucleotide length, GC content, gene number, tRNA number, average gene length, complete CDSs, no start CDSs, no stop CDSs, no start and no stop CDSs, single exon transcripts, multiple exon transcripts, average exon length, average protein length.

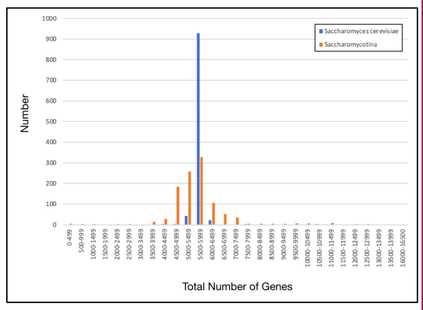
An online Information System is being developed to make available the genome annotation, the functional information on each gene and the Gene and Graph Pangenomes of yeast species to the research community. A pipeline developed in-house will also allow performing advanced Comparative Genomics approaches based on the gene family codes, including pairwise genome alignments and synteny analysis. To exemplify the power of this type of analysis, we have identified 20454 homologs of the *S. cerevisiae FLR1* gene and *C. albicans MDR1* gene in the sub-phylum Saccharomycotina and made the phylogenetic analysis of this Drug:H$^+$ Antiporter (DHA1) subfamily.

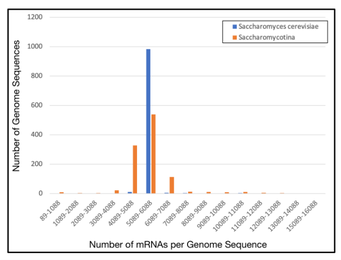
Genome Assembly and Annotation Metrics

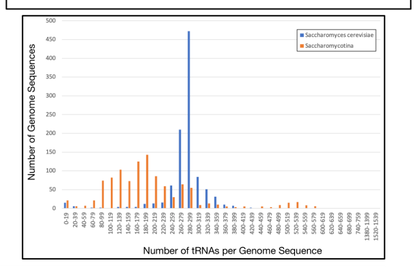


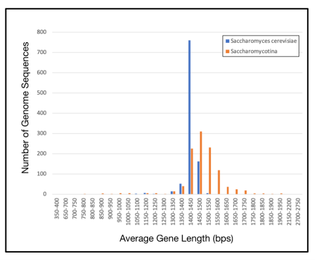


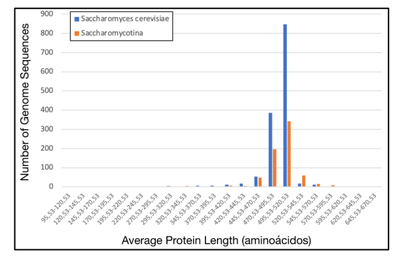


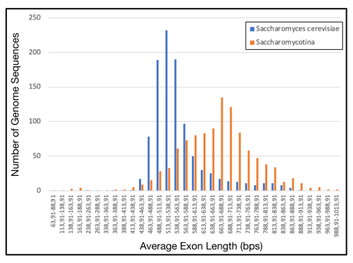


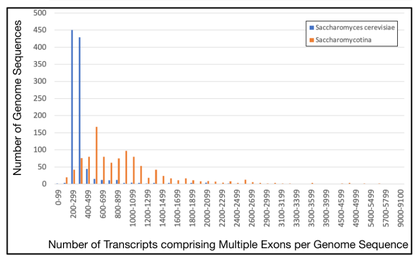


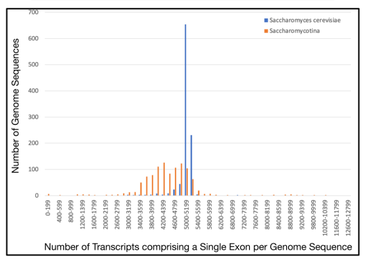


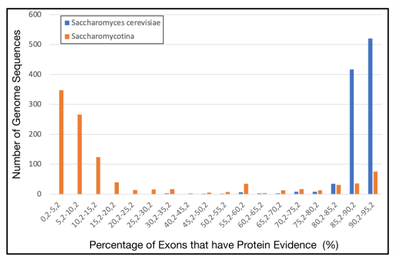


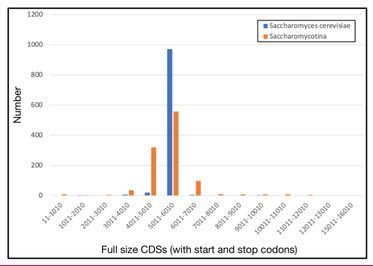


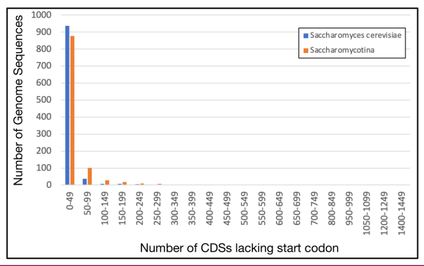


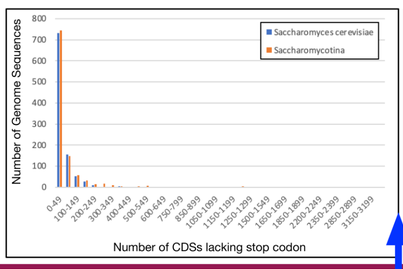


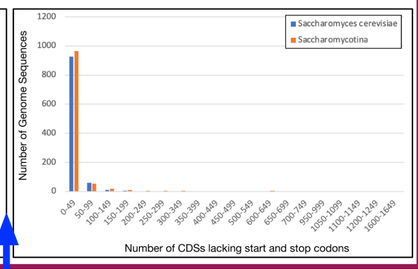


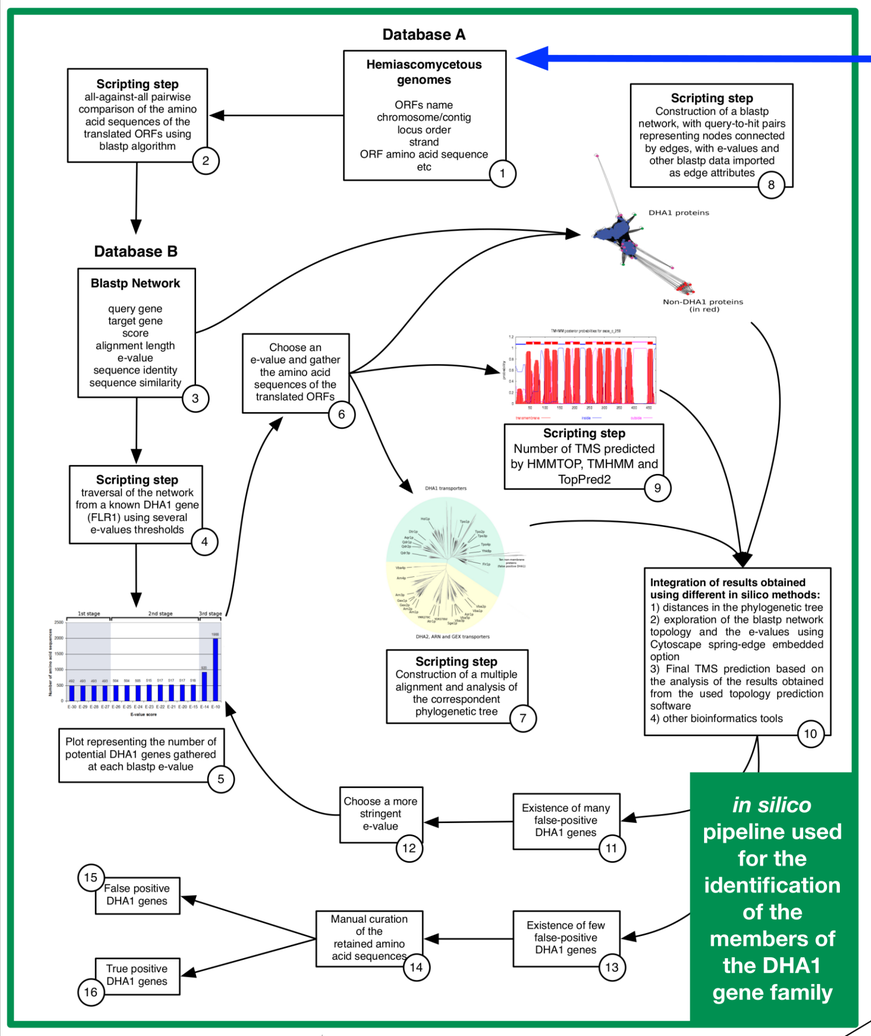




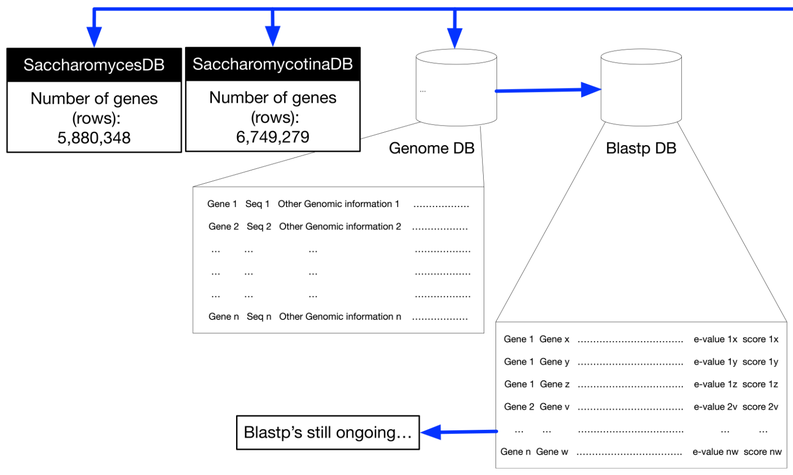


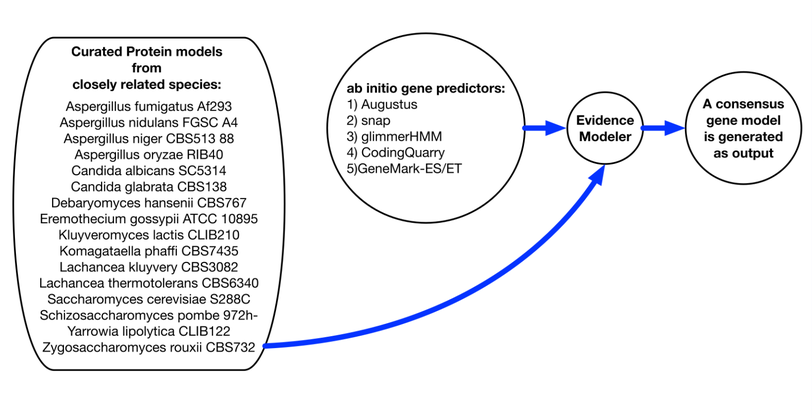


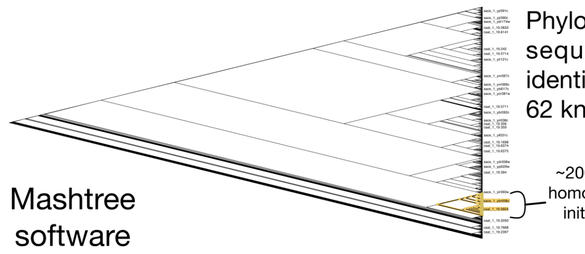




Gene annotation pipeline implemented by the Funannotate software suite





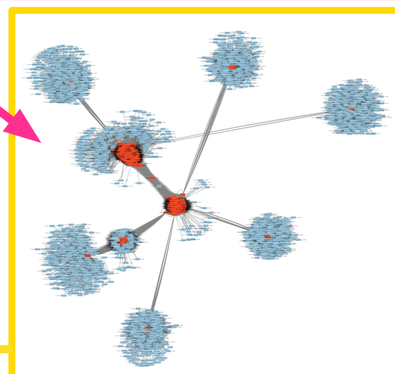


Phylogenetic tree constructed using the 20454 gene sequences of potential FLR1/MDR1 homologs identified using 4 traversals of the blastp network from 62 known homologs (Dias & Sá-Correia, 2010; 2014).

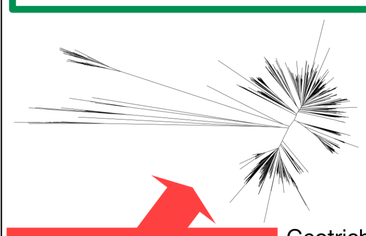
~2000 FLR1/MDR1 homologs among the initial 20454 DNA sequences

The Mashtree min-hash algorithm was used to convert the kmers into integers through a process involving hashing and a Bloom filter.

Mashtree software

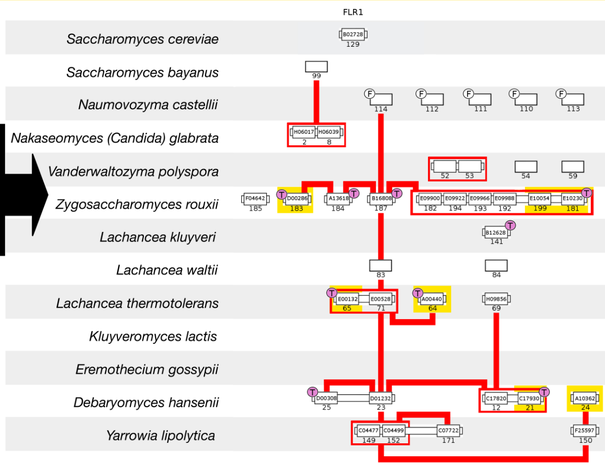


Network representing the blastp query hits between the 2129 FLR1/MDR1 homologs (red) and 4144 false positive proteins (blue). The layout of this network was calculated by the Cytoscape software suite using the number of blastps connections shared by these proteins. The topology of this network shows the retrieval of the false positives from a few FLR1/MDR1 homologs, showing that "noise" exists in the SaccharomycotinaDB.

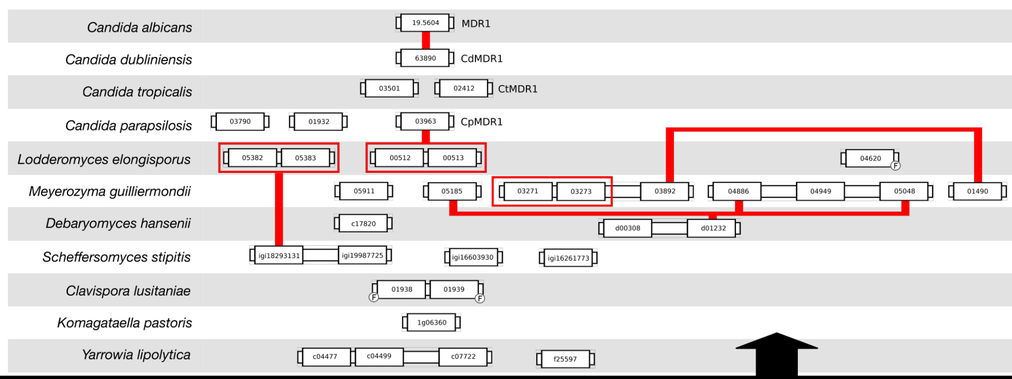


A) Phylogenetic tree representing the distances between the 2129 amino acid sequences of the FLR1/MDR1 homologs

B) Phylogenetic tree representing the topology between the 2129 amino acid sequences of the FLR1/MDR1 homologs



Synteny relationships between *FLR1* homologs encoded in the Saccharomycetaceae yeasts (Dias & Sá-Correia, 2010).



Synteny relationships between *MDR1* homologs encoded in the Debaryomycetaceae yeasts (Dias & Sá-Correia, 2014).