Inferring genome variation patterns in *Saccharomyces cerevisiae* using the eukaryote pan-genome toolset

Giltae Song[1], Benjamin Dickins[2], Stacia Engel[1], Travis Sheppard[1], Barbara Dunn[1], J. Michael Cherry[1]

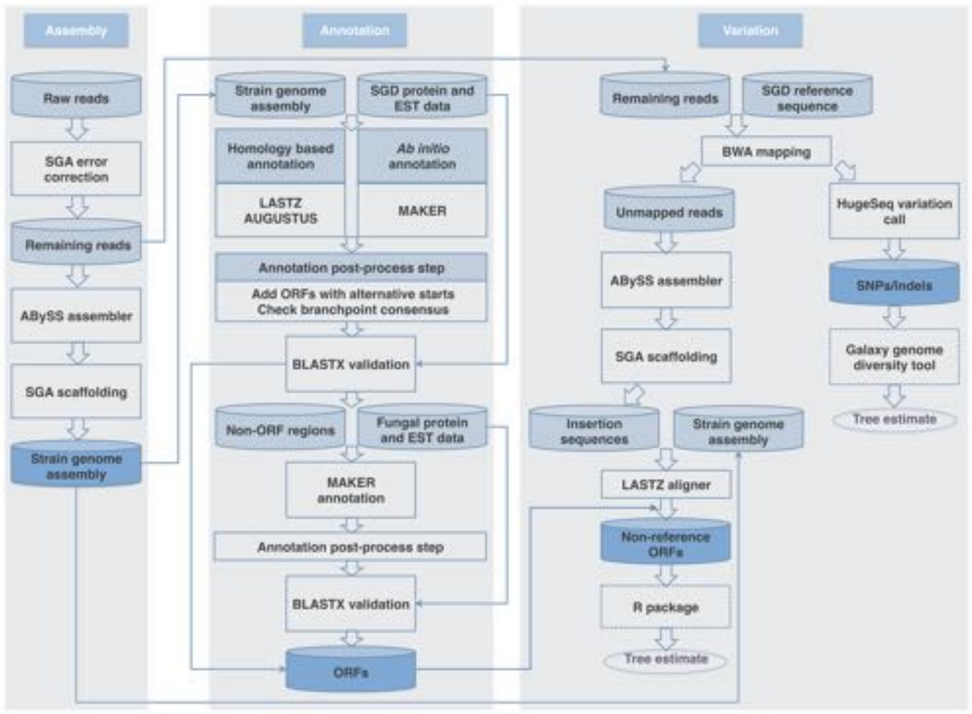
[1]Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305, USA.

[2]School of Science and Technology, Nottingham Trent University, Nottingham, UK

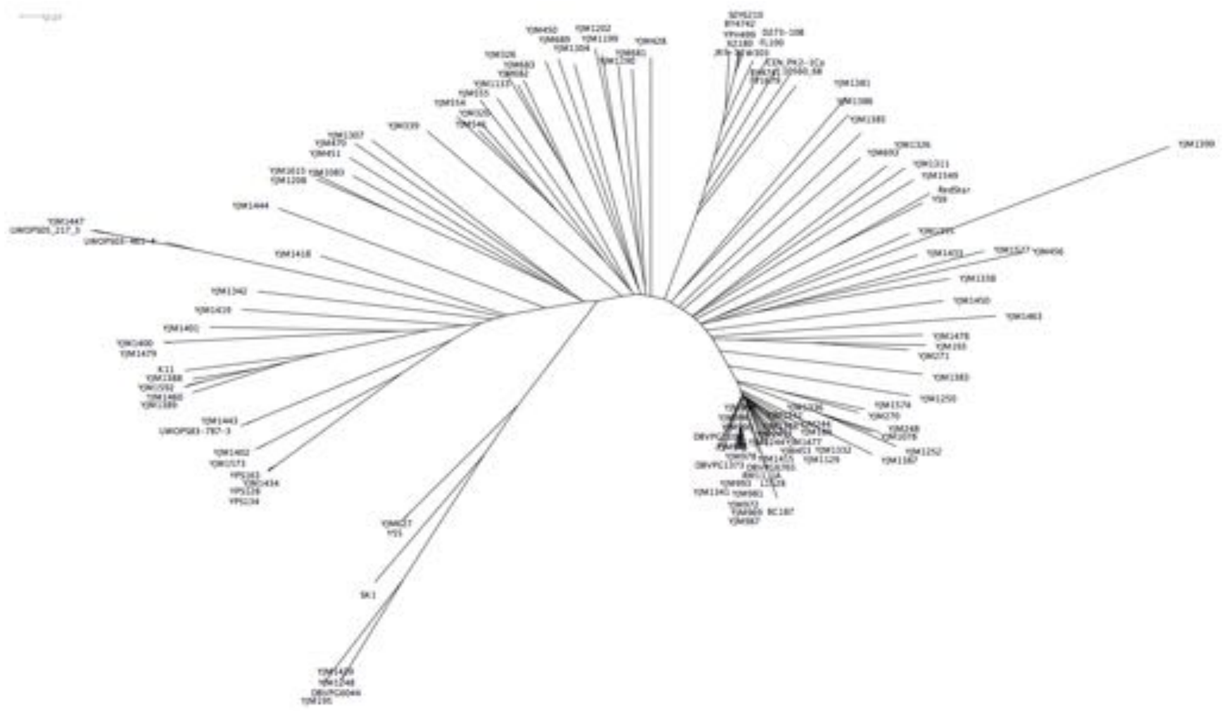
Despite the recent release of a hundred *Saccharomcyes cerevisiae* strain sequences, connecting population genome evolution with functional and phenotype variation in yeast remains challenging. This is due in part to the limitation of current comparative analyses, which rely on solely a reference genome defined from a single strain. Variation in genes that are not present in the reference genome is often overlooked, and cannot be examined with this method. Attempts have been made to overcome this problem in prokaryotes using pan-genome analysis, which aggregates sequence data from multiple strains to define a full set of genes within a species. However, these analyses have remained problematic for eukaryotic genomes due to their increasingly complex gene structures. We have developed a eukaryote pan-genome analysis pipeline for *S. cerevisiae*, and have made it freely accessible online. The pipeline includes steps for assembly, annotation, and variation-calling, and identifies novel genes that are not present in the S288C reference. Using the pipeline to analyze 125 *S. cerevisiae* strains, we have generated a composite pan-genome of *S. cerevisiae*. All genes can be assigned into core (always present), dispensable (sometimes present), and unique (rarely present) genomes. We investigate two aspects of variation within the pan-genome: (1) variation in the core genome, (2) patterns of absence or presence of genes in the dispensable and unique genomes. We have summarized this genetic variation with a phylogenetic tree based on Single Nucleotide Polymorphisms (SNPs) in the core genome.

AGAPE (Automated Genome Analysis PipelinE)

https://github.com/yeastgenome/AGAPE

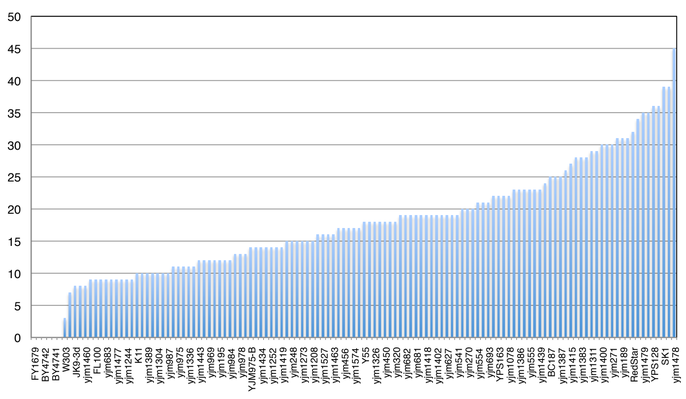


SNP based tree among 125 *S. cerevisiae* strains

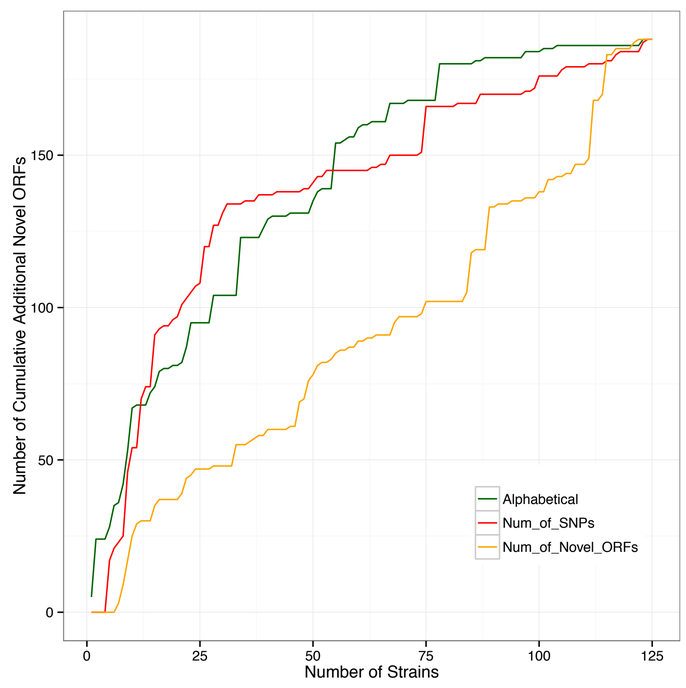


Number of Novel ORFs per strain

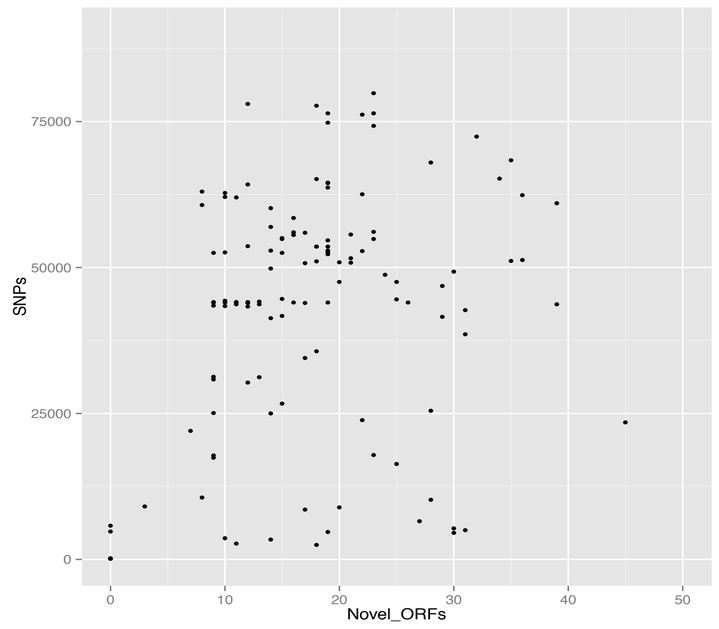
- Total 2206 novel ORFs are grouped to 188 homologues



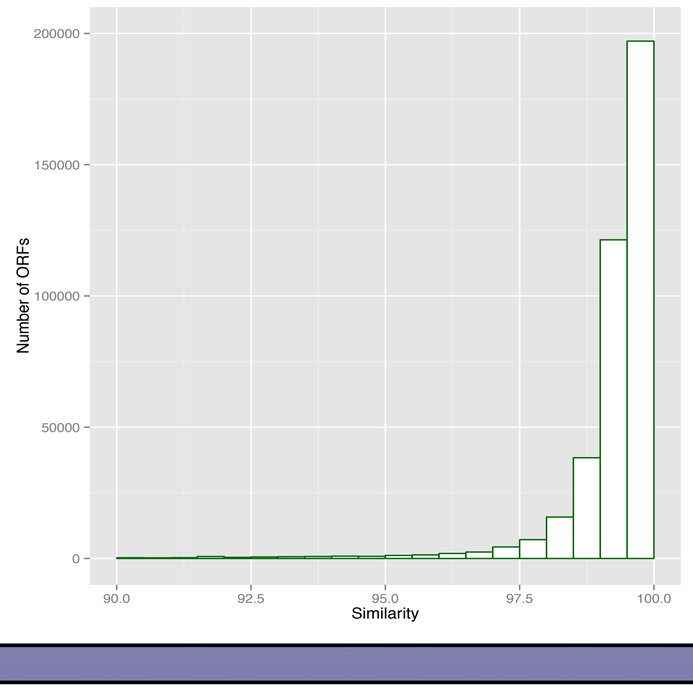
Saturation of the pan-genome



Correlation between number of novel ORFs and number of SNPs per strain

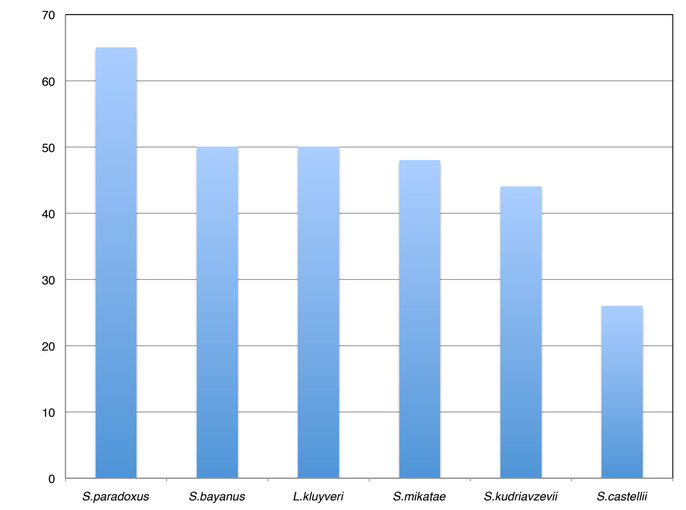


Majority of annotated ORFs are highly similar to the reference genome

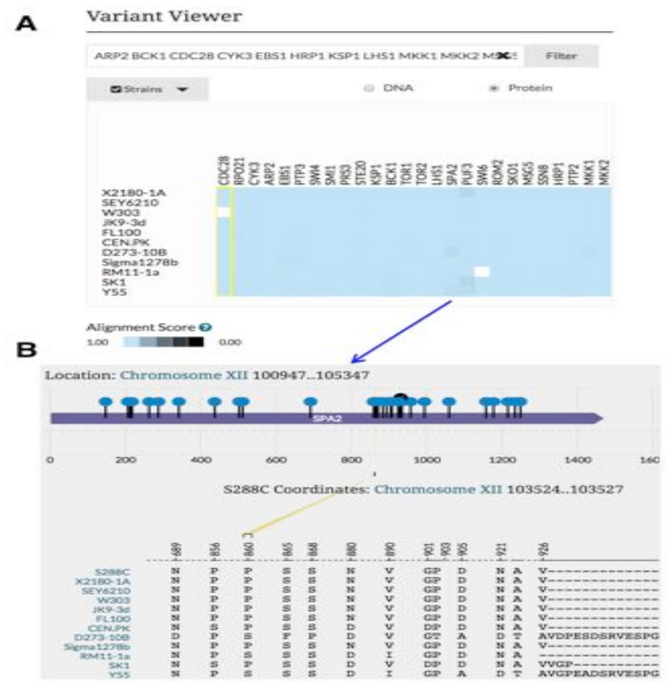


Novel ORFs also found in *S. sensu stricto* species

- 64 Novel ORFs are present in these fungi



SGD variant viewer



Visit SGD: http://www.yeastgenome.org/ https://twitter.com/yeastgenome Contact us: sgd-helpdesk@lists.stanford.edu