Beyond S288C: Incorporating Genomic Sequence Information from Large-Scale S. cerevisiae Population Surveys into SGD

Barbara Dunn, Stacia R. Engel, J. Michael Cherry, and The SGD Project
Department of Genetics, Stanford University School of Medicine, Stanford, CA

The Saccharomyces Genome Database (SGD; www.yeastgenome.org) began as a repository of the whole genome “reference sequence” of the S. cerevisiae S288C lab strain, the first sequenced eukaryotic genome. But there are now >1,500 different S. cerevisiae strains isolated from a wide variety of geographical locations and environmental niches with publicly-available whole genome sequences. Large-scale genomic comparisons using these data have allowed exploration of the genetic and phenotypic diversity of natural populations of yeast, helping elucidate the origins, present-day distribution, and standing genetic variation of this important organism. We are working toward incorporating the information from these data sets into SGD and we envision various ways we may store and display such information: the addition of new Locus Pages for open reading frames (ORFs) not found in the S288C reference genome, the identification and labeling of “core” ORFs (i.e., those shared by virtually all strains) vs. “variable” ORFs, and adding new information for sequenced strains, such as environmental niche, phylogenetic clade, and links to the genome sequence. We hope that providing easy access to information about strain variation as well as about the ecology and population dynamics of S. cerevisiae will be of use to the yeast community. Funded by the NHGRI, US NIH [5U41HG001315-18].

So many strains!

- Storing and accessing data
- How to incorporate the valuable info from these genomes into SGD?
  - ORFs not in S288C
  - Core vs. variable genes
  - Variation (SNPs, indels)
  - Ecological niche/phylogeny

Accessing 1011 strains study datasets via SGD’s reference page

Accessing other seq data by strain name

“Not-in-S288C” ORFs: Proposed nomenclature and new Locus tabs in SGD

Future plans

- Strain pages: add ecology and clade info; links to comparison tools, datasets.
- Many new strains in BLAST & other sequence tools.

Defining “Core” vs. “Variable” genes

- “Core” genes are those genes shared by all (or a vast majority) of strains within the S. cerevisiae species.
- “Variable” genes are all non-“core” genes and can range from very commonly-seen to extremely rare in the population.

The set of genes shared by >95% of strains overlaps very well with the core gene set defined by synteny (Dietrich unpub.; see below), making this a likely cutoff to use.