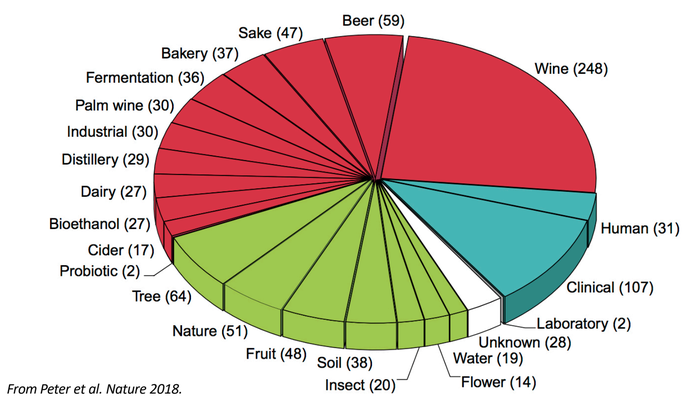
Beyond S288C: Incorporating new *S. cerevisiae* strain genomes and their associated not-in-S288C ORFs into SGD

Barbara Dunn, Stacia R. Engel, Gail Binkley, Stuart R. Miyasato, Travis K. Sheppard, J. Michael Cherry and The SGD Project
Stanford University, School of Medicine, Department of Genetics, Stanford, CA

The *Saccharomyces* Genome Database (SGD; www.yeastgenome.org) began as a repository of the genome sequence of *S. cerevisiae*, specifically the S288C lab strain, which was the first completely sequenced eukaryotic genome. There are currently >1500 different *S. cerevisiae* strains with whole genome sequences publicly available, with many more likely to be added in the coming months and years. Incorporating some or all of these data sets into SGD will eventually result in: the addition of many more "not-in-S288C" Locus Pages; the identification and labeling of "core" ORFs (i.e., those shared by virtually all whole-genome sequenced strains) vs. "variable" ORFs; the display of sequence variation in ORFs across many strains; and the creation of "Strain Pages" for sequenced strains, showing relevant isolation and phenotypic information and links to the genome sequence. We hope that the addition of these strain genomes and associated information will be of great use to the yeast community. This work is funded by the NHGRI, US NIH [5U41HG001315-18].

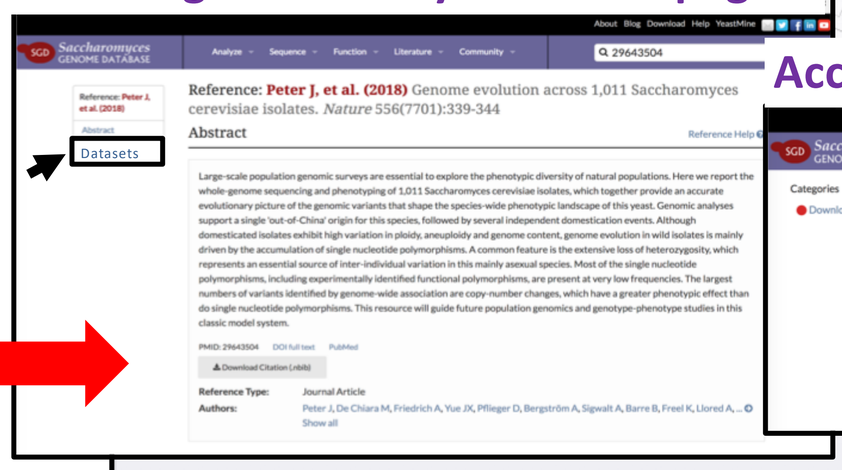
So many strains!



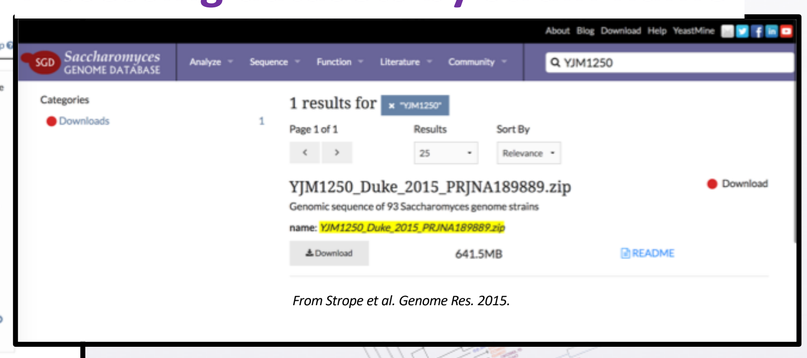
From Peter et al. Nature 2018.

- Storing and accessing data
- How incorporate the valuable info from these genomes into SGD?
  - -genes not in S288C
  - -variation (SNPs, indels)
  - -synteny

Accessing datasets by reference page



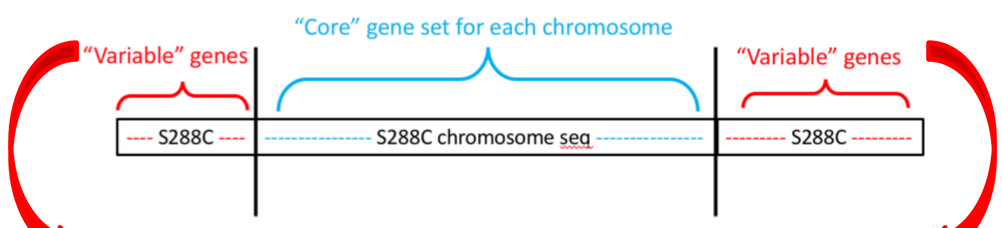
Accessing datasets by strain name



From Strope et al. Genome Res. 2015.

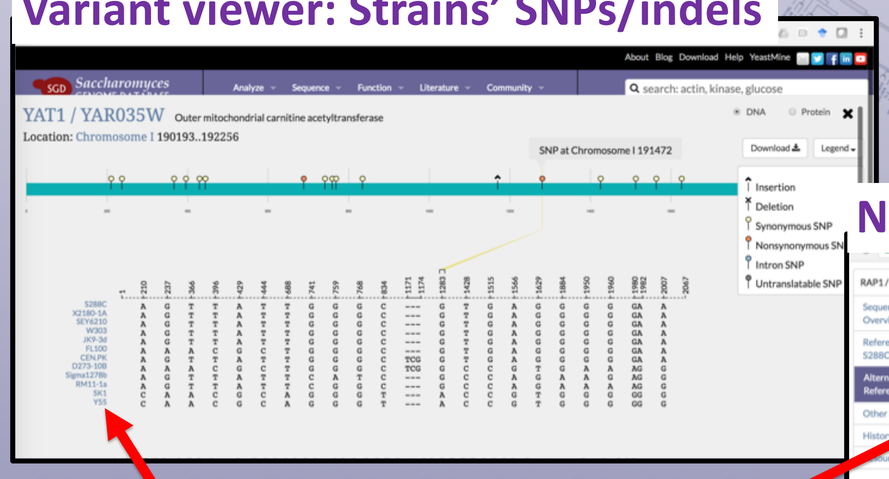
"Core" vs. "variable" genes

- A set of "core" genes for *S. cerevisiae* can be defined by shared synteny and homology (plus other characteristics) among a very large proportion of strains.
- One idea is that the **S288C** "core" genes would retain their systematic ORF names and standard gene names, and would be designated as "core" genes on Locus Pages…
- …while **non-S288C** "core" genes would be assigned systematic unique ORF names (not based on chromosome location, e.g., YSC####, with standard gene names also allowed). These would also have Locus Pages.



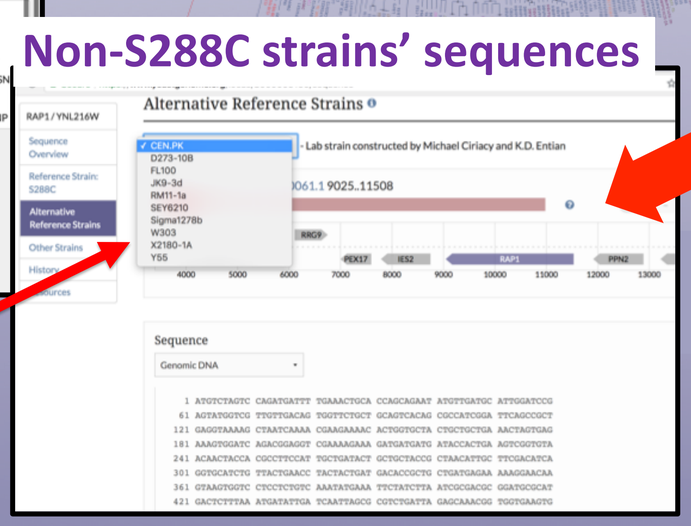
- Likewise, **S288C** genes **not** in the set of "core" genes would retain their systematic ORF names and standard gene names, and designated as "variable" genes on Locus Pages…
- …and **non-S288C** "variable" genes would be assigned systematic unique ORF names not based on chromosome location (standard gene names also allowed) and have Locus Pages.
- Core and variable genes could thus be obtained as separate gene lists.

Variant viewer: Strains' SNPs/indels



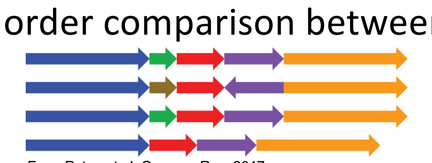
Representative strains from different ecological clades may be added in future

Non-S288C strains' sequences



Future plans/wish list

- Strain pages (all info about strain; links to comparison tools, datasets)
- Use 1-2 representative strains from each ecological clade for entry into JBrowse, Variant Viewer, Locus Sequence display…
- Many strains in BLAST & other seq tools
- Ability to input new strain sequence data and find out closest clade
- Display gene order comparison between strains, e.g.:



From Paten et al. Genome Res. 2017.

Background tree by Kristoffer Krogerus beer.suregork.com

Visit us: www.yeastgenome.org Contact us: sgd-helpdesk@lists.stanford.edu

Like: *Saccharomyces* Genome Database (SGD) Join: *Saccharomyces* Genome Database Subscribe: *Saccharomyces* Genome Database Follow: @yeastgenome