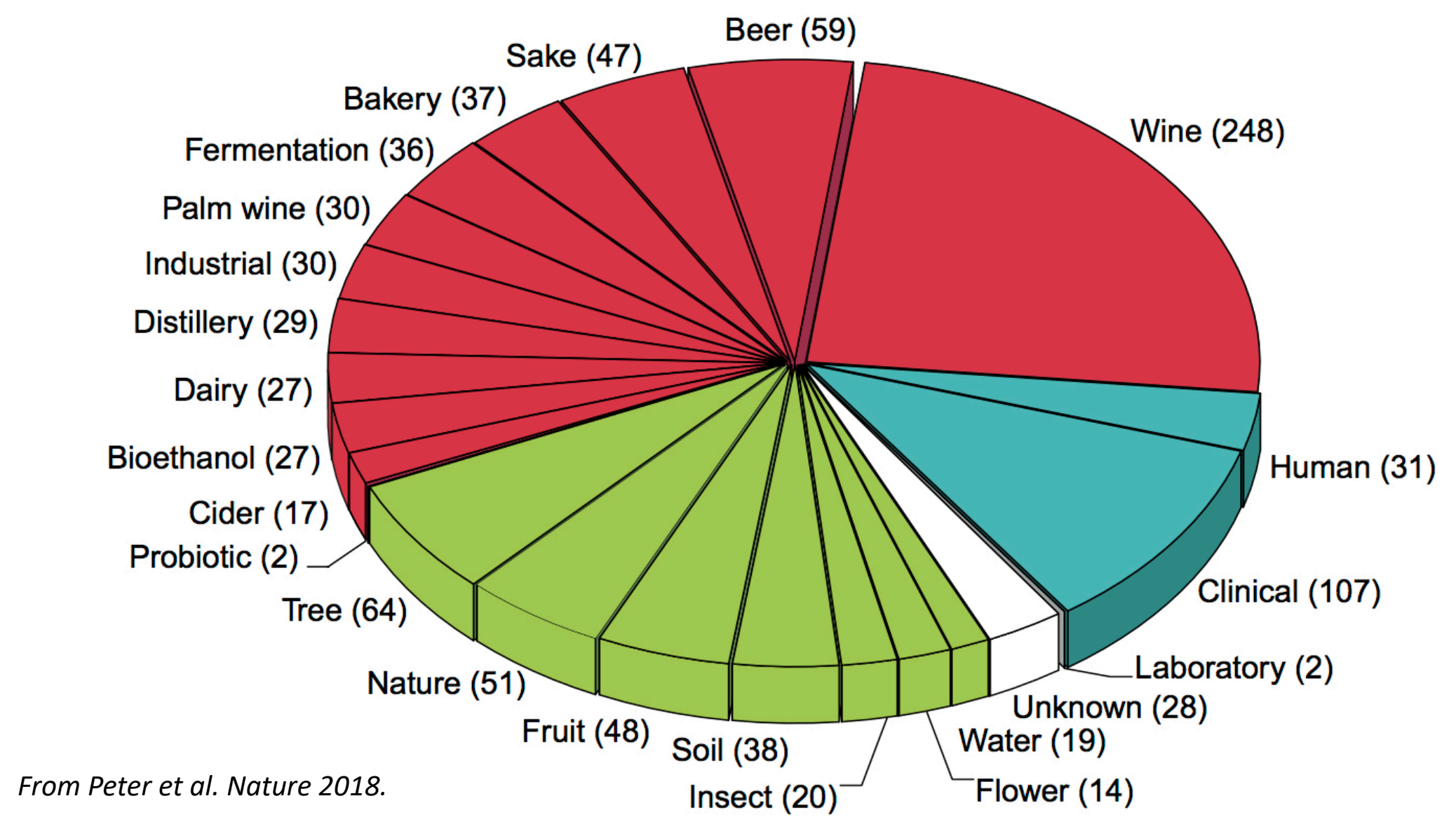
Beyond S288C: Incorporating Genomic Sequence Information from Large-Scale *S. cerevisiae* Population Surveys into SGD

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 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].

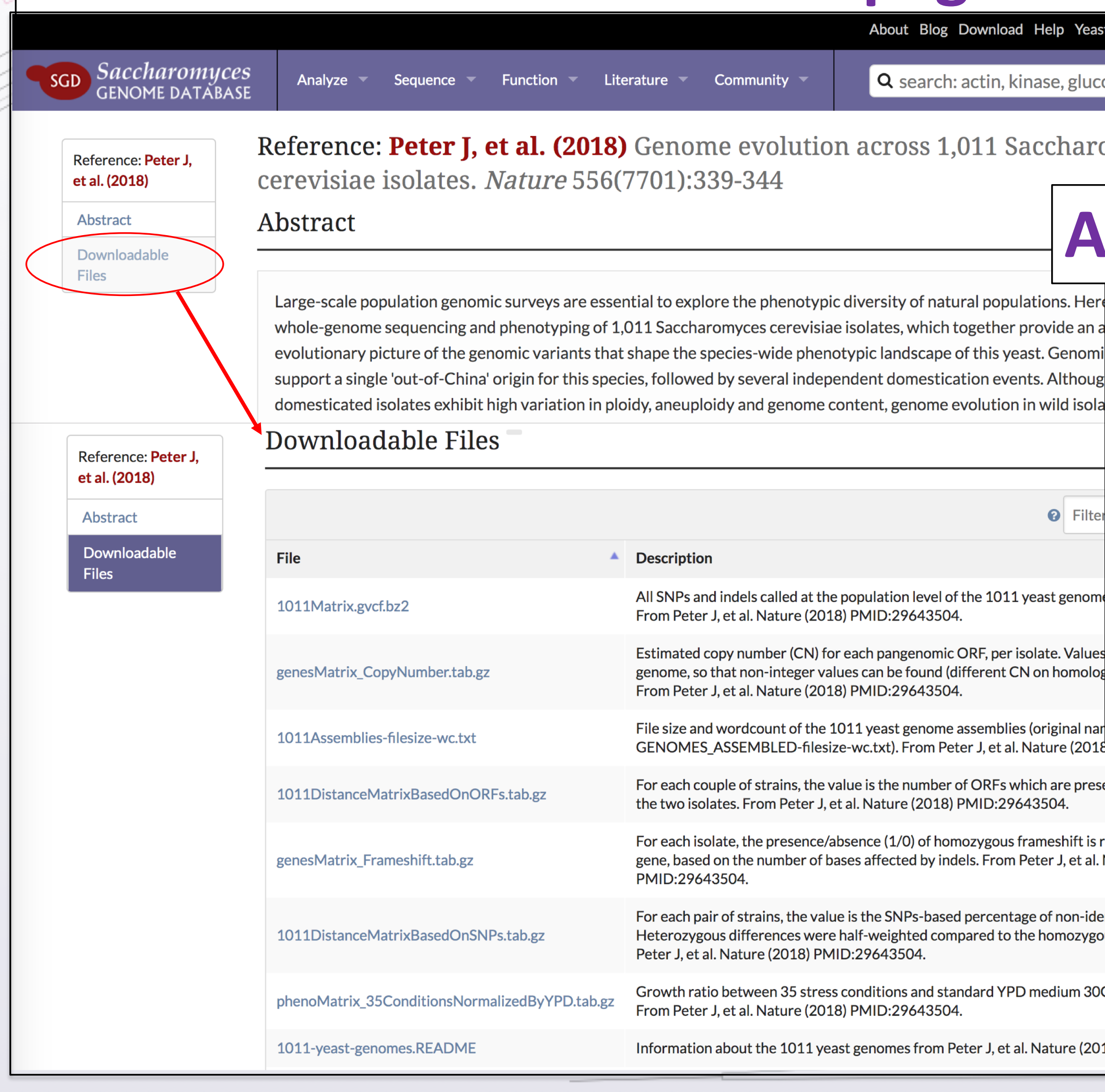
So many strains!



From Peter et al. Nature 2018.

- 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


About Blog Download Help YeastMine
Saccharomyces GENOME DATABASE
Analyze
Sequence
Function
Literature
Community
search: actin, kinase, glucose
Abstract
Downloadable Files
Filter

Reference: Peter J, et al. (2018) Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* 556(7701):339-344

Abstract

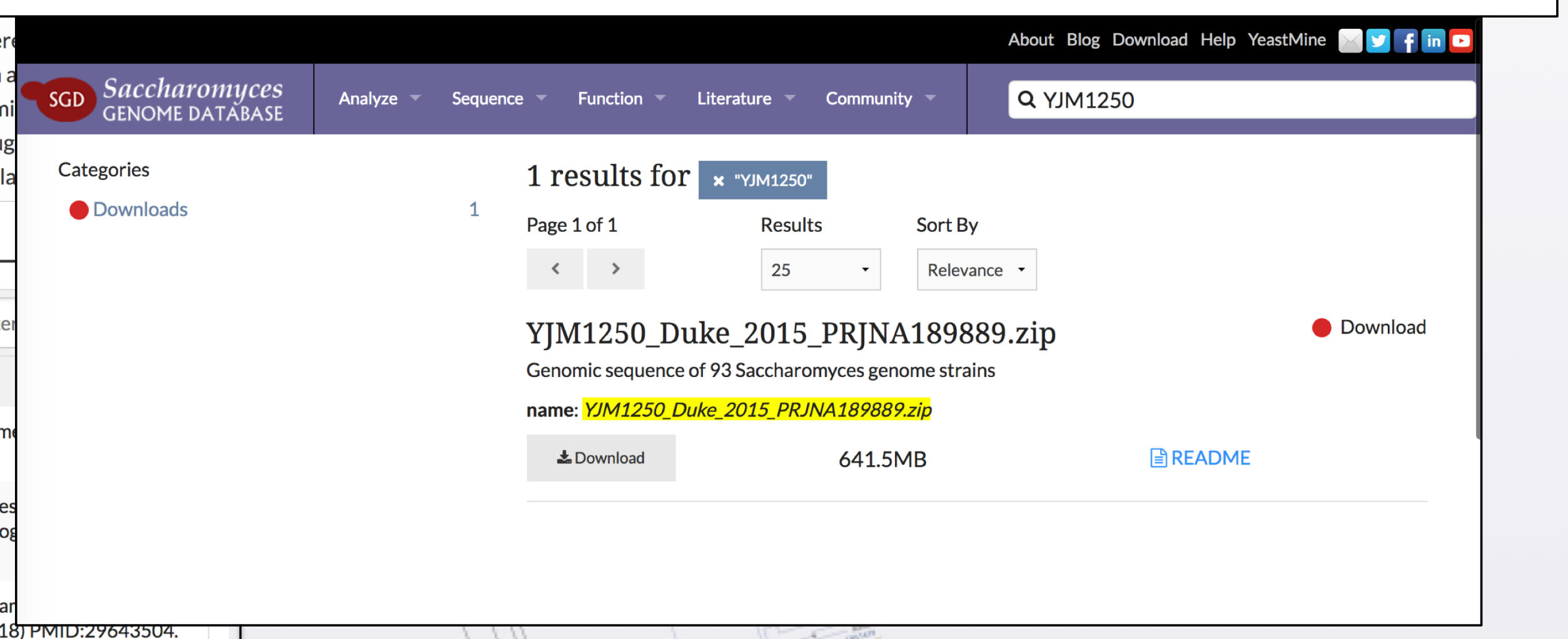
Large-scale population genomic surveys are essential to explore the phenotypic diversity of natural populations. Here whole-genome sequencing and phenotyping of 1,011 Saccharomyces cerevisiae isolates, which together provide an a evolutionary picture of the genomic variants that shape the species-wide phenotypic landscape of this yeast. Genomi support a single 'out-of-China' origin for this species, followed by several independent domestication events. Althoug domesticated isolates exhibit high variation in ploidy, aneuploidy and genome content, genome evolution in wild isola

Reference: Peter J, et al. (2018)

Downloadable Files

File	Description
1011Matrix.gvcf.bz2	All SNPs and indels called at the population level of the 1011 yeast genom From Peter J, et al. Nature (2018) PMID:29643504.
genesMatrix_CopyNumber.tab.gz	Estimated copy number (CN) for each pangenomic ORF, per isolate. Values genome, so that non-integer values can be found (different CN on homolog From Peter J, et al. Nature (2018) PMID:29643504.
1011Assemblies-filesize-wc.txt	File size and wordcount of the 1011 yeast genome assemblies (original nar GENOMES_ASSEMBLED-filesize-wc.txt). From Peter J, et al. Nature (2018) PMID:29643504.
1011DistanceMatrixBasedOnORFs.tab.gz	For each couple of strains, the value is the number of ORFs which are present only in one out of the two isolates. From Peter J, et al. Nature (2018) PMID:29643504.
genesMatrix_Frameshift.tab.gz	For each isolate, the presence/absence (1/0) of homozygous frameshift is reported in each gene, based on the number of bases affected by indels. From Peter J, et al. Nature (2018) PMID:29643504.
1011DistanceMatrixBasedOnSNPs.tab.gz	For each pair of strains, the value is the SNPs-based percentage of non-identical bases. Heterozygous differences were half-weighted compared to the homozygous differences. From Peter J, et al. Nature (2018) PMID:29643504.
phenoMatrix_35ConditionsNormalizedByYPD.tab.gz	Growth ratio between 35 stress conditions and standard YPD medium 30C for 971 isolates. From Peter J, et al. Nature (2018) PMID:29643504.
1011-yeast-genomes.README	Information about the 1011 yeast genomes from Peter J, et al. Nature (2018) PMID:29643504.

Accessing other seq data by strain name



1 results for "YJM1250"

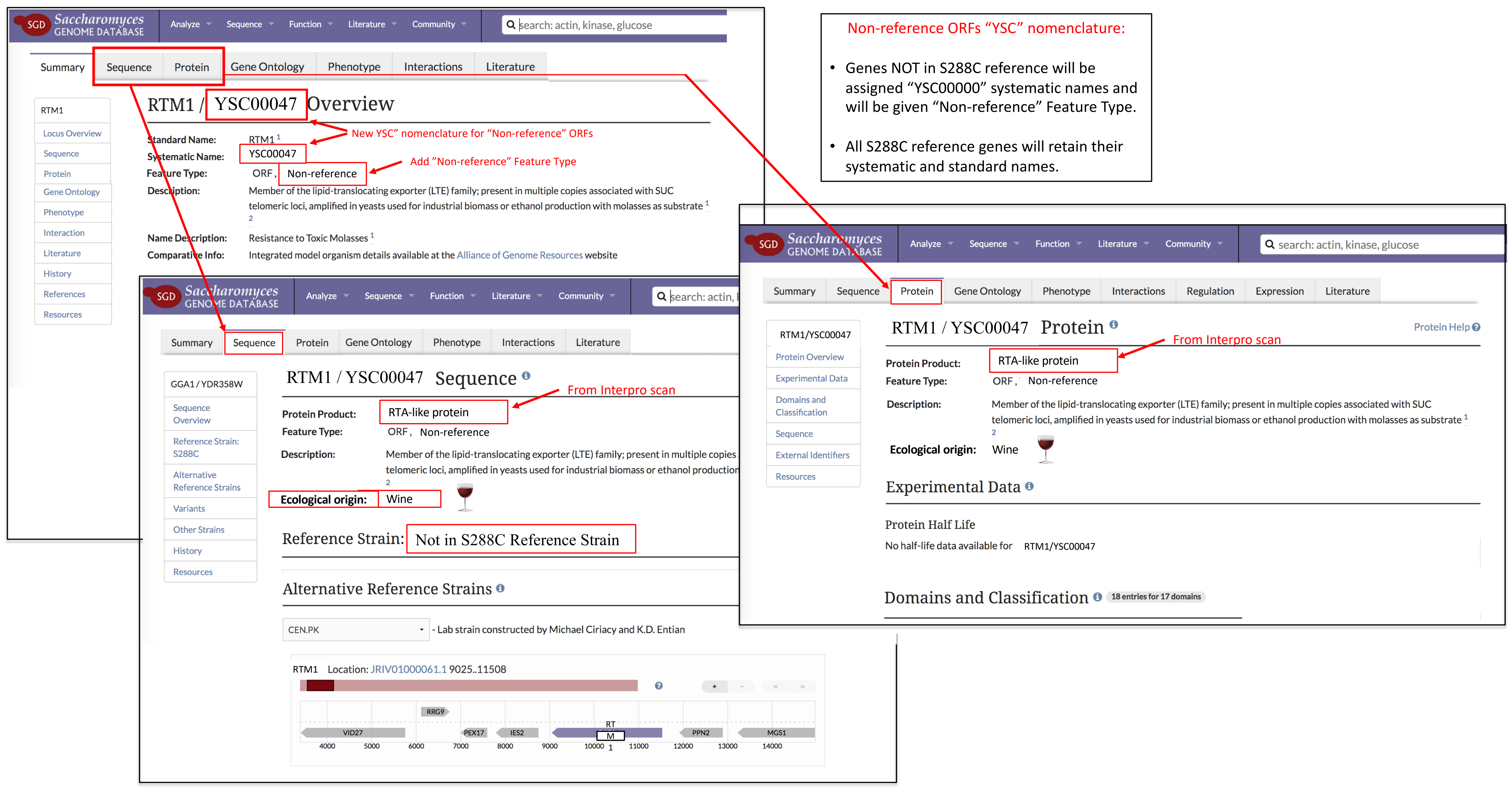
YJM1250_Duke_2015_PRJNA189889.zip

Genomic sequence of 93 Saccharomyces genome strains

name: *YJM1250_Duke_2015_PRJNA189889.zip*

641.5MB

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



RTM1 / YSC00047 Overview

New YSC" nomenclature for "Non-reference" ORFs

Standard Name: RTM1 [1]
Systematic Name: YSC00047
Feature Type: ORF , Non-reference

Add "Non-reference" Feature Type

Description: Member of the lipid-translocating exporter (LTE) family; present in multiple copies associated with SUC telomeric loci, amplified in yeasts used for industrial biomass or ethanol production with molasses as substrate [1] [2]
Name Description: Resistance to Toxic Molasses [1]
Comparative Info: Integrated model organism details available at the Alliance of Genome Resources website

RTM1 / YSC00047 Sequence

From Interpro scan

Protein Product: RTA-like protein
Feature Type: ORF , Non-reference
Description: Member of the lipid-translocating exporter (LTE) family; present in multiple copies telomeric loci, amplified in yeasts used for industrial biomass or ethanol production [2]
Ecological origin: Wine

Reference Strain: Not in S288C Reference Strain

Alternative Reference Strains

CEN.PK - Lab strain constructed by Michael Ciriacy and K.D. Entian

RTM1 Location: JRIV01000061.1 9025..11508

RTM1 / YSC00047 Protein

From Interpro scan

Protein Product: RTA-like protein
Feature Type: ORF , Non-reference
Description: Member of the lipid-translocating exporter (LTE) family; present in multiple copies associated with SUC telomeric loci, amplified in yeasts used for industrial biomass or ethanol production with molasses as substrate [1] [2]
Ecological origin: Wine

Experimental Data

Protein Half Life

No half-life data available for RTM1/YSC00047

Domains and Classification 18 entries for 17 domains

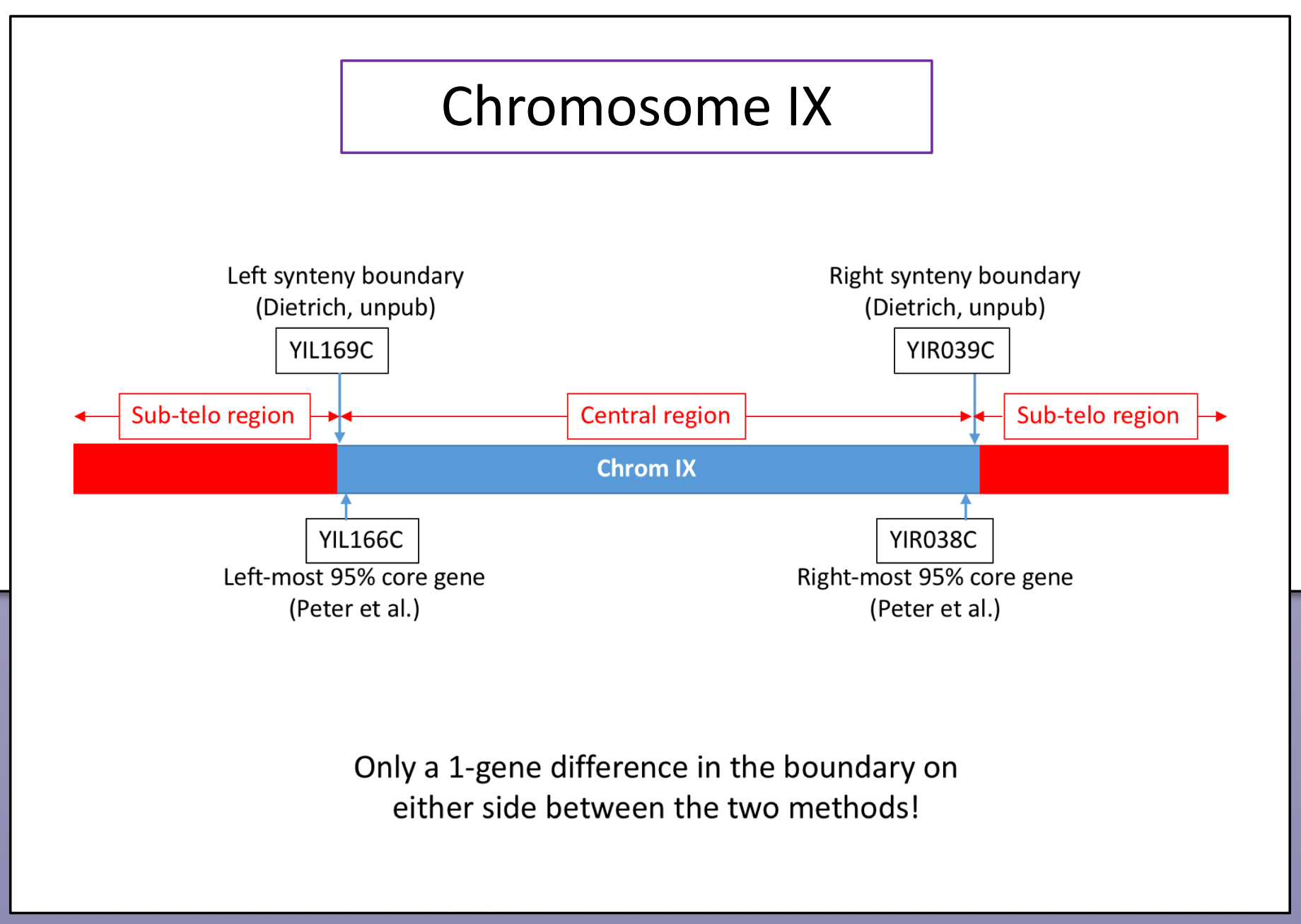
Non-reference ORFs "YSC" nomenclature:

- Genes NOT in S288C reference will be assigned "YSC00000" systematic names and will be given "Non-reference" Feature Type.
- All S288C reference genes will retain their systematic and standard names.

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.


Chromosome IX
Left synteny boundary (Dietrich, unpub)
YIL169C
Right synteny boundary (Dietrich, unpub)
YIR039C
Sub-telo region
Central region
Sub-telo region
Chrom IX
YIL166C
Left-most 95% core gene (Peter et al.)
YIR038C
Right-most 95% core gene (Peter et al.)

Only a 1-gene difference in the boundary on either side between the two methods!

Future plans

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

Background tree by Kristoffer Krogerus
beer.suregork.com

@yeastgenome
sgd-helpdesk@lists.stanford.edu
https://www.facebook.com/yeastgenome/
https://www.youtube.com/SaccharomycesGenomeDatabase

ALLIANCE of GENOME RESOURCES
FOUNDING MEMBER

Stanford MEDICINE