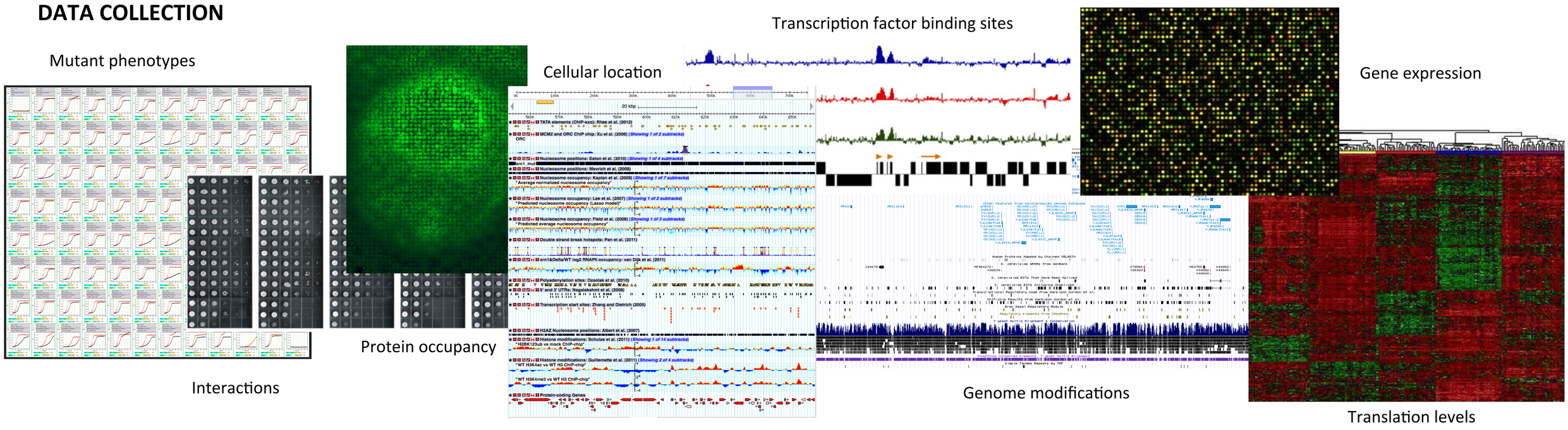
# Integrating multiple data types toward functional annotation of *Saccharomyces cerevisiae*

Edith D. Wong, Paul Lloyd, Janos Demeter, Giltae Song, Stacia Engel, and J. Michael Cherry
Stanford University, School of Medicine, Department of Genetics, Stanford, CA

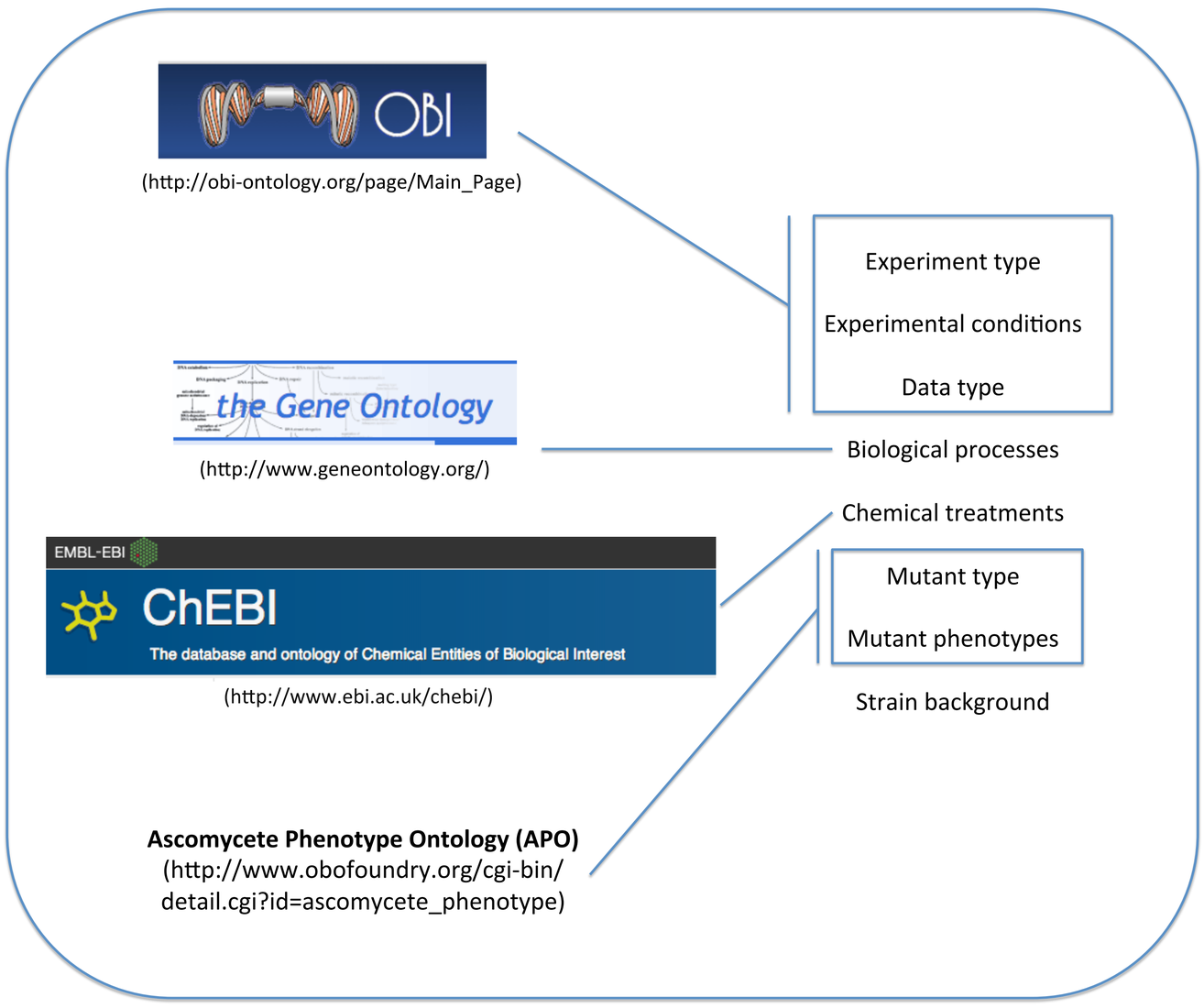
The *Saccharomyces* Genome Database (SGD; www.yeastgenome.org) is a comprehensive resource of curated molecular and genetic information on the genes and proteins of *Saccharomyces cerevisiae*. The emergence of large-scale, genome-wide technologies such as expression microarrays, RNA-seq and high-throughput sequencing have widened the scope of functional annotation beyond that of individual genes to entire genomes. As such, we have sequenced 25 *S. cerevisiae* strains, including both common laboratory and commercial strains, and are preparing these sequences to be incorporated into SGD. These new data allow us to compile and process sequence annotations in order to identify shared and divergent features between strains. In addition to new sequence data, we have collected published data from whole genome studies that employ a diverse set of techniques, including tiling arrays, cDNA clone libraries, TIF-seq, single and paired end RNA-seq, and serial analysis of gene expression (SAGE). These divergent techniques target different genomic regions, such as ncRNA, transcription start sites (TSS), transcripts, poly-A sites, antisense RNA, and Xrn1-sensitive unstable transcripts (XUTs), and add to our knowledge of biological processes in yeast. These data are being organized into a structured format, beginning with the compilation of various datasets to produce a complete yeast transcriptome. Here, we discuss computational approaches to process and to integrate these varying data with genomic sequences from different strains to form a more complete understanding of the complex processes of yeast biology.

## DATA COLLECTION



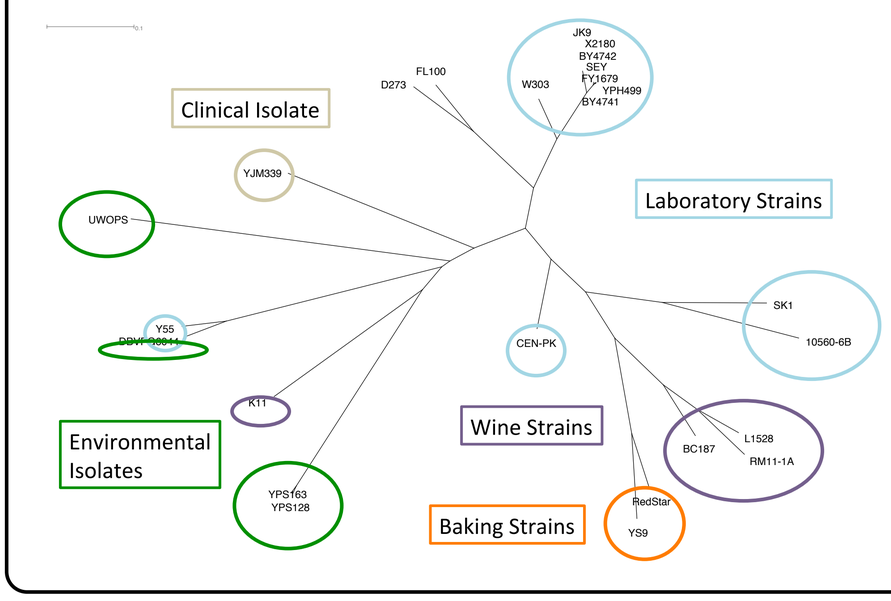

## LEVERAGING ONTOLOGIES AND METADATA

- Integrate datasets by capturing metadata using ontologies
- Use multiple ontologies to connect relevant datasets between strains and maximize information for users
- Focus on actively used and developed ontologies
- Capture metadata completely using a combination of ontologies




## FUNCTIONAL ANNOTATION OF MULTIPLE STRAINS

SNP based tree among 25 *S. cerevisiae* strains




- Use metadata to connect datasets to appropriate strain sequences
- Search for published data for genomic features that are not in the reference genome (S288C)
- Examine functional predictions of genomic features not in the S288C genomic sequence
- Create or update non-reference genome feature information
- Add additional strain sequences

Known features not present in the reference genome (S288C)

| | |
|---|---|
| **Standard Name** | *MEL1* |
| **Feature Type** | not in systematic sequence of S288C |
| **Description** | Secreted alpha-galactosidase; required for catabolic conversion of melibiose to glucose and galactose; regulated by several GAL genes (1, 2) |
| **Standard Name** | *RTM1* [1] |
| **Feature Type** | not in systematic sequence of S288C |
| **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) |
| **Standard Name** | *MPR1* [1] |
| **Feature Type** | not in systematic sequence of S288C |
| **Description** | L-azetidine-2-carboxylic acid acetyltransferase; reduces intracellular ROS and contributes to L-proline analog resistance and tolerance to ethanol and freezing; member of N-acetyltransferase superfamily; not found in S288C, only present in strains with Sigma1278b background; located on left arm of chromosome XIV approximately 15 kb from telomere (1, 2, 3, 4) |
| **Standard Name** | *BIO6* [1] |
| **Systematic Name** | YAR069W-A |
| **Feature Type** | not in systematic sequence of S288C |
| **Description** | Putative 7-keto-8-aminopelargonic acid (KAPA) synthetase; in sake strains and S. cerevisiae strains YJM627 and A364a; involved in the biotin biosynthesis pathway; homologs present in S. bayanus, S. paradoxus, S. mikatae and S. kudriavzevii (1, 2) |
| **Standard Name** | *TAT3* [1] |
| **Feature Type** | not in systematic sequence of S288C |
| **Description** | Permease identified in lager brewing yeast strain Weihenstephan Nr.34; also in RM11-1a, but not S288C; subject to nitrogen catabolite repression and post-translational control, cellular localization dependent on nitrogen source quality (1) |

Contact us at sgd-helpdesk@lists.stanford.edu

Visit us at www.yeastgenome.org

*Saccharomyces* Genome Database (SGD)

@yeastgenome

Subscribe to our YouTube Channel *Saccharomyces* Genome Database