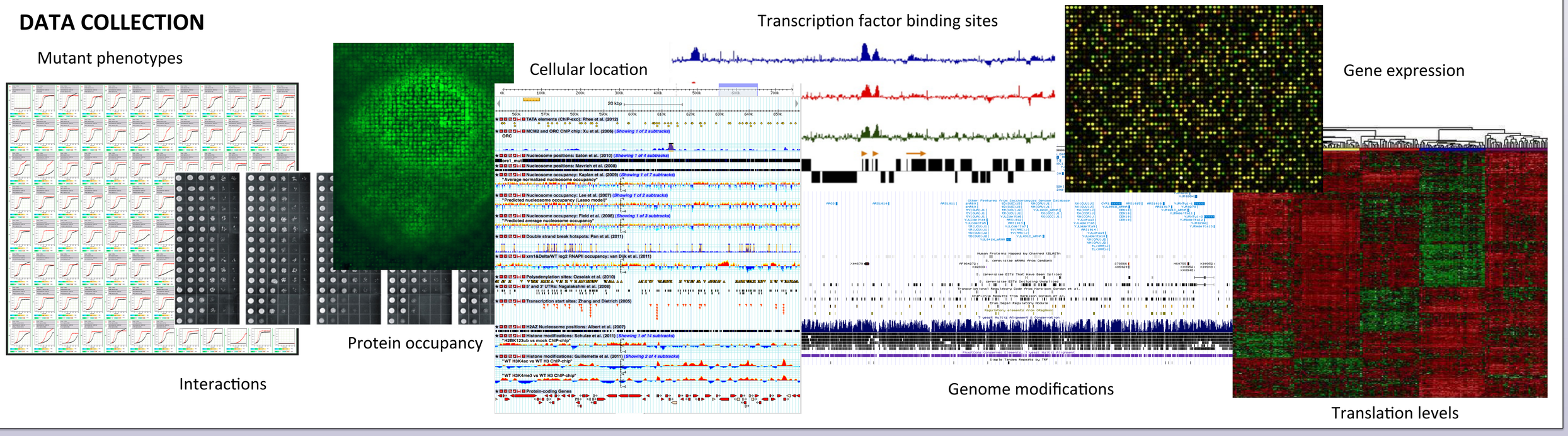
Integrating genome-wide datasets into the *Saccharomyces* Genome Database

Edith D. Wong, Janos Demeter, Robert S. Nash, Sage Hellerstedt, Kyla Dalusag, 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 has widened the scope of functional annotation beyond that of individual genes to entire genomes, allowing us to identify shared and divergent features between genomes. We have collected published data from whole-genome studies that employ a diverse set of modern techniques, including tiling arrays, cDNA clone libraries, TIF-seq, single and paired end RNA-seq, and serial analysis of gene expression (SAGE). These divergent methodologies target different genomic regions, such as ncRNA, transcription start sites (TSS), transcripts, poly(A) sites, and antisense RNA. Using ontologies and controlled vocabularies, metadata were curated from more than 1500 datasets from NCBI's GEO repository (Gene Expression Omnibus). These data will be available for straightforward querying at SGD via a faceted search tool to facilitate user access to yeast genomic data. This work is funded by the NHGRI, US NIH [5U41HG001315-18].

DATA COLLECTION



METADATA CURATION

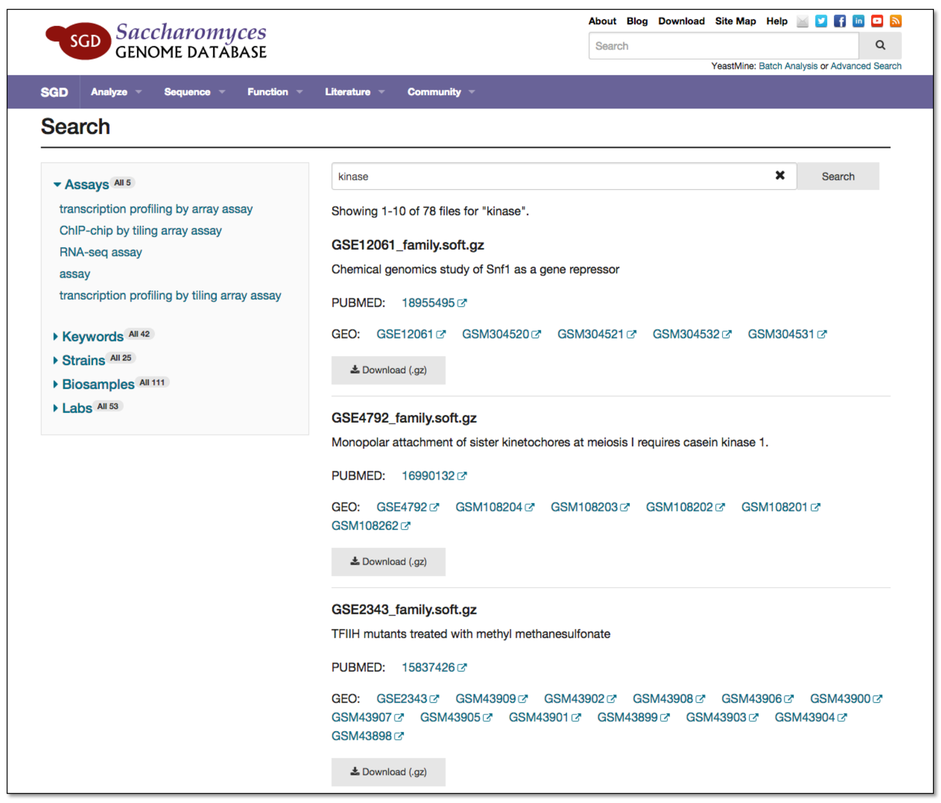
- Computationally parse metadata from SOFT files downloaded from Gene Expression Omnibus (GEO)
- Leverage multiple ontologies to maximize information for users
- Complete dataset coverage using combination of ontologies and controlled vocabularies
- Metadata from 1493 datasets from 1046 publications

Ontology	Metadata
http://obi-ontology.org	Experiment type Experimental conditions Data type
http://www.geneontology.org	Biological process
http://www.ebi.ac.uk/chebi/	Chemical treatments
Ascomycete Phenotype Ontology (APO) http://www.obofoundry.org	Mutant type Mutant phenotypes Strain background

SGD metadata key	GEO key
PMID	Series_pubmed_id
Author	First author
Lab	Last author
Dataset_description	Series summary|series title
Experiment description	Sample description
Assay term name (OBI)	Series_type_id
Biosample ID (OBI)	Series_type
Biosample Term name (OBI)	Sample_molecule_ch1_term
Strain background (controlled vocabulary)	Strain-ch1
Experiment dbxref	GEO accession
Datafile name	n/a
Category/Keyword (controlled vocabulary)	Manual curation

FACETED SEARCHING OF DATA

- Search across datasets using keywords
- Refine the search with facets:
  - Assays (ChIP-seq, ChIP-chip, etc)
  - Biosample (RNA, DNA, polyA RNA, etc)
  - Strain background (S288C, W303, etc)
- Retrieve datasets related to a process (e.g. cell aging, DNA damage, ubiquitin or UPL modification), or condition (e.g. heat shock, chemical stimulus, diauxic shift)
- Common ontologies (OBI, GO, and ChEBI) connect datasets to other relevant data types, such as phenotypes and gene functions



MAKING YOUR DATASETS AVAILABLE

- Submit datasets to the appropriate primary repository:
  - Gene Expression Omnibus (GEO)
  - Sequence Read Archive (SRA)
- Use ontology terms when submitting datasets:
  - Assays - OBI, Experimental Factor Ontology (EFO)
  - Biosamples - OBI, Sequence Ontology (SO)
  - Chemical treatments – ChEBI
- Include or update submissions with PMID after publication

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