Published online 15 October 2019 *Nucleic Acids Research, 2020, Vol. 48, Database issue* ***D743–D748***
doi: 10.1093/nar/gkz892

Transcriptome visualization and data availability at the *Saccharomyces* Genome Database

Patrick C. Ng[1], Edith D. Wong[1], Kevin A. MacPherson[2], Suzi Aleksander[1], Joanna Argasinska[1], Barbara Dunn[1], Robert S. Nash[1], Marek S. Skrzypek[1], Felix Gondwe[1], Sagar Jha[1], Kalpana Karra[1], Shuai Weng[1], Stuart Miyasato[1], Matt Simison[1], Stacia R. Engel[1] and J. Michael Cherry[1,*]

[1]Department of Genetics, Stanford University, Palo Alto, CA 94304-5477, USA and [2]Oregon Health Sciences University, Portland, OR 97239, USA

Received September 13, 2019; Editorial Decision September 27, 2019; Accepted October 07, 2019

ABSTRACT

The *Saccharomyces* Genome Database (SGD; www.yeastgenome.org) maintains the official annotation of all genes in the *Saccharomyces cerevisiae* reference genome and aims to elucidate the function of these genes and their products by integrating manually curated experimental data. Technological advances have allowed researchers to profile RNA expression and identify transcripts at high resolution. These data can be configured in web-based genome browser applications for display to the general public. Accordingly, SGD has incorporated published transcript isoform data in our instance of JBrowse, a genome visualization platform. This resource will help clarify *S. cerevisiae* biological processes by furthering studies of transcriptional regulation, untranslated regions, genome engineering, and expression quantification in *S. cerevisiae*.

INTRODUCTION

The annotation of >6000 genes in the reference genome of *Saccharomyces cerevisiae* is maintained by the *Saccharomyces* Genome Database (SGD; www.yeastgenome.org) (1), and is based on the common laboratory strain S288C (2). As a model organism database, SGD maintains a record of the sequence and chromosomal location of these gene features and manually curates functional annotation of their protein products in accordance with the guidelines of the Gene Ontology consortium (GO; www.geneontology.org) (3). This sequence information can be used to determine homology relationships across other organisms, and GO provides a controlled vocabulary and relational ontology for describing molecular functions, biological processes, or cellular components that may be shared by evolutionary conservation.

Precise gene mRNA sequence and coordinates are relevant to studies of mRNA stability (4), localization (5), and translational efficiency (6). Additionally, genome engineering projects seeking to alter or to vectorize the expression of *S. cerevisiae* genes (7) and transcript-based computational methods for measuring gene expression could also benefit from categorization of full-length mRNA transcripts (8). High-throughput next generation sequencing methodologies that measure the RNA expression of genes or map protein regulation of genomic DNA have become increasingly sensitive, making identification of these sequences easier.

In this paper we describe how SGD has taken data files and associated metadata from these RNA sequencing (RNA-seq) experiments, available at public repositories such as the Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo/) (9) and Array Express (www.ebi.ac.uk/arrayexpress/) (10), and visualized them using JBrowse (jbrowse.org) (11), a web-based genome browser application. We have divided datasets into tracks that either map assay values continuously across each position of the genome or highlight regions of interest identified experimentally. One of the categories of biochemical assays represented in SGD's JBrowse instance is the transcriptome: the identification of all RNA transcripts produced from the entire genome under particular conditions. The 5′ and 3′ untranslated regions flanking each gene are captured in these data tracks, where we aim to provide the research community with additional information about fundamental yeast cellular transcription.

INTEGRATION OF TRANSCRIPTOME DATA

A number of publications have separately sequenced the 5′ and/or 3′ ends of transcripts in *S. cerevisiae* (12–21). SGD has provided data from these studies to map the 5′ and 3′

[*]To whom correspondence should be addressed. Tel: +1 650 723 7541; Email: cherry@stanford.edu
Database URL: www.yeastgenome.org

Table 1. Filenames and descriptions of data tracks for transcript isoforms (GFF3 format) and coverage (bigWig format)

Data Track Filename	Description
longest_full-ORF_transcripts_ypd.gff3	This track contains the longest transcript overlapping each individual ORF completely for WT cells grown in glucose (ypd) media.
longest_full-ORF_transcripts_gal.gff3	This track contains the longest transcript overlapping each individual ORF completely for WT cells grown in galactose (gal) media.
most_abundant_full-ORF_transcripts_ypd.gff	This track contains the most abundant transcript overlapping each individual ORF completely for WT cells grown in glucose (ypd) media.
most_abundant_full-ORF_transcripts_gal.gff	This track contains the most abundant transcript overlapping each individual ORF completely for WT cells grown in galactose (gal) media.
unfiltered_full-ORF_transcripts.gff3	This track contains all transcripts that overlapped individual open reading frame (ORF) completely for WT cells grown in either glucose (ypd) or galactose (gal) media.
plus_strand_coverage_ypd.bw	For WT cells grown in glucose media (ypd), the amount of transcripts covering each position on the plus strand is represented in this track.
plus_strand_coverage_gal.bw	For WT cells grown in galactose media (gal), the amount of transcripts covering each position on the plus strand is represented in this track.
minus_strand_coverage_ypd.bw	For WT cells grown in glucose media (ypd), the amount of transcripts covering each position on the minus strand is represented in this track.
minus_strand_coverage_gal.bw	For WT cells grown in galactose media (gal), the amount of transcripts covering each position on the minus strand is represented in this track.

boundaries of mRNAs. However, Pelechano *et al.* developed Transcript Isoform Sequencing (TIF-seq) to characterize each individual transcript of the S288C strain (22). With this method, both ends of a single RNA molecule are identified at the same time. Thus, single complete transcripts are determined, rather than being inferred by matching the 5′ and 3′ ends that have been sequenced from different experiments. We have incorporated the data from Pelechano's study of two separate metabolic conditions (glucose- or galactose-containing media) to generate a representative view of the *S. cerevisiae* transcriptome (23).

First, we downloaded a text file of transcript coordinates and raw counts from GEO (accession GSE39128). We selected the subset in which each transcript's chromosomal location fully overlapped the protein coding sequence of a single open reading frame (ORF) annotated by SGD on the same strand. In order to compare across conditions, we combined transcripts from both conditions and gave them unique identifiers containing the systematic name of the associated SGD ORF. We first ordered transcripts by distance upstream of the start site of the associated ORF and then in descending order by transcript length, and finally created an output file using the General Feature Format (GFF) annotation format. Transcript identifiers are consistent across all files. For example, the YAL008W_id199 transcript isoform in the file with the most abundant transcripts found in yeast grown in glucose media (most_abundant_full-ORF_transcripts_ypd.gff3) corresponds to the same isoform in all other files; most_abundant_full-ORF_transcripts_gal.gff3, which contains the most abundant transcripts found under the galactose condition.

To begin to define a transcriptome, we initially created a set with all full-length transcripts for all ORFs (unfiltered_full-ORF_transcripts.gff3). Because growth condition affects what is being transcribed, we split this set into two different transcript sets, based on growth conditions (galactose or glucose). To depict the full range of what is transcribed for a particular ORF, for each growth condition, we then filtered the dataset for the longest transcript for each ORF. Finally, to indicate the predominant transcript isoform under each condition, we also created a most abundant transcript set. The GFF annotation filename suffixes for the longest and most abundant transcript sets denote whether they refer to the glucose nutritional condition (_ypd.gff3) or galactose nutritional condition (_gal.gff3). All filenames and descriptions can be found in Table 1.

To reflect how many transcripts covered each individual nucleotide of the *S. cerevisiae* genome, coverage tracks were generated for each condition from the raw transcript text file. Because of the size of the files, we split transcript coverage into plus and minus strands and created separate bigWig (.bw) files. The presence of intergenic, truncated and polycistronic transcription are also reflected in the provided files (Table 1).

TRACK VISUALIZATION AND ANNOTATION DOWNLOAD

SGD's JBrowse instance is accessible through the 'Genome Browser' link within the 'Sequence' menu in the purple toolbar that runs across the top of most SGD webpages, or via direct URL (browse.yeastgenome.org). Within the JBrowse browser window, the TIF-seq transcriptome tracks can be viewed by using the 'Select tracks' button in the top left corner. In the resulting slide out window, nine data tracks comprise the transcriptome (unfiltered transcripts that fully overlap ORFs, longest transcript in each of two conditions, most abundant transcript in each of two conditions, and both plus and minus strand coverage in each of two conditions), and can be navigated to in several ways using the categorical track selector to the left or the text query box at the top. Choosing 'Pelechano' within the 'First Author' category or '23615609' within the 'PMID' category and checking the leftmost boxes for each track in the metadata display table results in the tracks being viewable in the JBrowse navigation window. These instructions are also reviewed in a video tutorial on SGD's YouTube page (www.youtube.com/SaccharomycesGenomeDatabase). A recent post on the SGD Blog (www.yeastgenome.org/blog/explore-the-s288c-transcriptome-in-jbrowse) has the YouTube tutorial embedded and provides direct links to the tracks in JBrowse,

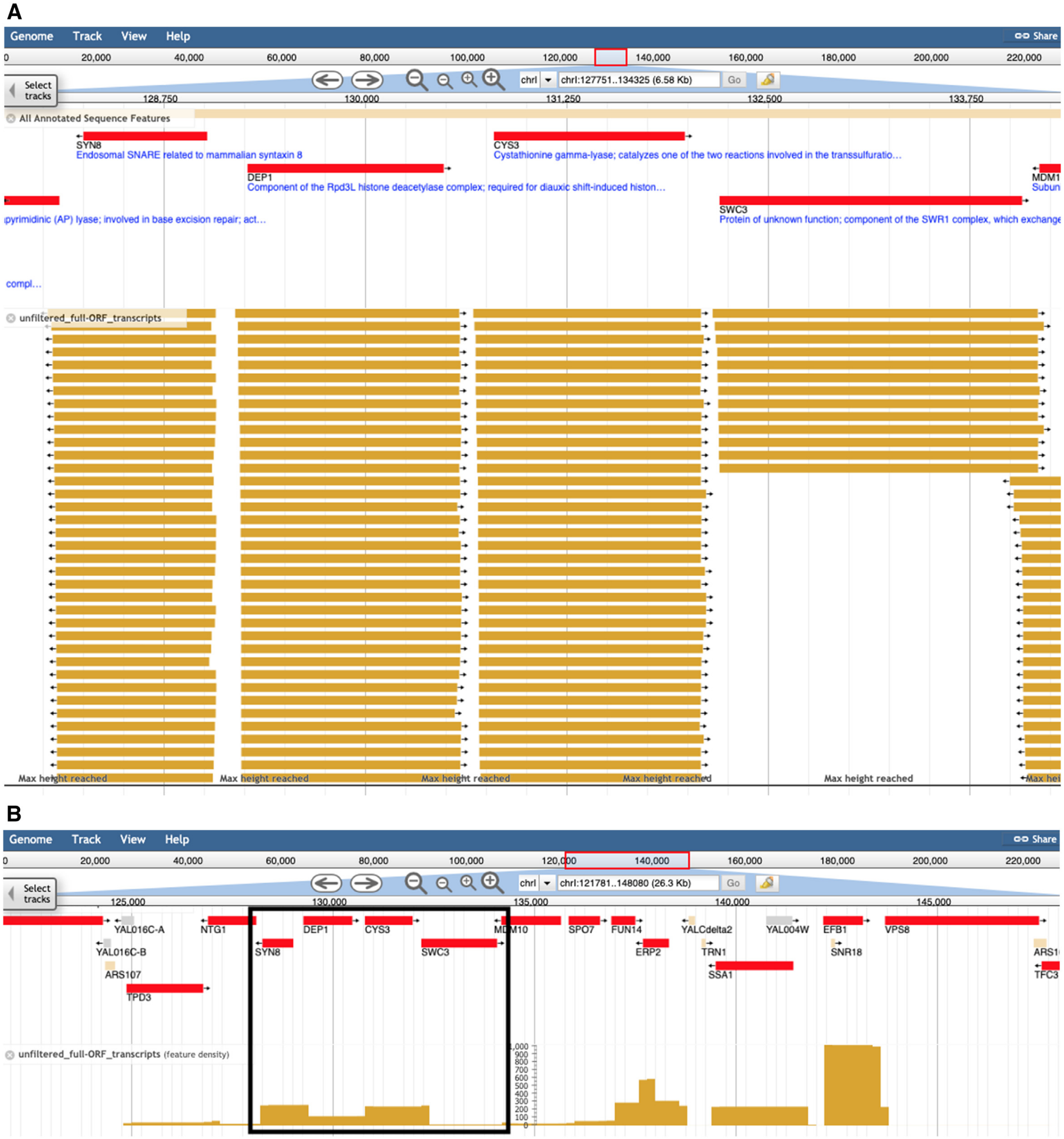
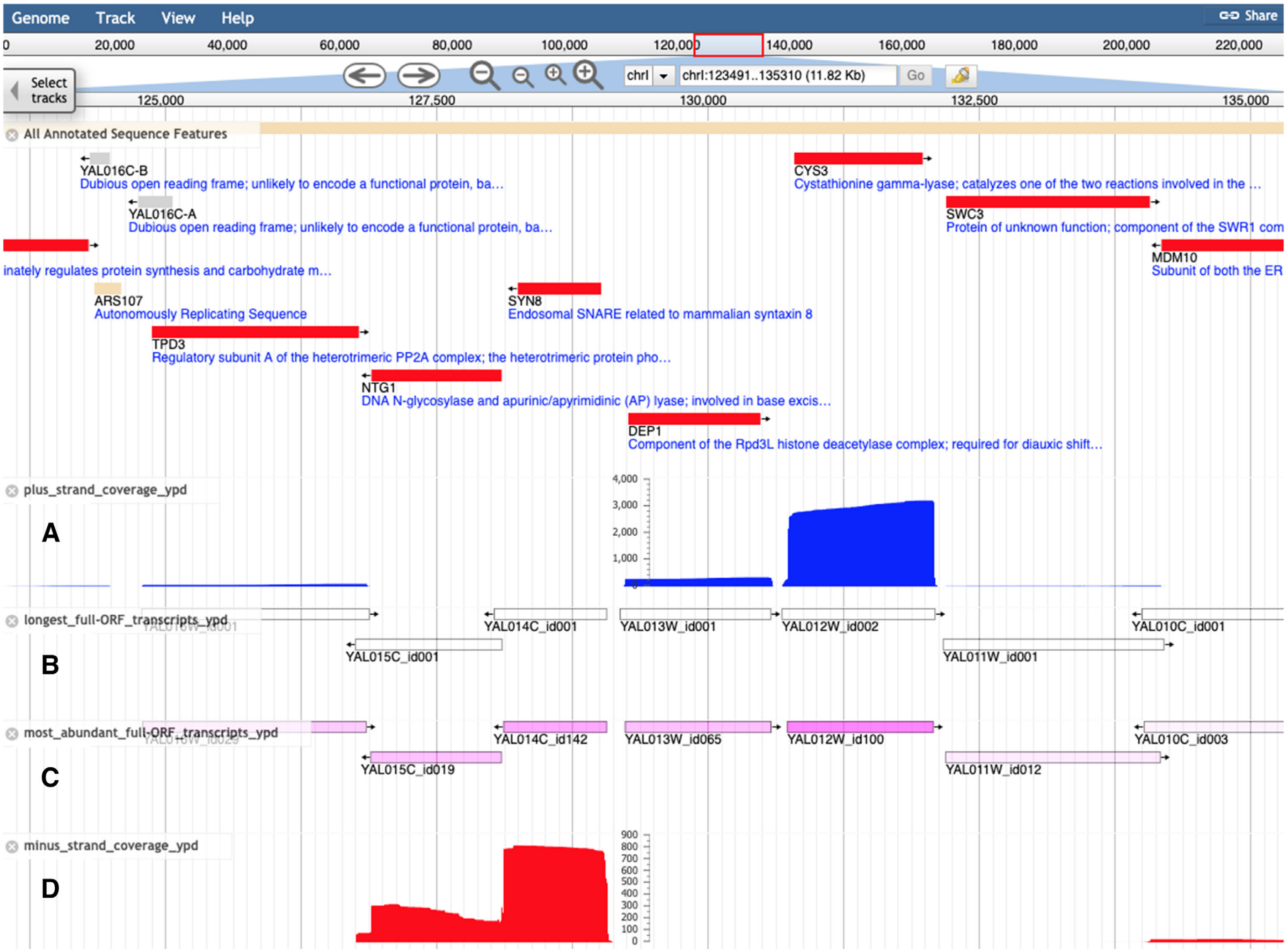


Figure 1. Unfiltered transcript isoforms that overlap ORFs displayed at various levels of zoom in JBrowse. (**A**) Individual glyphs representing each transcript isoform are visible at close zoom. (**B**) At lower magnification, the same region (outlined in the box) is displayed as a collapsed histogram.

and includes a direct download link to a zipped folder of all the track files (https://sgd-dev-upload.s3.amazonaws.com/S000246061/Pelechano_2013_PMID_23615609.zip).

Once displayed in the JBrowse navigation window, tracks can be distinguished by color. Unfiltered transcripts are displayed in solid yellow and can number up to the hundreds. By default, the browser clips the number of transcripts viewed at close zoom or collapses the track into a 'density' view at far zoom (Figure 1). For the most abundant and longest transcript tracks, transcript identifiers are displayed beneath the glyph (Figure 2B, C). Pink shading represents the logarithmically scaled transcript abundance; darker shading reflects higher abundance. Clicking on the glyph for a transcript reveals a popup listing its exact coor-



YAL016C-B
Dubious open reading frame; unlikely to encode a functional protein, ba…

YAL016C-A
Dubious open reading frame; unlikely to encode a functional protein, ba…

inately regulates protein synthesis and carbohydrate m…

ARS107
Autonomously Replicating Sequence

TPD3
Regulatory subunit A of the heterotrimeric PP2A complex; the heterotrimeric protein pho…

NTG1
DNA N-glycosylase and apurinic/apyrimidinic (AP) lyase; involved in base excis…

SYN8
Endosomal SNARE related to mammalian syntaxin 8

DEP1
Component of the Rpd3L histone deacetylase complex; required for diauxic shift…

CYS3
Cystathionine gamma-lyase; catalyzes one of the two reactions involved in the …

SWC3
Protein of unknown function; component of the SWR1 com

MDM10
Subunit of both the ER

YAL016W_id001, YAL015C_id001, YAL014C_id001, YAL013W_id001, YAL012W_id002, YAL011W_id001, YAL010C_id001

YAL016W_id029, YAL015C_id019, YAL014C_id142, YAL013W_id065, YAL012W_id100, YAL011W_id012, YAL010C_id003

Figure 2. Representative tracks for transcript isoforms and coverage in glucose containing media. (**A**) Plus-strand transcript coverage in blue. (**B**) Longest transcript isoform for each ORF. (**C**) Most abundant transcript isoform for each ORF. (**D**) Minus-strand transcript coverage in red.

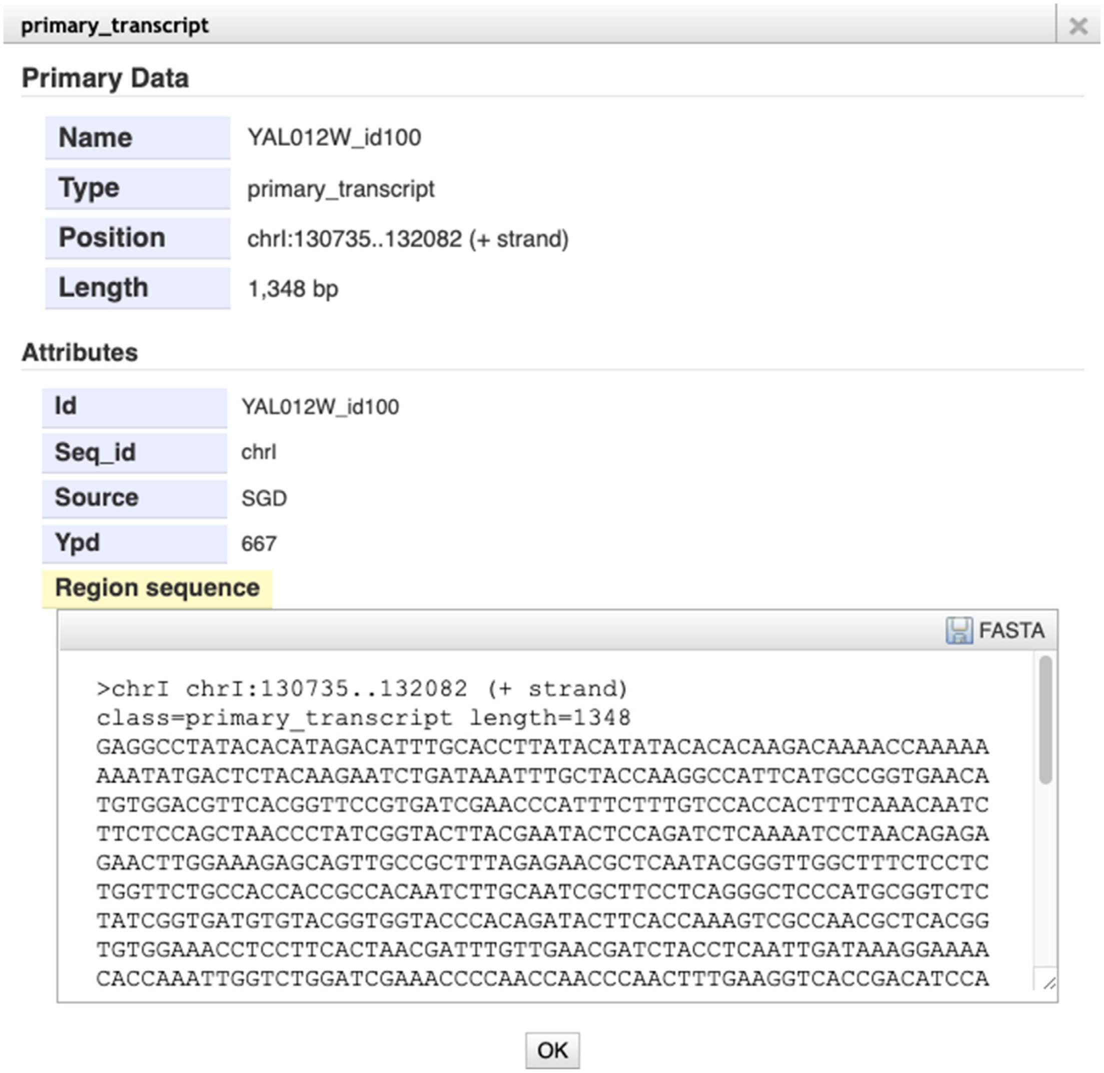
dinates, raw abundance in the particular media condition, and predicted sequence based on the reference genome (Figure 3). Quantitative coverage tracks for each condition are presented as histograms; blue for the plus strand and red for the minus strand (Figure 2A, D). These tracks represent the cumulative raw abundances of all transcripts at each position of the S288C reference sequence.

FUTURE DIRECTIONS

There are multiple ways to expand the transcriptome data that SGD provides. Pelechano's dataset examines transcripts and their abundance at specific glucose/galactose concentrations for mid-log phase cells. However, additional datasets, such as those from experiments that examine conditions utilizing different chemical, genetic or epigenetic perturbations, as well as extended time courses, can be incorporated. Large heterogeneity between individual transcripts for the same gene was a key observation of the Pelechano study. Incorporation of single cell sequencing methodologies could clarify the varied transcriptional landscape between individual cells and determine the existence of burgeoning subpopulations over time (24,25). Multiple studies exist that also profile the transcriptional heterogeneity of untranslated regions (UTR) and transcription start sites (TSS) utilizing alternative deep sequencing technologies (26,27). Overlaying existing ribosome profiling (Ribo-seq) studies with the transcriptome data could expand our understanding of transcriptional dynamics (28). SGD will continue to integrate the aforementioned research in a systematic way and depict them informatively to help to gain insight into the *S. cerevisiae* transcriptome.

DATA AVAILABILITY

JBrowse is an open source genome browser available in the GitHub repository (https://github.com/GMOD/jbrowse). SGD software is open source and available from the GitHub



primary_transcript

Primary Data

Name	YAL012W_id100
Type	primary_transcript
Position	chrI:130735..132082 (+ strand)
Length	1,348 bp

Attributes

Id	YAL012W_id100
Seq_id	chrI
Source	SGD
Ypd	667

Region sequence

```
>chrI chrI:130735..132082 (+ strand)
class=primary_transcript length=1348
GAGGCCTATACACATAGACATTTGCACCTTATACATATACACACAAGACAAAACCAAAAA
AAATATGACTCTACAAGAATCTGATAAATTTGCTACCAAGGCCATTCATGCCGGTGAACA
TGTGGACGTTCACGGTTCCGTGATCGAACCCATTTCTTTGTCCACCACTTTCAAACAATC
TTCTCCAGCTAACCCTATCGGTACTTACGAATACTCCAGATCTCAAAATCCTAACAGAGA
GAACTTGGAAAGAGCAGTTGCCGCTTTAGAGAACGCTCAATACGGGTTGGCTTTCTCCTC
TGGTTCTGCCACCACCGCCACAATCTTGCAATCGCTTCCTCAGGGCTCCCATGCGGTCTC
TATCGGTGATGTGTACGGTGGTACCCACAGATACTTCACCAAAGTCGCCAACGCTCACGG
TGTGGAAACCTCCTTCACTAACGATTTGTTGAACGATCTACCTCAATTGATAAAGGAAAA
CACCAAATTGGTCTGGATCGAAACCCCAACCAACCCAACTTTGAAGGTCACCGACATCCA
```

Figure 3. Transcript isoform dialog popup. The ‘Ypd’ or ‘Gal’ attribute lists the raw abundance in the glucose- or galactose-media condition, respectively. ‘Region sequence’ displays the predicted sequence based on the S288C reference genome.

repository (https://github.com/yeastgenome). The TIF-seq data from Pelechano *et al.*, 2013, is accessible at NCBI GEO archive (accession GSE39128).

ACKNOWLEDGEMENTS

We would like to thank Wei Wu and Lars Steinmetz for their feedback on transcript annotation and making these data available. Thank you to Sage T. Hellerstedt and Micheal J. Alexander for help in preparing this manuscript.

FUNDING

United States National Institutes of Health, National Human Genome Research Institute [U41HG001315]; Gene Ontology Consortium (GOC) [U41HG002273]; Alliance of Genome Resources [U41HG02223]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health. The funders had no role in design, data processing, implementation, decision to publish or preparation of the manuscript. Funding for open access charge: the National Human Genome Research Institute at the United States National Institutes of Health [U41HG001315].

Conflict of interest statement. None declared.

REFERENCES

1. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R. *et al.* (2012) *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.
2. Engel,S.R., Dietrich,F.S., Fisk,D.G., Binkley,G., Balakrishnan,R., Costanzo,M.C., Dwight,S.S., Hitz,B.C., Karra,K., Nash,R.S. *et al.* (2013) The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 (Bethesda)*, **4**, 389–398.
3. The Gene Ontology Consortium (2019) The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
4. Miller,C., Schwalb,B., Maier,K., Schulz,D., Dümcke,S., Zacher,B., Mayer,A., Sydow,J., Marcinowski,L., Dölken,L. *et al.* (2011)

Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.*, **7**, 458.

5. Pizzinga,M. and Ashe,M.P. (2014) Yeast mRNA localization: protein asymmetry, organelle localization and response to stress. *Biochem. Soc. Trans.*, **42**, 1256–1260.
6. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
7. Gardner,J.M. and Jaspersen,S.L. (2014) Manipulating the yeast genome: deletion, mutation, and tagging by PCR. *Methods Mol. Biol.*, **1205**, 45–78.
8. Babarinde,I.A., Li,Y. and Hutchins,A.P. (2019) Computational methods for mapping, assembly and quantification for coding and Non-coding transcripts. *Comput. Struct. Biotechnol. J.*, **17**, 628–637.
9. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res.*, **41**, D991–D995.
10. Athar,A., Füllgrabe,A., George,N., Iqbal,H., Huerta,L., Ali,A., Snow,C., Fonseca,N.A., Petryszak,R., Papatheodorou,I. *et al.* (2019) ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.*, **47**, D711–D715.
11. Buels,R., Yao,E., Diesh,C.M., Hayes,R.D., Munoz-Torres,M., Helt,G., Goodstein,D.M., Elsik,C.G., Lewis,S.E., Stein,L. and Holmes,I.H. (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
12. Zhang,Z. and Dietrich,F.S. (2005) Mapping of transcription start sites in Saccharomyces cerevisiae using 5′ SAGE. *Nucleic Acids Res.*, **33**, 2838–2851.
13. Miura,F., Kawaguchi,N., Sese,J., Toyoda,A., Hattori,M., Morishita,S. and Ito,T. (2006) A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 17846–17851.
14. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
15. Xu,Z., Wei,W., Gagneur,J., Perocchi,F., Clauder-Münster,S., Camblong,J., Guffanti,E., Stutz,F., Huber,W. and Steinmetz,L.M. (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature*, **457**, 1033–1037.
16. Neil,H., Malabat,C., d'Aubenton-Carafa,Y., Xu,Z., Steinmetz,L.M. and Jacquier,A. (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, **457**, 1038–1042.
17. Yassour,M., Kaplan,T., Fraser,H.B., Levin,J.Z., Pfiffner,J., Adiconis,X., Schroth,G., Luo,S., Khrebtukova,I., Gnirke,A. *et al.* (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 3264–3269.
18. Yassour,M., Pfiffner,J., Levin,J.Z., Adiconis,X., Gnirke,A., Nusbaum,C., Thompson,D.A., Friedman,N. and Regev,A. (2010) Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. *Genome Biol.*, **11**, R87.
19. Ozsolak,F., Kapranov,P., Foissac,S., Kim,S.W., Fishilevich,E., Monaghan,A.P., John,B. and Milos,P.M. (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, **143**, 1018–1029.
20. Lardenois,A., Liu,Y., Walther,T., Chalmel,F., Evrard,B., Granovskaia,M., Chu,A., Davis,R.W., Steinmetz,L.M. and Primig,M. (2011) Execution of the meiotic noncoding RNA expression program and the onset of gametogenesis in yeast require the conserved exosome subunit Rrp6. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 1058–1063.
21. van Dijk,E.L., Chen,C.L., d'Aubenton-Carafa,Y., Gourvennec,S., Kwapisz,M., Roche,V., Bertrand,C., Silvain,M., Legoix-Né,P., Loeillet,S. *et al.* (2011) XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature*, **475**, 114–117.
22. Pelechano,V., Wei,W., Jakob,P. and Steinmetz,L.M. (2014) Genome-wide identification of transcript start and end sites by transcript isoform sequencing. *Nat. Protoc.*, **9**, 1740–1759.
23. Pelechano,V., Wei,W. and Steinmetz,L.M. (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, **497**, 127–131.
24. Gasch,A.P., Yu,F.B., Hose,J., Escalante,L.E., Place,M., Bacher,R., Kanbar,J., Ciobanu,D., Sandor,L., Grigoriev,I.V. *et al.* (2017) Single-cell RNA sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. *PLoS Biol.*, **15**, e2004050.
25. Nadal-Ribelles,M., Islam,S., Wei,W., Latorre,P., Nguyen,M., de Nadal,E., Posas,F. and Steinmetz,L.M. (2019) Sensitive high-throughput single-cell RNA-seq reveals within-clonal transcript correlations in yeast populations. *Nat. Microbiol.*, **4**, 683–692.
26. Kang,Y.N., Lai,D.P., Ooi,H.S., Shen,T.T., Kou,Y., Tian,J., Czajkowsky,D.M., Shao,Z. and Zhao,X. (2015) Genome-wide profiling of untranslated regions by paired-end ditag sequencing reveals unexpected transcriptome complexity in yeast. *Mol. Genet. Genomics*, **290**, 217–224.
27. McMillan,J., Lu,Z., Rodriguez,J.S., Ahn,T-H. and Lin,Z. (2019) YeasTSS: An Integrative Web Database of Yeast Transcription Start Sites.*Database (Oxford)*, doi:10.1093/database/baz048.
28. Michel,A.M., Kiniry,S.J., O'Connor,P.B.F., Mullan,J.P. and Baranov,P.V. (2018) GWIPS-viz: 2018 update. *Nucleic Acids Res.*, **46**, D823–D830.