DATABASE
The Journal of Biological Databases and Curation

*Database*, 2020, 1–6
doi: 10.1093/database/baaa008
Database update

Database update

# Incorporation of a unified protein abundance dataset into the *Saccharomyces* genome database

**Robert S. Nash, Shuai Weng, Kalpana Karra, Edith D. Wong, Stacia R. Engel, J. Michael Cherry* and the SGD Project**

Department of Genetics, Stanford University, 3165 Porter Drive, Palo Alto, CA 94304, USA

***Corresponding author**: Tel: +(650) 723-7541; Email: cherry@stanford.edu

Citation details: Nash,R.S., Weng,S., Karra,K. *et al.* Incorporation of a unified protein abundance dataset into the *Saccharomyces* genome database. *Database* (2020) Vol. 2020: article ID baaa008; doi:10.1093/database/baaa008

## Abstract

The identification and accurate quantitation of protein abundance has been a major objective of proteomics research. Abundance studies have the potential to provide users with data that can be used to gain a deeper understanding of protein function and regulation and can also help identify cellular pathways and modules that operate under various environmental stress conditions. One of the central missions of the *Saccharomyces* Genome Database (SGD; https://www.yeastgenome.org) is to work with researchers to identify and incorporate datasets of interest to the wider scientific community, thereby enabling hypothesis-driven research. A large number of studies have detailed efforts to generate proteome-wide abundance data, but deeper analyses of these data have been hampered by the inability to compare results between studies. Recently, a unified protein abundance dataset was generated through the evaluation of more than 20 abundance datasets, which were normalized and converted to common measurement units, in this case molecules per cell. We have incorporated these normalized protein abundance data and associated metadata into the SGD database, as well as the SGD YeastMine data warehouse, resulting in the addition of 56 487 values for untreated cells grown in either rich or defined media and 28 335 values for cells treated with environmental stressors. Abundance data for protein-coding genes are displayed in a sortable, filterable table on Protein pages, available through Locus Summary pages. A median abundance value was incorporated, and a median absolute deviation was calculated for each protein-coding gene and incorporated into SGD. These values are displayed in the Protein section of the Locus Summary page. The inclusion of these data has enhanced the quality and quantity of protein experimental information presented at SGD and provides opportunities for researchers to access and utilize the data to further their research.
Website URL: https://www.yeastgenome.org

## Introduction

The *Saccharomyces* Genome Database (SGD) collects, organizes and presents biological information about the genes and proteins of the budding yeast *Saccharomyces cerevisiae* (1–2). This information can be used to direct experimental research aimed at elucidating protein function and biological role in the context of the cell. Currently, Protein pages contain a descriptive overview of each predicted protein, experimental data such as protein abundance and half-life, structural domain information, primary amino acid sequence from a variety of strains with overlaid experimental post-translational modification (PTM) data, physicochemical properties derived from the protein sequence, a list of external identifiers and links to other resources that may be of use to researchers. As a community resource, one of the core missions of SGD is to interact with users in a variety of ways to assess their needs and future research directions. One aspect of this interaction involves working with researchers to incorporate datasets of interest to the wider scientific community.

Although many genes are controlled at the transcriptional level, others are controlled translationally or post-translationally, and yet others, including rate-limiting regulators, are controlled at multiple levels. As a result, one goal of proteomics research is to reliably measure and quantitate protein content under standard growth conditions. Doing so provides researchers with context regarding abundance levels relative to other proteins in the proteome and provides baseline information that can then be extended to answer questions regarding the regulation of protein levels when cells are grown under stress conditions.

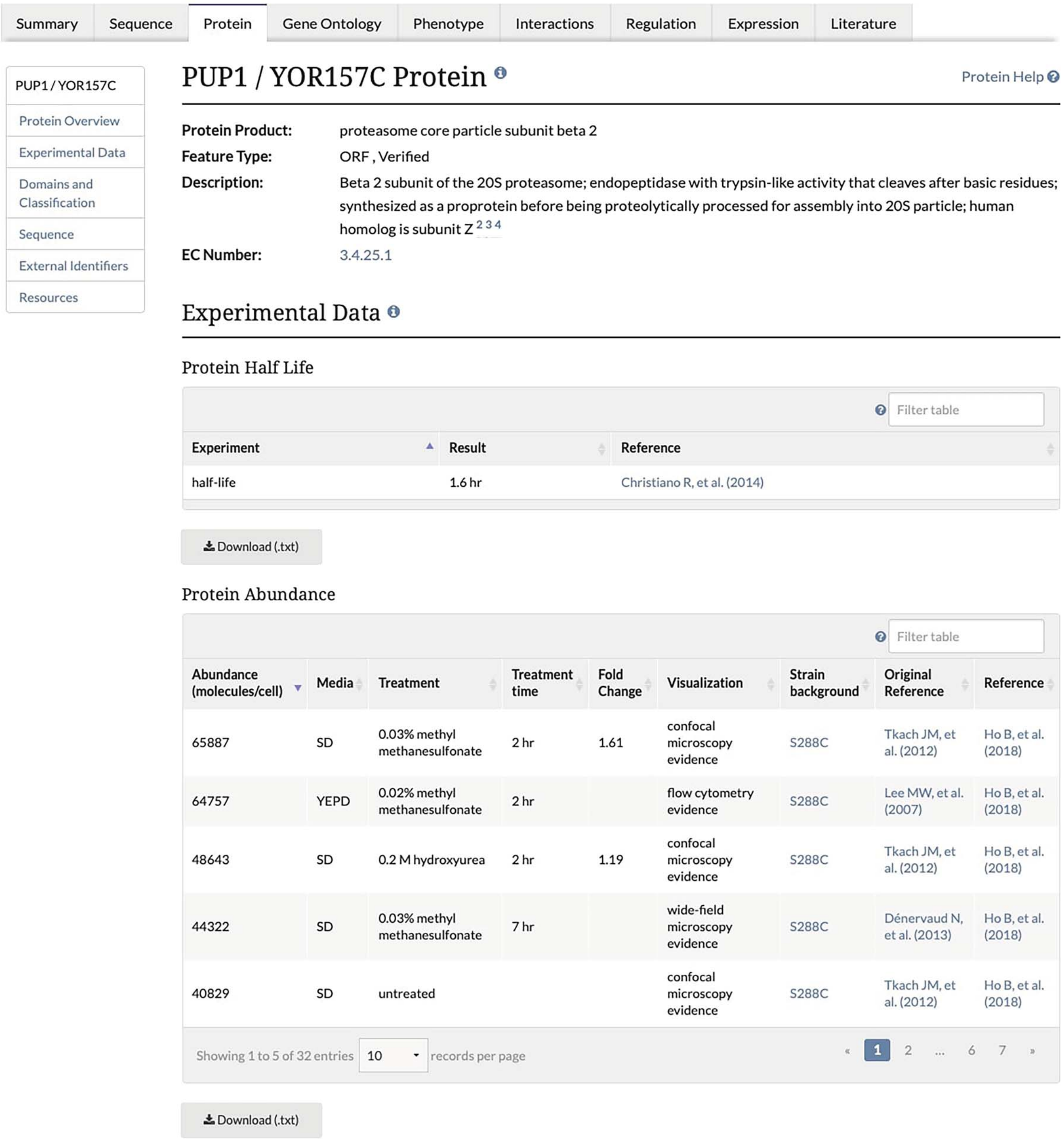
Recent advances in peptide labeling, sample preparation and both sensitivity and accuracy of mass spectrometry-based methods, coupled with advances in high-throughput imaging techniques, robotics and computational approaches to image analysis, have led to significant improvements in both the identification and quantification of proteins (3,4). These improvements have resulted in a proliferation of papers providing protein abundance datasets and provide an opportunity for the comprehensive analysis of protein abundance (5–25).

We became interested in updating our protein abundance data, with the goal of improving the quality and the quantity of experimental protein information available to our scientific community. As we surveyed the literature to collect protein abundance datasets, we concurrently learned that the laboratory of Grant Brown at the University of Toronto was collecting and evaluating protein abundance datasets with the goal of first normalizing the data presented in these studies and then converting it into the common units of molecules per cell (26). Since this dovetailed nicely with our desire to enhance experimental protein information presented on SGD Protein pages, we embarked on a collaborative effort to integrate this information and associated metadata into SGD. We also agreed that it would be advantageous to include median abundance values, and the Brown lab proposed calculating median absolute deviation (MAD). The median abundance provides a measure of the midpoint and makes it easier to compare the relative abundance of two or more proteins. The MAD provides a robust statistical measure of the variability within the abundance values. These values were added to Locus Summary pages for protein-coding genes and additionally integrated into the YeastMine data warehouse so that the abundance data and median values, even for large sets of genes, could be easily retrieved with templated queries (27).

## Data, metadata and ontologies

We focused our curation efforts on the unified dataset obtained from the Brown lab (26). This paper contains abundance data collected from the unified dataset published by the Brown lab in Ho *et al.* (26), where protein abundance data from 21 separate previously published proteomic studies were collected and analyzed. These previous studies had generated protein abundance values by any one of several independent methods, including mass spectrometry, GFP tagging coupled with either fluorescence microscopy or flow cytometry and tandem affinity purification coupled with immunoblot analysis. Since the unit space of the original data was either relative (abundance units) or absolute (molecules per cell), Ho *et al.* (26) used mode-shift normalization and scaling to convert all measurements of protein abundance from these publications into the intuitive units of molecules per cell. After filtering values to remove background autofluorescence from fluorescence microscopy-based studies, they removed low-abundance GFP-fusion protein signals, reducing coverage of the unified dataset from 97% to 92% of the proteome, which represented 5391 proteins and improved correlation with the calibration dataset (26). In addition to the baseline data obtained under standard growth conditions, a subset of GFP-based studies containing abundance data gathered from cells exposed to various environmental stressors were also analyzed (16–20, 23–24). For treated cells, abundance values were also normalized and unified. When the value in stressed cells was more than two standard deviations away from the untreated average abundance, a fold change was also calculated (26).

Metadata associated with the primary datasets used in Ho *et al.* (26) was reviewed and verified. This included the growth media, strain background, visualization method and, for treated cells, the treatment (including the concen-



PUP1 / YOR157C Protein

| | |
|---|---|
| **Protein Product:** | proteasome core particle subunit beta 2 |
| **Feature Type:** | ORF , Verified |
| **Description:** | Beta 2 subunit of the 20S proteasome; endopeptidase with trypsin-like activity that cleaves after basic residues; synthesized as a proprotein before being proteolytically processed for assembly into 20S particle; human homolog is subunit Z [2 3 4] |
| **EC Number:** | 3.4.25.1 |

Experimental Data

Protein Half Life

| Experiment | Result | Reference |
|---|---|---|
| half-life | 1.6 hr | Christiano R, et al. (2014) |

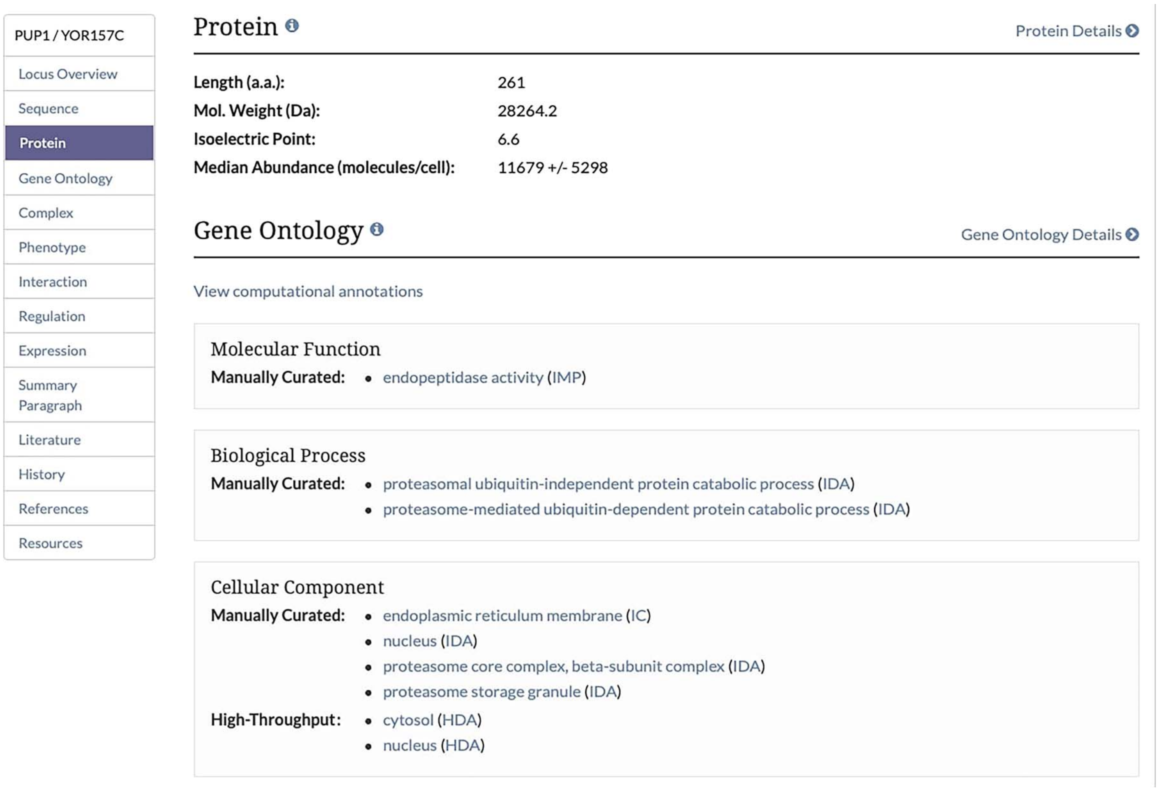
Protein Abundance

| Abundance (molecules/cell) | Media | Treatment | Treatment time | Fold Change | Visualization | Strain background | Original Reference | Reference |
|---|---|---|---|---|---|---|---|---|
| 65887 | SD | 0.03% methyl methanesulfonate | 2 hr | 1.61 | confocal microscopy evidence | S288C | Tkach JM, et al. (2012) | Ho B, et al. (2018) |
| 64757 | YEPD | 0.02% methyl methanesulfonate | 2 hr | | flow cytometry evidence | S288C | Lee MW, et al. (2007) | Ho B, et al. (2018) |
| 48643 | SD | 0.2 M hydroxyurea | 2 hr | 1.19 | confocal microscopy evidence | S288C | Tkach JM, et al. (2012) | Ho B, et al. (2018) |
| 44322 | SD | 0.03% methyl methanesulfonate | 7 hr | | wide-field microscopy evidence | S288C | Dénervaud N, et al. (2013) | Ho B, et al. (2018) |
| 40829 | SD | untreated | | | confocal microscopy evidence | S288C | Tkach JM, et al. (2012) | Ho B, et al. (2018) |

Showing 1 to 5 of 32 entries 10 records per page « 1 2 ... 6 7 »

**Figure 1.** Experimental Data Section of the Protein Page. This section of the Protein page contains two tables, one containing protein half-life data and the second containing the protein abundance data, and associated metadata, along with the original reference and the reference for the combined unified dataset. This table is both sortable and filterable.

tration of chemical applied to the cells, when applicable), units and treatment time (26). To standardize the metadata representation and enhance computational analysis, several different ontologies were investigated. We used the Experimental Factor Ontology (EFO; https://www.ebi.ac.uk/efo/), originally developed to describe experimental variables for expression studies, to represent yeast growth media (28). We used the Evidence & Conclusion Ontology (ECO; http://www.evidenceontology.org), a controlled vocabulary that describes scientific evidence, to describe the various visualization methods (29). When chemical treatments were used to induce environmental stress, terms from the Chemical Entities of Biological Interest (ChEBI; https://www.ebi.ac.uk/chebi/), an ontology used to classify chemicals based on both shared structural features and activities, were used (30). Experimental treatments that involved nitrogen starvation or cellular quiescence were represented by Gene Ontology process terms (GO; http://www.geneontology.org) (31). Finally, strain backgrounds were recorded to document the genetic environment in which abundance was measured (https://wiki.yeastgenome.org/index.php/Commonly_used_strains).



Protein

Length (a.a.): 261
Mol. Weight (Da): 28264.2
Isoelectric Point: 6.6
Median Abundance (molecules/cell): 11679 +/- 5298

Gene Ontology

Molecular Function
Manually Curated:
- endopeptidase activity (IMP)

Biological Process
Manually Curated:
- proteasomal ubiquitin-independent protein catabolic process (IDA)
- proteasome-mediated ubiquitin-dependent protein catabolic process (IDA)

Cellular Component
Manually Curated:
- endoplasmic reticulum membrane (IC)
- nucleus (IDA)
- proteasome core complex, beta-subunit complex (IDA)
- proteasome storage granule (IDA)

High-Throughput:
- cytosol (HDA)
- nucleus (HDA)

**Figure 2.** Protein Section of the Locus Summary Page. The protein section of the Locus Summary pages, located between the Sequence and Gene Ontology sections, contains the calculated median and MAD for the protein of interest expressed in molecules/cell in addition to basic sequence-derived information (length, molecular weight and isoelectric point). Median was calculated based on all values for a given protein from untreated cells, and MAD was calculated using the same values. When the median value was generated based on a single value, a MAD could not be calculated.

### Integration of protein abundance data into SGD and YeastMine

To store this novel unified protein abundance data in the SGD database, we created a new database table containing fields for recording the protein entity to which the specific abundance value is associated, an identifier to indicate the annotation source, a taxonomy ID indicating the strain background and two reference IDs, one for the original data source and a second for the data normalization and unification paper. In addition, there are fields for the data value, data unit, assay ID (ECO identifier) and media ID (EFO identifier) for the various growth media used. For cells treated with an environmental stress-inducing agent or condition, the table contains fields for chemical ID (CHEBI identifier), concentration value, concentration unit, time value and time unit. For cases in which the environmental stress was not a chemical treatment, this was captured using a Gene Ontology process term and is stored as a GO identifier. The fold change is also included in cases where the value in stressed cells is more than two standard deviations from the untreated average abundance. Finally, a median value was calculated from all values for a given protein from untreated cells and was used to calculate MAD using all values from untreated datasets and a constant of C = 1 (26). In cases where the median value was generated based on data from a single study, the MAD could not be calculated. Scaled protein abundance data for untreated and treated cells were loaded into the SGD database, which houses the data, metadata, original and unified data reference, median and newly calculated MAD. Abundance data, metadata, median abundance and MAD values were also added to our YeastMine data warehouse, using the data integrated into the SGD database as the source.

## Accessing protein abundance data at SGD

Unified protein abundance data stored in our database are displayed on our public website on the Protein page in the experimental data section for each visualized protein-coding gene (Figure 1). The table is located below a table containing experimentally determined proteome-wide protein half-life data. This table, consistent with others on the SGD website, can be both filtered and sorted. The data in each table can be retrieved using the 'Download' button located under the table (Figure 1). The median and MAD values are displayed in the Protein section of the Locus

Summary page, below sequence-derived values of protein length, molecular weight and isoelectric point (Figure 2).

Additionally, these data can be explored and downloaded from YeastMine (https://yeastmine.yeastgenome.org). Specifically, there are two templated pre-generated queries in the protein category; 'Gene to Protein Abundance', where abundance values for one or more proteins or a user customized list of proteins can be retrieved, and 'Gene to Median Protein Abundance', where median and MAD values for one or more proteins can be downloaded. These data are downloadable as tab- (.tsv) or comma-delimited text files (.csv), XML or JSON formats. Data is also downloadable using the YeastMine API (https://yeastmine.yeastgenome.org/yeastmine/api.do) or using SGD's web services (e.g. https://www.yeastgenome.org/webservice/locus/S000000364/protein_abundance_details).

## Future directions

We are currently investigating ways to better visualize the protein abundance data. To provide users with an overview of the abundance values and variance of protein(s) of interest, we are exploring the use of scatter plots to visualize the abundance value or median value for a given protein relative to all other proteins and to visualize the effect of treatment with stress on relative abundance. We will also need to explore how best to update these data if additional abundance datasets become available.

## Acknowledgements

We thank Grant W. Brown and Brandon Ho (Department of Biochemistry and Donnelly Center, University of Toronto, Toronto, Ontario, Canada) for initiating this project, providing datasets, accommodating our requests and providing feedback and advice during the integration of these data into SGD. Thanks are due to Micheal J. Alexander for help in preparing this manuscript.

## Funding

National Human Genome Research Institute at the United States National Institutes of Health [U41 HG001315]. Our efforts are also supported via the Gene Ontology Consortium (GOC) funded by the NHGRI [U41 HG002273], and the Alliance of Genome Resources, also funded by the NHGRI [U24 HG010859]. 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.

*Conflict of interest*. The authors declare that they have no conflict of interest.

## References

1. Cherry,J.M., Hong,E.L., Amundsen,C. *et al.* (2012) *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.
2. Hellerstedt,S.T., Nash,R.S., Weng,S. *et al.* (2017) Curated protein information in the *Saccharomyces* genome database. *Database*, **2017**, 1–6.
3. Mann,M., Kulak,N.A., Nagaraj,N. *et al.* (2013) The coming age of complete, accurate, and ubiquitous proteomes. *Mol. Cell*, **49**, 583–590.
4. Torres,N.P., Ho,B. and Brown,G.W. (2016) High-throughput fluorescence microscopic analysis of protein abundance and localization in budding yeast. *Crit. Rev. Biochem. Mol. Biol.*, **51**, 110–119.
5. Lu,P., Vogel,C., Wang,R. *et al.* (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.*, **25**, 117–124.
6. Peng,M., Taouatas,N., Cappadona,S. *et al.* (2012) Protease bias in absolute protein quantitation. *Nat. Methods.*, **9**, 524–525.
7. Kulak,N.A., Pichler,G., Paron,I. *et al.* (2014) Encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods.*, **11**, 319–324.
8. Lawless,C., Holman,S.W., Brownridge,P. *et al.* (2016) Direct and absolute quantification of over 1800 yeast proteins via selected reaction monitoring. *Mol. Cell Proteomics*, **15**, 1309–1322.
9. Lahtvee,P.J., Sánchez,B.J., Smialowska,A. *et al.* (2017) Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast. *Cell Syst.*, **4**, 495–504.e5.
10. de Godoy,L.M., Olsen,J.V., Cox,J. *et al.* (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, **455**, 1251–1254.
11. Picotti,P., Bodenmiller,B., Mueller,L.N. *et al.* (2009) Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell*, **138**, 795–806.
12. Lee,M.V., Topper,S.E., Hubler,S.L. *et al.* (2011) A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Mol. Syst. Biol.*, **7**, 514.
13. Thakur,S.S., Geiger,T., Chatterjee,B. *et al.* (2011) Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol. Cell Proteomics*, **10**, M110.003699.
14. Nagaraj,N., Kulak,N.A., Cox,J. *et al.* (2012) System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell Proteomics*, **11**, M111.013722.
15. Webb,K.J., Xu,T., Park,S.K. *et al.* (2013) Modified MuDPIT separation identified 4488 proteins in a system-wide analysis of quiescence in yeast. *J. Proteome Res.*, **12**, 2177–2184.
16. Tkach,J.M., Yimit,A., Lee,A.Y. *et al.* (2012) Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. *Nat. Cell Biol.*, **14**, 966–976.
17. Breker,M., Gymrek,M. and Schuldiner,M. (2013) A novel single-cell screening platform reveals proteome plasticity during yeast stress responses. *J. Cell Biol.*, **200**, 839–850.

18. Dénervaud,N., Becker,J., Delgado-Gonzalo,R. *et al.* (2013) A chemostat array enables the spatio-temporal analysis of the yeast proteome. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 15842–15847.
19. Mazumder,A., Pesudo,L.Q., McRee,S. *et al.* (2013) Genome-wide single-cell-level screen for protein abundance and localization changes in response to DNA damage in *S. cerevisiae*. *Nucleic Acids Res.*, **41**, 9310–9324.
20. Chong,Y.T., Koh,J.L., Friesen,H. *et al.* (2015) Yeast proteome dynamics from single cell imaging and automated analysis. *Cell*, **161**, 1413–1424.
21. Yofe,I., Weill,U., Meurer,M. *et al.* (2016) One library to make them all: streamlining the creation of yeast libraries via a SWAp-Tag strategy. *Nat. Methods*, **13**, 371–378.
22. Newman,J.R., Ghaemmaghami,S., Ihmels,J. *et al.* (2006) Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature*, **441**, 840–846.
23. Lee,M.W., Kim,B.J., Choi,H.K. *et al.* (2007) Global protein expression profiling of budding yeast in response to DNA damage. *Yeast*, **24**, 145–154.
24. Davidson,G.S., Joe,R.M., Roy,S. *et al.* (2011) The proteomics of quiescent and nonquiescent cell differentiation in yeast stationary-phase cultures. *Mol. Biol. Cell*, **22**, 988–998.
25. Ghaemmaghami,S., Huh,W.K., Bower,K. *et al.* (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.
26. Ho,B., Baryshnikova,A. and Brown,G.W. (2018) Unification of protein abundance datasets yields a quantitative *Saccharomyces cerevisiae* proteome. *Cell Syst.*, **6**, 192–205.e3.
27. Balakrishnan, R., Park, J., Karra, K. *et al.* (2012) Yeast mine–an integrated data warehouse for *Saccharomyces cerevisiae* data as a multipurpose tool-kit. *Database (Oxford)*, **2012**, bar 062.
28. Malone,J., Holloway,E., Adamusiak,T. *et al.* (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **2010**, 1112–1118.
29. Giglio,M., Tauber,R., Nadendla,S. *et al.* (2019) ECO, the evidence & conclusion ontology: community standard for evidence information. *Nucleic Acids Res*, **47**, D1186–D1194.
30. Hastings,J., Owen,G., Dekker,A. *et al.* (2016) ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res*, **44**, D1214–D1219.
31. The Gene Ontology Consortium (2019) The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.