YMDB 2.0: a significantly expanded version of the yeast metabolome database

Miguel Ramirez-Gaona[1], Ana Marcu[1], Allison Pon[1], An Chi Guo[1], Tanvir Sajed[1], Noah A. Wishart[2], Naama Karu[2], Yannick Djoumbou Feunang[2], David Arndt[1] and David S. Wishart[1,2,3,*]

[1]Departments of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada, [2]Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada and [3]National Institute for Nanotechnology, 11421 Saskatchewan Drive, Edmonton, AB T6G 2M9, Canada

ABSTRACT

YMDB or the Yeast Metabolome Database (http://www.ymdb.ca/) is a comprehensive database containing extensive information on the genome and metabolome of *Saccharomyces cerevisiae*. Initially released in 2012, the YMDB has gone through a significant expansion and a number of improvements over the past 4 years. This manuscript describes the most recent version of YMDB (YMDB 2.0). More specifically, it provides an updated description of the database that was previously described in the 2012 NAR Database Issue and it details many of the additions and improvements made to the YMDB over that time. Some of the most important changes include a 7-fold increase in the number of compounds in the database (from 2007 to 16 042), a 430-fold increase in the number of metabolic and signaling pathway diagrams (from 66 to 28 734), a 16-fold increase in the number of compounds linked to pathways (from 742 to 12 733), a 17-fold increase in the numbers of compounds with nuclear magnetic resonance or MS spectra (from 783 to 13 173) and an increase in both the number of data fields and the number of links to external databases. In addition to these database expansions, a number of improvements to YMDB's web interface and its data visualization tools have been made. These additions and improvements should greatly improve the ease, the speed and the quantity of data that can be extracted, searched or viewed within YMDB. Overall, we believe these improvements should not only improve the understanding of the metabolism of *S. cerevisiae*, but also allow more in-depth exploration of its extensive metabolic networks, signaling pathways and biochemistry.

INTRODUCTION

Saccharomyces cerevisiae, commonly known as baker's yeast, is a single-celled eukaryote used in biofuel production, winemaking, baking and brewing. In addition to being a cornerstone to several key industries that are collectively worth >$1 trillion/year, this humble microbe is also one of the principal model organisms used in genetics labs around the world. This is because *S. cerevisiae* can be easily grown, easily manipulated and many of its genes, proteins, cellular structures and signaling pathways resemble those found in the cells of higher eukaryotes. Hundreds of *S. cerevisiae* strains are known, of which more than 50 have been fully sequenced (1). The most commonly studied isolate of *S. cerevisiae* is S288C, a lab derived strain (2). The *S. cerevisiae* S288C genome consists of 16 chromosomes with 12 156 677 base pairs encoding for 5885 protein-coding genes, 275 tRNA genes, 140 genes specifying ribosomal RNA and 40 genes for small nuclear RNA molecules (3). Being one of the most intensively studied organisms on the planet, a number of online *S. cerevisiae* databases have appeared that archive experimentally measured data on the genetics, the biochemistry and the metabolism of this microbe. Some of the most important yeast databases include the Saccharomyces Genome Database or SGD (4), the MIPS Comprehensive Yeast Genome Database (CYGD) (5) and YeastCyc (6). More general pathway and metabolism databases such as Reactome (7), MetaCyc (6) and the Kyoto Encyclopedia of Genes and Genomes or KEGG (8) have also annotated detailed metabolic knowledge obtained from *S. cerevisiae*. Collectively, these resources have recently been used to construct the Yeast 6 consensus yeast metabolic network (9), which consists of 1888 reactions involving 1458 metabolites and 900 genes.

However, most of today's *S. cerevisiae* metabolism and pathway databases were designed specifically for systems biology or comparative metabolic network applications. Consequently they are focused on the annotation of water-

soluble primary metabolites and the description of primary metabolism in *S. cerevisiae*. With the rapid growth in metabolomics research around the world and the growing interest in identifying or understanding secondary metabolites, volatiles and lipid-soluble molecules, the need for a different kind of *S. cerevisiae* database has emerged. In particular, organism-specific metabolomic databases require much more detailed information on primary and secondary metabolites, referential nuclear magnetic resonance (NMR) and MS spectral data, metabolite concentration data, substrate growth conditions, very detailed pathway diagrams and many other features not commonly found in classical pathway databases or metabolic reconstructions (10). In an effort to address the shortcomings with existing databases and with a goal to create a resource more targeted to the needs of yeast metabolomics, we developed YMDB (11). The first version of YMDB was well received by the metabolomics community. However, there were still a number of shortcomings to the database that prevented it from being appealing to a broader user community. These included incomplete coverage of the yeast lipidome, limited viewing/searching tools, incomplete referential spectral data and relatively sparse pathway data.

Thanks to a number of recent developments in on-line rendering and viewing of MS and NMR spectra, of complex chemical structures and machine-readable pathways, we have been able to correct some of these previously mentioned shortcomings. Furthermore, we have been able to collect additional data on the chemistry and biochemistry of *S. cerevisiae* and have been able to incorporate a number of advanced web interface technologies to improve the both the content and the capabilities of YMDB. This manuscript provides a detailed description of the improvements made to the database as well as a point-by-point comparison between the original version of YMDB and the most recent version (YMDB 2.0).

Database updates and enhancements

Major improvements featured in YMDB 2.0 include: (i) significantly increased metabolite coverage, especially for the yeast lipidome; (ii) significant improvements to YMDB's pathways—including the number, the content and viewability; (iii) significantly improved spectral coverage including more experimental (and predicted) spectral data from more analytical platforms along with improved spectral searching and spectral viewing features; (iv) more detailed information on metabolites including their chemical classifications, physical properties, organoleptic features and more comprehensive descriptions; and (v) improvements to the database interface which enhance its ease of use, interactivity and search capabilities. Many of these updates are outlined in Table 1, where we compare the original version of YMDB with YMDB 2.0. Details for each major improvement are provided below.

Increased metabolite coverage

The original version of YMDB contained 2007 metabolites. While YMDB's coverage of water-soluble primary and secondary yeast metabolites was quite extensive, its coverage

of lipid molecules was very minimal, with just 672 lipids. With the growing interest in lipidomics (12) and the improvements in lipid detection technology (13) we decided to undertake a comprehensive update to YMDB's collection of lipids and lipid intermediates. The newly added lipid data in YMDB 2.0 was compiled from a combination of experimental data and known lipid synthesis pathways in yeast. In particular, using the known fatty acid composition reported in yeast (14), the known lipid species in yeast (14,15) and the known substrate processing capabilities of yeast lipid synthases (14,15), we generated all biochemically expected (feasible) lipid structures and deposited these into YMDB 2.0. This was the same process used in assembling the lipid collections published for ECMDB (10) and HMDB (16). This process led to the addition of more than 13 000 lipid molecules along with hundreds of new proteins and reactions, particularly those involving phospholipids and lysophospholipids. This also led to the addition of many metabolic intermediates and transport proteins for lipid biosynthesis. While only a portion of these yeast lipids have been identified/quantified experimentally so far, improvements in lipid detection technology will likely lead to most of these being formally identified and/or quantified in the coming years. To help users distinguish between already observed lipids and lipids predicted to exist via biochemical inference, we have added a filter function on the 'Browse Metabolites' page for users to select/filter between 'Observerd' and 'Expected' metabolites. In addition to expanding the degree of lipidome coverage, we also undertook an effort to substantially improve the quality of the lipid structure renderings. Previously, lipid structure diagrams had a 'squashed bug' appearance due to the random way that standard rendering programs (such as ChemAxon) generated lipid acyl or alkyl chains. As a result we developed Metbuilder, which is able to generate any lipid structure (using SMILES strings or IUPAC names as input) with the appropriate chirality and the degree of symmetry typically seen in textbook lipid diagrams.

In addition to our work on expanding YMDB's lipid coverage, our curation team also worked on scanning existing online *S. cerevisiae* metabolism databases for recent updates as well as reviewing recent literature on *S. cerevisiae* metabolism and *S. cerevisiae* metabolomic studies to identify additional metabolites, enzymes, proteins, transporters and reactions that were newly described or not already in YMDB. This also included the identification of metabolites and proteins involved in metabolite signaling. This work led to the addition of 421 non-lipid metabolites as well as concentration data for another 170 metabolites. Many of the metabolites identified from this literature review include volatile compounds associated with the production of wine and beer. Given the importance of volatiles in defining the flavor and aroma of many foods and drinks, we have also included a new data field in YMDB 2.0, which covers the organoleptic features of 446 yeast metabolites. No doubt, this expanded list of secondary metabolites will require to the identification and generation of many more pathways, reactions and enzyme components—some of which will likely be added to the next release of YMDB. Overall, this effort to expand the metabolite coverage in YMDB led to the addition of 14 035 metabolites and 30 708 reactions in-

Table 1. Comparison between YMDB and YMDB 2.0

Features	YMDB	YMDB 2.0
Number of metabolite entries	2007	16 042
Number of data fields	81	99
Number of external DB links	15	17
Number of synonyms	26 920	41 686
Number of pathways	66	9595
Number of reactions	916	31 624
Number of enzymes	857	919
Number of transporters	137	159
Compounds with concentration data	627	797
Number of experimental NMR spectra	930	1572
Number of compounds with Expt NMR	466	788
Number of experimental MS/MS spectra	1346	3154
Number of Compounds with Expt MS/MS	317	333
Number of experimental GC/MS spectra	208	1079
Number of Compounds with Expt GC/MS	208	448
Number of predicted MS/MS spectra	0	22 649
Number of predicted GC/MS spectra	0	1143

volving 155 more proteins. This represents an increase of 700, 3352 and 15% respectively over what was originally contained in the original version of YMDB.

The decision to include metabolite transport proteins/reactions as well as metabolites and proteins involved in signaling in YMDB 2.0 shows that there has been an important shift in focus within the metabolomics community. Traditionally, metabolites have been viewed as the bricks and mortar of cells, leading to a rather limited focus on their anabolic and catabolic reactions. However, many recent discoveries in metabolite uptake, cell–cell communication (quorum sensing), in cell signaling, in oncometabolism, in onco-immunology and in morphogenesis are pointing to the fact that metabolites play a vital role in defining cellular fate, function and control.

Pathway improvements

The original version of YMDB contained links to just 66 *Saccharomyces cerevisiae* KEGG pathways. This set of pathways provided coverage for just 37% of the 2007 compounds in YMDB 1.0. While the KEGG database is clearly an exceptional resource, it lacks key information about most aspects of lipid metabolism, metabolite localization, metabolite transport and metabolite signalling. Furthermore, KEGG pathways used in YMDB do not have reciprocal links to the metabolite data contained in YMDB. With the addition of some 13 000 lipids and lipid intermediates to YMDB 2.0 (see above), the need to create yeast-specific pathways became particularly acute. As a result we decided to correct these pathway problems through the creation of a large set of yeast-specific pathways using Path-Whiz (17). Pathwhiz is a web-enabled system that exploits recent advances in JavaScript technology to allow the creation of rich, interactive user interfaces that can be used to display complex detailed, colorful, interactive and machine-readable (SBML, BioPAX, SBGN-ML, PWML) pathway diagrams.

Using PathWhiz we generated a total of 9556 interactive *S. cerevisiae* metabolic and signaling pathways. This means that nearly every metabolite (80%) in YMDB 2.0 is now linked to a pathway and each element in the pathway is linked either to a compound in YMDB 2.0 or to a protein in UniProt (18). The pathways have been compiled with information from various sources, such as YeastCyc (6) and

other published literature. Each PathWhiz pathway shows detailed data about the cellular and subcellular location of the pathway components, each protein's quartenary structures and cofactors (if known), reaction processes guided by arrows, protein locations within the cell and the metabolite products or reactants. Every pathway has a meaningful title along with a detailed description of the pathway shown in the diagram. These yeast pathways also include transport proteins and transport reactions occurring between the membranes and multiple biological organelles of *S. cerevisiae,* which are not depicted in any other pathway database.

Each YMDB pathway is viewable through PathWhiz's interactive viewer and each pathway can be displayed as a colored image with explicit metabolite and protein structure depictions, a black and white version or a simplified KEGG-style wire-like pathway with proteins shown as boxes and metabolites as circles (see Figure 1). Metabolites in every pathway can be recolored to display concentration levels or metabolite locations, and each pathway can be zoomed in and out interactively (similar to a standard Google Maps Viewer). All of the PathWhiz pathways can be downloaded as PNG files, PDF files or even higher resolution SVG files for slide presentations or figures in posters and papers. Note that YMDB users are not expected to edit or annotate the PathWhiz pathways through the PathWhiz viewer. Instead, static PathWhiz pathway images are available by clicking on the thumbnail images found in YMDB. Users can use commercial image manipulation software to annotate these images if they wish.

By trying to generate as many complete and accurate pathway diagrams of *S. cerevisiae* covering as much of its metabolism as possible, we believe have created a much more solid foundation for understanding *S. cerevisiae's* biochemistry. In addition to the 9556 PathWhiz pathways, YMDB 2.0 still contains links to the KEGG pathways. As a result, the total number of pathway diagrams in YMDB 2.0 is 28 668 (see Table 1).

Enhanced spectral data

Metabolomics is critically dependent on the availability of reference NMR, MS/MS or GC-MS spectral data of pure compounds to facilitate compound identification or quantification. Many metabolomic databases such as Metlin
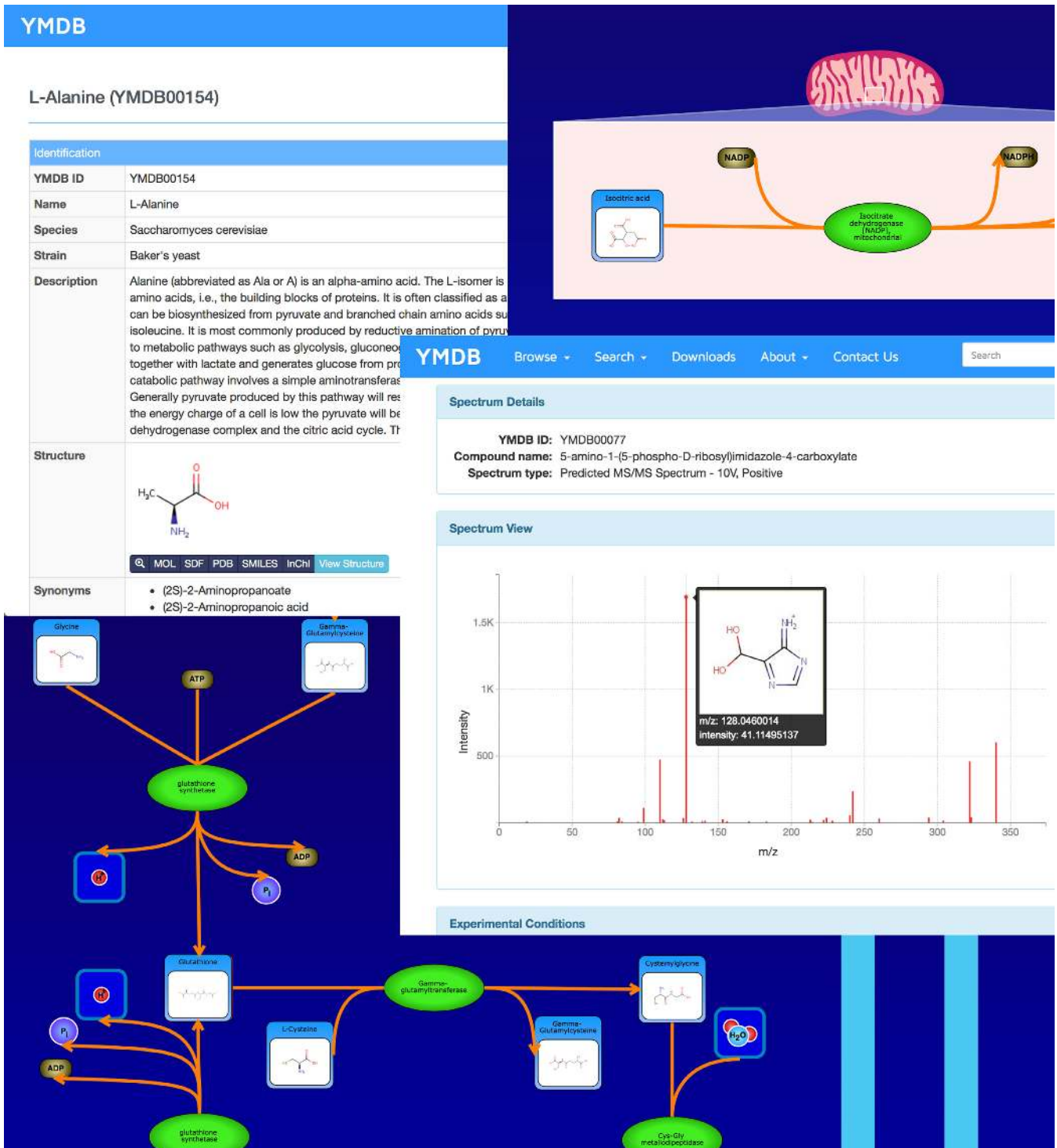
Figure 1. A screenshot montage of several different 'views' available in YMDB 2.0. These include the new pathway views (generated via PathWhiz), more detailed compound description pages and improved viewing tools for both NMR and MS/MS spectra with the capacity to interactively view predicted fragment ions by mousing over the MS/MS peaks.

(19), ECMDB (10) or HMDB (16) put considerable effort to maintaining reference spectral data that can be easily searched, queried or viewed. The first release of YMDB contained a total of 2484 experimentally acquired NMR spectra and MS/MS spectra from a total of 783 compounds. With the deployment of many new open-access spectral database resources such as the BioMagResBank (20) and MassBank (21) as well as significant improvements to the ability to predict MS/MS spectra with programs such as CFM-ID (22) and CFM-EI (23), we decided to take this opportunity to significantly update and upgrade the spectral data in YMDB. In particular, YMDB 2.0 now contains 2711 additional experimental MS, MS/MS and GC/MS spectra and 13 146 predicted (via software such as CFM-ID) MS/MS spectra as well as 1143 predicted (via CFM-EI) GC-MS spectra (Table 1). A total of 13 153 (82%) metabolites in YMDB 2.0 now have either experimental or predicted MS/MS spectra. In addition, 1166 (7%) metabolites in YMDB 2.0 have either experimental or predicted GC-MS spectra.

The expansion of YMDB 2.0′s spectral coverage was also done in conjunction with efforts to improve the quality and viewability of its reference spectral images. In the original version of YMDB all images of the experimentally acquired NMR spectra were static PNG files. YMDB 2.0 now displays all of its NMR spectra in an interactive JavaScript spectral viewer called JSpectraViewer. Using this on-line viewer all NMR spectra can be interactively zoomed, rescaled or manipulated with a simple click of a button. In addition, views for all of the MS/MS and EI-MS spectra in YMDB 2.0 have been recently regenerated so that they may be viewed as interactive 'stick' images (see Figure 1). Many of these updated MS/MS and EI-MS images are also annotated with fragment ion structures so that as the user mouses over the peaks, the predicted fragment ions appear.

Richer data content

The YMDB curation team has developed several in-house tools to automatically gather, annotate, predict and generate metabolites with their appropriate chemical structure. Using a locally developed program called DataWrangler the YMDB curation team validated, re-annotated and corrected every compound in YMDB 2.0. DataWrangler uses the compound name, CAS (Chemical Abstract Service) number or InChi identifier as a starting point to search through dozens of online chemical databases and resources. It also performs a variety of physical property predictions, annotations and 'sanity checks' on each query before compiling a summary of its findings, corrections, predictions and suspected inconsistencies. DataWrangler also captures synonyms, external links for other databases and chemical references thereby allowing for a more comprehensive compound annotation. Using the data generated by DataWrangler, members of the YMDB curation team then manually inspected the results and uploaded the appropriate corrections and/or updates. This work led to 14 766 more synonyms and 8839 more references than the original version of YMDB.

In addition to these data content upgrades we also used a software tool called ClassyFire (accessible at http://classyfire.wishartlab.com/) to perform a detailed taxonomic classification of all the compounds in YMDB 2.0. ClassyFire uses a structure-based classification algorithm combined with a detailed text description (an ontology) to facilitate and rationalize compound classifications and descriptions. In addition to the taxonomic classification by ClassyFire, all chemicals in YMDB 2.0 were also assigned (where possible) to taxonomic classes provided by ChEBI (24), KEGG (8,25), MetaCyc (6) or LIPID MAPS (26). The use of a standardized chemical taxonomy/ontology allows one to more readily compare, cluster or describe groups of metabolites. It also helps in compound identification and in a variety of machine learning applications in cheminformatics.

Improved user interface

A number of improvements to YMDB's user interface to enhance YMDB 2.0′s responsiveness, query speed and ease-of-use. For instance, we used a number of code changes to accelerate the NMR, GC-MS and MS/MS spectral search speeds by up to 500× over the original version. The use of JavaScript tools to display NMR spectra, MS/MS and GC-MS spectra as well as the PathWhiz pathway image data has also made for a much more colorful and interactive viewing experience. YMDB's text searches have been modified to support Elasticsearch thereby allowing 'approximate' text queries where typographical errors and mis-spellings are both tolerated and properly re-interpreted.

YMDB 2.0 also exploits of a number of recent improvements to web-based tools, frameworks and caching systems to make the website more user friendly and responsive. YMDB 2.0′s design is inspired by the Twitter Bootstrap framework. This makes for much easier navigation and a more appealing user experience. YMDB 2.0 also uses Redis-based caching that makes the loading of compounds, protein reactions and pathways very fast with wait times for page refreshes being reduced by a factor of four or more. Each YMDB compound view page now includes a jump-tab (the row of buttons on top of a compound annotation page) that allows users to quickly jump to specific data fields. In addition to these jump-tabs we have also added filters to the 'Browse Compounds' page. These filters allow for the rapid selection of a more specialized list of compounds according to the user's needs.

In addition to the many enhancements to the user interface, we also improved YMDB's data accessibility. Each compound annotation page has an 'XML' button that when clicked will display all the compound's annotations in XML format. Individual pages or the entire YMDB 2.0 database may be downloaded as an XML-formatted file. For those who prefer a simpler or less verbose data format, all of the data in YMDB 2.0 can also be viewed or downloaded in the more compact 'JSON' format. To further facilitate downloading, the entire YMDB database is now downloadable via 'HTTP Get' access. This allows users to easily download all of the YMDB 2.0 data including concentration information, chemical classification data, synonyms, spectral data and structural data.

CONCLUSION

YMDB 2.0 represents a substantial improvement to the original version of YMDB. The most significant enhancement has been the enormous increase ($7\times$) in the number of compounds in the database. This expansion led to a 435-fold increase in the number of metabolic pathway diagrams (from 66 to 28 734) now housed in the database. These pathways now cover almost all aspects of *S. cerevisiae* metabolism as well as many aspects of *S. cerevisiae*'s metabolite signalling and metabolite transport activities. Compared to other resources such as KEGG (with 72 *S. cerevisiae*-specific pathways) or YeastCyc (with 287 *S. cerevisiae*-specific pathways), YMDB 2.0 with its 9595 metabolic and signalling pathways now provides the most complete and comprehensive metabolic pathway collection for *S. cerevisiae*.

Many other important improvements were also implemented into YMDB 2.0. These include a 17-fold increase in the number of compounds linked to pathways (from 742 to 12 733), a significant increase in the quantity of NMR and MS spectra (from 2484 to 15 060) and a notable increase in the number of data fields and the number of types of external links to other *S. cerevisiae* or chemical resources. These additions, along with user interface enhancements allow for an improved speed in querying, searching and viewing data in YMDB 2.0. The improvements should also encourage a more in-depth exploration and understanding of *S. cerevisiae*'s extensive metabolic networks, its signaling pathways and its essential biochemistry.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Song,G., Balakrishnan,R., Binkley,G., Costanzo,M.C., Dalusag,K., Demeter,J., Engel,S., Hellerstedt,S.T., Karra,K., Hitz,B.C. *et al.* (2016) Integration of new alternative reference strain genome sequences into the saccharomyces genome database. *Database (Oxford)*, **2016**, baw074.
2. Mortimer,R.K. and Johnston,J.R. (1986) Genealogy of principal strains of the yeast genetic stock center. *Genetics*, **113**, 35–43.
3. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 563–567.
4. Sheppard,T.K., Hitz,B.C., Engel,S.R., Song,G., Balakrishnan,R., Binkley,G., Costanzo,M.C., Dalusag,K.S., Demeter,J., Hellerstedt,S.T. *et al.* (2016) The Saccharomyces Genome Database Variant Viewer. *Nucleic Acids Res.*, **44**, D698–D702.
5. Güldener,U., Münsterkötter,M., Kastenmüller,G., Strack,N., van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., García-Martínez,J., Pérez-Ortín,J.E. *et al.* (2005) CYGD: the comprehensive yeast genome database. *Nucleic Acids Res.*, **33**, D364–D368.
6. Caspi,R., Billington,R., Ferrer,L., Foerster,H., Fulcher,C.A., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **44**, D471–D480.
7. Fabregat,A., Sidiropoulos,K., Garapati,P., Gillespie,M., Hausmann,K., Haw,R., Jassal,B., Jupe,S., Korninger,F., McKay,S. *et al.* (2005) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.
8. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
9. Heavner,B.D., Smallbone,K., Price,N.D. and Walker,L.P. (2013) Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. *Database (Oxford)*, **2013**, bat059.
10. Sajed,T., Marcu,A., Ramirez,M., Pon,A., Guo,A.C., Knox,C., Wilson,M., Grant,J.R., Djoumbou,Y. and Wishart,D.S. (2016) ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli. *Nucleic Acids Res.*, **44**, D495–D501.
11. Jewison,T., Knox,C., Neveu,V., Djoumbou,Y., Guo,A.C., Lee,J., Liu,P., Mandal,R., Krishnamurthy,R., Sinelnikov,I. *et al.* (2012) YMDB: the yeast metabolome database. *Nucleic Acids Res.*, **40**, D815–D820.
12. Han,X. (2016) Lipidomics for studying metabolism. *Nat. Rev. Endocrinol.*, **12**, 668–679.
13. Sitepu,I.R., Ignatia,L., Franz,A.K., Wong,D.M., Faulina,S.A., Tsui,M., Kanti,A. and Boundy-Mills,K. (2012) An improved high-throughput Nile red fluorescence assay for estimating intracellular lipids in a variety of yeast species. *J. Microbiol. Methods*, **91**, 321–328.
14. Rattray,J.B, Schibeci,A. and Kidby,D.K. (1975) Lipids of yeasts. *Bacteriol. Rev.*, **39**, 197–231.
15. Baile,M.G., Lu,Y.W. and Claypool,S.M. (2014) The topology and regulation of cardiolipin biosynthesis and remodeling in yeast. *Chem. Phys. Lipids*, **179**, 25–31.
16. Wishart,D.S., Jewison,T., Guo,A.C., Wilson,M., Knox,C., Liu,Y., Djoumbou,Y., Mandal,R., Aziat,F., Dong,E. *et al.* (2013) HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res.*, **41**, D801–D807.
17. Pon,A., Jewison,T., Liang,Y., Knox,C., Maciejewski,A., Wilson,M. and Wishart,D.S. (2015) Pathways with PathWhiz. *Nucleic Acids Res.*, **43**, W552–W559.
18. UniProt Consortium (2007) The universal protein resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
19. Smith,C.A., O'Maille,G., Want,E.J., Qin,C., Trauger,S.A., Brandon,T.R., Custodio,D.E., Abagyan,R. and Siuzdak,G. (2005) METLIN: a metabolite mass spectral database. *Ther. Drug Monit.*, **27**, 747–751.
20. Ulrich,E.L., Akutsu,H., Doreleijers,J.F., Harano,Y., Ioannidis,Y.E., Lin,J., Livny,M., Mading,S., Maziuk,D., Miller,Z. *et al.* (2008) BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.
21. Horai,H., Arita,M., Kanaya,S., Nihei,Y., Ikeda,T., Suwa,K., Ojima,Y., Tanaka,K., Tanaka,S., Aoshima,K. *et al.* (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.*, **45**, 703–714.
22. Allen,F., Pon,A., Wilson,M., Greiner,R. and Wishart,D. (2014) CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res.*, **42**, W94–W99.
23. Allen,F., Pon,A., Greiner,R. and Wishart,D. (2016) Computational prediction of electron ionization mass spectra to assist in GC/MS compound Identification. *Anal. Chem.*, **88**, 7689–7697.
24. Degtyarenko,K., de Matos,P., Ennis,M., Hastings,J., Zbinden,M., McNaught,A., Alcántara,R., Darsow,M., Guedj,M. and Ashburner,M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.
25. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
26. Fahy,E., Subramaniam,S., Murphy,R.C., Nishijima,M., Raetz,C.R.H., Shimizu,T., Spener,F., van Meer,G., Wakelam,M.J.O. and Dennis,E.A. (2009) Update of the LIPID MAPS comprehensive classification system for lipids. *J. Lipid Res.*, **50**(Suppl), S9–S14.