The 2015 *Nucleic Acids Research* Database Issue and Molecular Biology Database Collection

Michael Y. Galperin¹,*, Daniel J. Rigden² and Xosé M. Fernández-Suárez³

¹National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, ²Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK and ³Thermo Fisher Scientific, Inchinnan Business Park, Paisley, Renfrew PA4 9RF, UK

Received November 10, 2014; Accepted November 11, 2014

ABSTRACT

The 2015 *Nucleic Acids Research* Database Issue contains 172 papers that include descriptions of 56 new molecular biology databases, and updates on 115 databases whose descriptions have been previously published in *NAR* or other journals. Following the classification that has been introduced last year in order to simplify navigation of the entire issue, these articles are divided into eight subject categories. This year’s highlights include RNA-central, an international community portal to various databases on noncoding RNA; ValidatorDB, a validation database for protein structures and their ligands; SASBDB, a primary repository for small-angle scattering data of various macromolecular complexes; MoonProt, a database of ‘moonlighting’ proteins, and two new databases of protein–protein and other macromolecular complexes, ComPPI and the Complex Portal. This issue also includes an unusually high number of cancer-related databases and other databases dedicated to genomic basics of disease and potential drugs and drug targets. The size of *NAR* online Molecular Biology Database Collection, [http://www.oxfordjournals.org/nar/database/a/](http://www.oxfordjournals.org/nar/database/a/), remained approximately the same, following the addition of 74 new resources and removal of 77 obsolete web sites. The entire Database Issue is freely available online on the *Nucleic Acids Research* web site ([http://nar.oxfordjournals.org/](http://nar.oxfordjournals.org/)).

NEW AND UPDATED DATABASES

The current *Nucleic Acids Research* (*NAR*) Database Issue is the 22nd annual collection of brief descriptions of bioinformatics databases, some of which are already well known, while others are described here for the first time. It includes 172 papers, of which 56 describe new databases (Table 1), 98 provide updates on the progress of the databases that have been previously described in the *NAR* Database Issue, and 17 contain updates on the status of the databases whose descriptions have previously been published elsewhere (Table 2).

To simplify navigation within the issue, we introduced last year the division of the entire Database Issue into eight sections: (i) nucleic acid sequence and structure, transcriptional regulation; (ii) protein sequence and structure, motifs and domains, protein-protein interactions; (iii) metabolic and signalling pathways, metabolites, enzymes, protein modification; (iv) viruses, bacteria, protozoa and fungi; (v) human genome, model organisms, comparative genomics; (vi) genomic variation, diseases and drugs; (vii) plant databases and (viii) other molecular biology databases. After mostly positive feedback, this year’s issue is again divided into the same sections. It must be noted, however, that many databases transcend the traditional borders between different areas of research and cannot be easily assigned to a single bin. In contrast, the *Nucleic Acids Research* online Molecular Biology Database Collection, [http://www.oxfordjournals.org/nar/database/a/](http://www.oxfordjournals.org/nar/database/a/), still retains the same 15 categories and 41 subcategories as before.

The most notable feature of this year’s issue is the increased number of databases that exploit RNA-seq data, using them for such diverse tasks as mapping transcription start sites (1), analysing gene co-expression data (2, 3), and cataloguing chimeric transcripts (4). Co-expression data are naturally focused principally on model organisms, but the move away from microarray-based data sets allows a more objective sampling of transcripts that encompass newly discovered genes. RNA-seq data, cheaply obtained, also offer a convenient way to obtain valuable comparative information on non-model organisms. The non-human primate reference transcriptome resource [NHPTRTR, (5)], for example, takes in relatives out to lemurs, enabling study of what is shared or distinct between us and our distant cousins. Similarly, DataBase of Apicomplexa Transcriptomes [DB-AT, formerly Full-Malaria, (6)] now includes RNA-Seq data for 14 species, shedding light on lesser known species with the context from the better studied apicomplexan parasites like *Plasmodium falciparum*.
Table 1. Descriptions of new online databases in the 2015 NAR Database Issue

| Database name | URL | Brief description |
|--------------|-----|-------------------|
| Addgene Vector Database* | http://www.addgene.org/vector-database/ | Plasmid vectors from publications and commercial sources |
| ADReCS | http://bioinf.xmu.edu.cn/ADReCS | Adverse Drug Reaction Classification System |
| AHTPdb | http://cedd.ossd.net/raghaa/ahtpdb/ | Antihypertensive peptides database |
| APASd | http://mosas.sysu.edu.cn/utr | Alternative poly(A) sites and downstream cleavage sites |
| BARCdbs | http://www.barcdbs.org | BioBanking Analysis Resource Database |
| BARD+ | https://bard.nih.gov/ | Breast Cancer Tissue Bank Bioinformatics portal |
| BCTTBmp | http://bioinformatics.brestcancertissuebank.org | Breast Cancer Tissue Bank Bioinformatics portal |
| Cancer3D | http://www.cancer3d.org | Mapping of cancer mutations to protein structures |
| CancerPDD | http://cedd.ossd.net/raghaa/cancerpdd/ | Experimentally validated anticancer peptides |
| Candidate Cancer Gene Database | http://cgcd-starfish.otf.umn.edu/ | Cancer genes identified in transposon-based genetic screens in mice |
| CeCaFDDB | http://www.cecaldb.org | Carbon flux data of central metabolism in various organisms |
| CFam | http://0137.132.71.120/cfam | Similarity-based classification of chemical compounds |
| CMPD | http://cgcbl.gwu.edu/tcmpd | Cancer Mutant Proteome Database |
| Coffee Genome Hub | http://coffee-genome.org/ | Coffee genomics, genetics and breeding data and tools |
| ComPPI | http://compgen.chem.uw.edu.pl/Protein | Protein-protein interactions linked to subcellular/localization |
| DBTMEEM | http://biocell.brc.duke.edu/ | Database of Transcriptome in Mouse Early Embryos |
| DDMGD | http://www.crbc.kaust.edu.sa/ddmgd/ | Associations between gene methylation and disease |
| Digital Ageing Atlas | http://ageing-map.org/ | Human age-related data |
| DoGSD | http://dogpd.big.ac.cn | Dog and wolf genome SNP database |
| EHPI | http://biotech.bmi.ac.cn/ehip/ | Essential Host Factors for Pathogenic Infection |
| EpilepsyGene | http://0122.228.158.106/EpilepsyGene | Genes and mutations related to epilepsy |
| eul.1dbd | http://eul.1dbd.unice.fr | L1 retrotransposon insertions in humans |
| GbShape | http://roshdb.emb.usc.edu/GbShape/ | Genome Browser for DNA shape annotations |
| i5k Workspace@NAL | http://i5k.nal.usda.gov/ | Arthropod genome projects |
| iBeetle-Base | http://beetle-base.uni-goettingen.de | RNAi screen phenotypes in the red flour beetle Tribolium castaneum |
| ImmunoCe | http://immunoces.ucsc.edu/ | Gene expression in immuno cells |
| KnotProt | http://biocomp.chem.uw.edu.pl/Protop | Proteins with topological knots and slipknots |
| LncRNA2Target | http://mgl.hit.edu.cn/lncrna2target | Differentially expressed genes after LncRNA knockdown or overexpression |
| lncRNAsNP | http://bioln.hit.edu.cn/lncRNAsNP | SNPs in human lncRNAs |
| LncRNAWiki | http://lncrna.big.ac.cn | Wiki database of human lncRNAs |
| MeT-DB | http://compgenomics.uta.edu/methylation/ | N6-methyladenine in human and mouse mRNA |
| MethBank | http://methdb.bsc.mctu.edu.tw | Nucleotide methylomes of gametes and early embryos |
| MethHC | http://methhc.mbc.nctu.edu.tw | DNA methylation in human cancer |
| MoonProt | http://www.moonlightingproteins.org/ | Moonlighting proteins |
| MyMpn | http://mycoplasma.org.cn/ | Mycoplasma pneumoniae as a model organism |
| NRChEd | http://proline.biochem.iseeernet.nri/chEd/ | Sequence databases enriched with computationally designed protein-like sequences |
| NutriChem | http://chc.dtu.dk/services/NutriChem-1.0/ | Nutritional and medicinal value of plant-based foods |
| Open TG-GATEs | http://home.nibb.go.jp/english/ | Gene expression and toxology data for rat liver and human and rat hepatocytes |
| Organ System Heterogeneity DB | http://mips.helmholtz-muenchen.de/organSystemHeterogeneity/ | Phenotypic effects of diseases and drugs on different organisms |
| Plantid-LGBase | http://lgbase.big.ac.cn/plantid-LGBase/ | Plantid-Linearly-based Conserved Gene-pair database |
| Platinum | http://structure.boc.csm.ac.uk/platinum | Experimentally measured effects of mutations on protein-ligand complexes |
| Protoscript | http://megaploton.org/aasp | Public angiosgenesis research portal |
| PyIgClassify | http://dunbrack2.fccc.edu/PyIgClassify/default.aspx | Clusters of conformations of antibody complementarity determining regions |
| RafProt | http://lipid-raft-database.du.eau.edu/ | Lipid raft-associated proteins |
| RiceVrMap | http://ricevmap.cgrt.org/ | Rice Variation Map, SNPs and indels |
| SABSDB | http://www.sabsdb.org/ | Small Angle Scattering Biological Data Bank |
| SNP-Seek | http://www.oryzasnp.org/snp-portal | Rice SNPs database |
| SuperFly | http://superfly.crg.eu | Spatio-temporal gene expression patterns in dipteran embryos |
| The Complex Portal | http://www.ebi.ac.uk/int/cmpact | Macromolecular complexes from key model organisms |
| tRFdb | http://genome.bioch.virginia.edu/trfdb/ | Short (14-32 nt) tRNA-related fragments |
| TrypanoCyc | http://www.metsystems.fr/trypanocyc/ | Biochemical pathways of Trypanosoma brucei |
| TSMTI | http://trmt hit-a-sta.tsi.brandenburg.de/ | Triplex Target DNA Sites in the human genome |
| VDE | http://bms-tokai.jp/vdE/ | VarySoDB Disease Edition: disease-associated genomic polymorphisms |
| ValidatorDB | http://uch.nmun.cn/ValidatorDB | Validation results for ligands and residues in the PDB |
| VRBase | http://www.rna-society.org/vracnadb/ | Virus-host interaction-associated ncRNAs |
| W3DB | http://w3db.phkz.ehu.es/dwp | W3Dp domain proteins and structure predictions |

*This database is part of this issue's Resource Collection (see text for details).

Table 2. Previously published databases that are new for the NAR Database Issue

| Database name | URL | Brief description |
|--------------|-----|-------------------|
| AraNet | http://www.aranet.org | Functional gene networks in Arabidopsis and other plants |
| ArrayMap | http://www.arraymap.com | Gene copy number profiling in human cancers |
| CareMap | http://www.caremap.com | Human Dephosphorylation database |
| CEPH | http://www.cephb.fr | A database of protein S-nitrosylation |
| CFam | http://superfly.crg.eu | A database of eukaryotic gene sequencing projects |
| CODINGDB | http://www.codingdb.cam.ac.uk | RNAseq based human gene and transcript co-expression map |
| DEGON | http://www.degond.org | Short (14-32 nt) tRNA-related fragments |
| DiSNNo | http://0140.138.144.14~/dSNNo/index.php | Short (14-32 nt) tRNA-related fragments |
| diArk | http://superfly.crg.eu | Spatio-temporal gene expression patterns in dipteran embryos |
| DNAVASH | http://vms-tokai.jp/vdE/ | VarySoDB Disease Edition: disease-associated genomic polymorphisms |
| GRASP | http://apps.shibb.nih.gov/grasp/ | RNAseq based human gene and transcript co-expression map |
| GeneFriends | http://www.genefriends.org | VarySoDB Disease Edition: disease-associated genomic polymorphisms |
| GenomeBase* | http://celulliauris.jp/ | RNAseq based human gene and transcript co-expression map |
| mRDB | http://msearch.org | Short (14-32 nt) tRNA-related fragments |
| MobIDB | http://mobidb.bio.unipd.it/ | Short (14-32 nt) tRNA-related fragments |
| PLoMGE* | http://plasmogen.sanger.ac.uk/ | Short (14-32 nt) tRNA-related fragments |
| PLAZA | http://bioinformatics.pub.agentur/plaza/ | Short (14-32 nt) tRNA-related fragments |
| ProteomeScout | https://proteomescout.wustl.edu | Short (14-32 nt) tRNA-related fragments |
| RNAcentral | http://rnacentral.org/ | Short (14-32 nt) tRNA-related fragments |
| sc-PDB | http://bioinfo-pharma.u-strasbg.fr/scPDB/ | Short (14-32 nt) tRNA-related fragments |

*This database is part of this issue's Resource Collection (see text for details).
Among the nucleic acid sequence databases, the major new entry is RNACentral (http://rnacentral.org), an international community portal to various databases on noncoding RNA (7). This new web site collects the data from (and provides links to) 10 major ncRNA databases, including ENA, RFam, RefSeq, tRNA Website, and LncRNAdb, updates on which are presented as separate papers in this issue. The growing interest in LncRNAs led to the inclusion, in addition to an update on LncRNAdb (8), of four other LncRNA-related databases: LncRNA2Target, LncRNASENP, LNCipedia, and LncRNAWiki. The last is structured as wiki, attempting to encourage community annotation, as advocated in the 2012 NAR editorial by Finn and colleagues (9).

The protein database section includes two thought-provoking papers describing the recent changes and a new vision of the UniProt and InterPro databases (10,11). The InterPro article is an interesting narrative of attempts to efficiently deal with data that are challenging not only in quantity but also in heterogeneity (deriving from carefully curated to fully automatic sources), as well as reporting on a valuable domain architecture-based search tool. They are accompanied by regular updates on protein domain databases, such as CDD, SMART and SUPERFAMILY (12–14), and protein family/orthology databases, such as Inparanoid, OMA, OrthoDB, COG and MBGD (15–19). While most of these databases have been featured in the NAR Database Issue in recent years, the COG database (18) receives its first update since 2003. The new version of COGs (whose description shares an author with this editorial) expands its coverage to 711 bacterial and archaeal genomes and provides improved annotations for more than 500 protein families (18). For two other databases in this issue, the tmRNA website and Islander (20,21), previous descriptions have been published in 2004.

Various aspects of the common goal of providing reliable and useful functional annotation are addressed by the updates of such well-known databases as GO, GOA, HAMAP, BRENDA, Rhea, Ensembl, FlyBase, XenBase, Mouse Genome Database, Rat Genome Database, neXtProt and the UCSC Genome Browser. Functional annotation is often complicated by ‘moonlighting’, the ability of certain proteins to perform two or more unrelated functions depending, for example, on their localization inside or outside the cell, or expression in specific tissues. We expect MoonProt (22), a database of such ‘moonlighting’ proteins, together with MultiTaskProtDB (23), published last year, to serve as guiding light in the analyses of this interesting phenomenon.

The protein structure category features an update on the status of the RCSB Protein Data Bank [PDB, (24)] and an update on a valuable collection of PDB-derived databases and structure-related tools (e.g., DSSP, HSSP, PDBReport, PDB_RedO) (25). Improvements to the CATH hierarchical database of structural domains include downloadable superpositions of superfamily representatives and improved parsing of ‘functional families’ (26). The Genome3D database also reports new developments (27) with sequences from 10 model organisms and from the Pfam database (28) now annotated with structural domain assignments from CATH, SCOP and several fold recognition programs. Four distinct methods now contribute structural models which can be visualized online or downloaded (27). These papers are accompanied by a major new entrant in this area, the ValidatorDB (http://ncbr.muni.cz/ValidatorDB) database, which provides results of a thorough validation of the properties of bound ligands and non-standard residues (e.g., phosphoserine), reporting on various problems in the analyzed structures, such as instances of missing atoms and incorrect chiralities (29).

Another important new arrival is SASBDB (http://www.sasbdb.org), a repository for small-angle scattering data of proteins, nucleic acids, and various macromolecular complexes, obtained using beams of X-rays or neutrons (30). Improvements in both the brilliance of the latest synchrotrons and in instrumentation have considerably reduced the sample and time requirements of such experiments. Scattering experiments can address biologically important targets such as macromolecular complexes and intrinsically disordered proteins that can be difficult or even plainly intractable for other structural methods; and can span a huge target size range. Given these factors, it is not surprising that scattering experiments have seen a surge in popularity justifying a new, bespoke and (currently) curated primary database.

The importance of post-translational modifications (PTMs) of proteins is reflected in continued efforts to capture the diversity of PTMs and their biological significance. This issue includes dbDNO, which specifically deals with protein S-nitrosylation, and ProteomeScout, which neatly adds context to raw PTM data with context from various databases (31,32). The important involvement of PTMs in modulating protein–protein interactions is addressed by the PTMcode database, which combines different analytic routes to describe and predict functional associations between PTM sites in either the same protein or two interacting proteins (33).

A significant fraction of new and updated databases reflect the efforts to use the genomic data to advance human health. These include updates on such well-established databases as Online Mendelian Inheritance in Man® (OMIM®), Catalog of Somatic Mutations in Cancer (COSMIC), and the UCSC Cancer Genomics Browser (34–36). Several databases focus on diverse aspects of cancers, from identification of candidate genes in mice [Mouse Tumor Biology Database and the Candidate Cancer Gene Database, (37,38)] via the impact of DNA methylation (MethHC) or copy number variation (ArrayMap) in cancer cells, to the consequences of mutations at the protein sequence or structural level (Cancer Mutant Proteome Database, Cancer3D) (39–42).

While many databases featured in this issue have had a long and successful presence on the web, we are glad to note the 25th anniversary of BRENDA and the international ImMunoGeneTics information system [IMGT®, (43)], the 15th anniversary of InterPro (11), and the 10th anniversary updates on BioModels and the Comparative Toxicogenomics Database (44,45). That said, the database with the most consistent presence in NAR is REBASE, a database of restriction-modification systems which has been featured in the first NAR Database Issue and described in NAR even before that. It now presents its 15th update (46), the first since 2010.
Finally, this year’s issue also includes a selection of papers that, in addition to virtual resources (i.e. online databases), provide descriptions of tangible resources for molecular biology, which are linked to these databases. These include Addgene Vector Database, Biobanking Analysis Resource Catalogue, BioAssay Research Database, and Breast Cancer Tissue Bank (see Table 1). Among the updates on previously described resources (Table 2), it is worth mentioning GenoBase, a description of the renowned Keio collection of Escherichia coli K-12 single-gene knockout mutants and associated bioinformatics resources (47); PlasmoGEM, a collection of DNA vectors and data for genetic manipulation of malaria parasites (48); Standard European Vector Architecture (SEVA), a database and a set of genetic tools for analysis and the engineering of Gram-negative bacteria for research or biotechnological purposes (49), and INFRAFRONTIER (formerly European Mouse Mutant Archive), which promotes the use of mouse models of disease by systematically phenotyping mouse mutants, as well as archiving and distributing mouse mutant lines (50).

**NAR ONLINE MOLECULAR BIOLOGY DATABASE COLLECTION**

This year’s update of the NAR online Molecular Biology Database Collection (which is freely available at [http://www.oxfordjournals.org/nar/database/a/](http://www.oxfordjournals.org/nar/database/a/)), involved inclusion of the 56 new databases (Table 1) and 15 databases not described previously in the NAR Database Issue (Table 2). In addition, the Collection has been expanded by including such databases as the European Variation Archive (http://www.ebi.ac.uk/eva/), cBioPortal for Cancer Genomics (http://www.cbioportal.org/public-portal/), and the ExAC Browser from the Exome Aggregation Consortium (http://exac.broadinstitute.org/). On the other hand, 76 discontinued databases have been removed from the Collection, which kept its size almost unchanged. After contacting their authors, 177 database entries have been updated by their authors with respect to new URLs, new descriptions, and/or other kinds of metadata.

We welcome suggestions for inclusion in the Collection of additional databases that have been published in other journals. Such suggestions should be addressed to XMFS at xose.m.fernandez@gmail.com and should include database summaries in plain text, organized in accordance with the [http://www.oxfordjournals.org/nar/database/summary/1](http://www.oxfordjournals.org/nar/database/summary/1) template.

**ACKNOWLEDGEMENTS**

We thank Dr Martine Bernardes-Silva and the Oxford University Press team led by Jennifer Boyd and Caomhie Ni Dhonáill for their help in compiling this issue.

**FUNDING**

The NIH Intramural Research Program at the National Library of Medicine [to M.Y.G.]. Funding for open access charge: Oxford University Press.

**Conflict of interest statement.** The authors’ opinions do not necessarily reflect the views of their respective institutions. X.M.F.S. is an employee of Thermo Fisher Scientific Inc.

**REFERENCES**

1. Suzuki, A., Wakaguri, H., Yamashita, R., Kawano, S., Tsuhirasha, K., Sugano, S., Suzuki, Y. and Nakai, K. (2015) DTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. *Nucleic Acids Res.*, 43, D87–D91.
2. Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T. and Kinoshita, K. (2015) COXPRESdb in 2015: coexpression database for animal species by DNA microarray and RNA-seq-based expression data with multiple quality assessment systems. *Nucleic Acids Res.*, 43, D82–D86.
3. van Dam, S., Craig, T. and de Magalhaes, J.P. (2015) GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Res.*, 43, D1124–D1132.
4. Frenkel-Morgenstern, M., Gorohovski, A., Vucenovic, D., Maestre, L. and Valencia, A. (2015) ChiTaRS 2.1 – an improved database of the Chimeric Transcripts and RNA-seq Data with novel sense–antisense chimeric RNA transcripts. *Nucleic Acids Res.*, 43, D6–D7.
5. Peng, X., Thierry-Mieg, J., Thierry-Mieg, D., Nishida, A., Pipes, L., Bozinoski, M., Thomas, M.J., Kelly, S., Weiss, J.M., Raveendran, M. et al. (2015) Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHRPR). *Nucleic Acids Res.*, 43, D73–D743.
6. Jakalski, M., Wakaguri, H., Kischka, T.G., Nishikawa, Y., Kawazu, S.-I., Matsubayashi, M., Kawahara, F., Tsuji, N., Cao, S., Sunaga, F. et al. (2015) DB-AT: a 2015 update to the full-parasites database brings a multitude of new transcriptomic data for apicomplexan parasites. *Nucleic Acids Res.*, 43, D631–D636.
7. The RNAcentral Consortium. (2015) RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Res.*, 43, D123–D129.
8. Quek, X.C., Thomson, D.W., Maag, J.L., Bartonick, N., Signal, B., Clark, M.B., Gloss, B.S. and Dinger, M.E. (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.*, 43, D168–D173.
9. Finn, R.D., Gardner, P.P. and Bateman, A. (2012) Making your database available through Wikipedia: the pros and cons. *Nucleic Acids Res.*, 40, D9–D12.
10. The UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, 43, D204–D212.
11. Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAmulla, C., McMenamin, C., Nuka, G., Pease, S. et al. (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, 43, D213–D221.
12. Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I. et al. (2015) CDD: NCBI’s Conserved Domains Database. *Nucleic Acids Res.*, 43, D222–D226.
13. Letunic, I., Doerks, T. and Bork, P. (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.*, 43, D257–D260.
14. Oates, M.E., Stahlhacke, J., Vavoulis, D.V., Smithers, B., Rackham, O.J.L., Sardar, A.J., Zaucha, J., Thurlby, N., Fang, H. and Gough, J. (2015) The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic Acids Res.*, 43, D227–D231.
15. Sonnhammer, E.L.L. and Oxtudl, G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, 43, D234–D239.
16. Altenhoff, A.M., Skunca, N., Glover, N., Train, C.-M., Suik, A., Pilzlo, i., Gori, K., Timiczeqk, B., Muller, S., Redestig, H. et al. (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view, and other improvements. *Nucleic Acids Res.*, 43, D240–D249.
17. Kriventseva, E.V., Tegendorf, F., Petry, T.J., Waterhouse, R.M., Simao, F.A., Poizneryak, I.A., Ioannidis, P. and Zdobnov, E.M. (2015) OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.*, 43, D250–D256.
18. Galperin, M.Y., Makarova, K.S., Wolf, Y.I. and Koonin, E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, 43, D261–D269.
19. Uchiyama, I., Mihaara, M., Nishi, H. and Chiba, H. (2015) MBGD update 2015: microbial genome database for flexible ortholog
