Published online 2 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D477–D485***
doi: 10.1093/nar/gkx1019

The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces

Adrian M. Altenhoff[1,2], Natasha M. Glover[1,3,4], Clément-Marie Train[1,3,4], Klara Kaleb[5], Alex Warwick Vesztrocy[1,5], David Dylus[1,3,4], Tarcisio M. de Farias[1,3,4], Karina Zile[1,5], Charles Stevenson[5], Jiao Long[6], Henning Redestig[6], Gaston H. Gonnet[1,2] and Christophe Dessimoz[1,3,4,5,7,*]

[1]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, [2]ETH Zurich, Computer Science, Universitätstrasse 6, 8092 Zurich, Switzerland, [3]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland, [4]Dept. of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland, [5]Dept. of Genetics, Evolution & Environment, University College London, Gower St, London WC1E 6BT, UK, [6]Bayer Crop Science NV, Technologiepark 38, 9052 Gent, Belgium and [7]Dept. of Computer Science, University College London, Gower St, London WC1E 6BT, UK

Received September 15, 2017; Revised October 16, 2017; Editorial Decision October 17, 2017; Accepted October 27, 2017

ABSTRACT

The Orthologous Matrix (OMA) is a leading resource to relate genes across many species from all of life. In this update paper, we review the recent algorithmic improvements in the OMA pipeline, describe increases in species coverage (particularly in plants and early-branching eukaryotes) and introduce several new features in the OMA web browser. Notable improvements include: (i) a scalable, interactive viewer for hierarchical orthologous groups; (ii) protein domain annotations and domain-based links between orthologous groups; (iii) functionality to retrieve phylogenetic marker genes for a subset of species of interest; (iv) a new synteny dot plot viewer; and (v) an overhaul of the programmatic access (REST API and semantic web), which will facilitate incorporation of OMA analyses in computational pipelines and integration with other bioinformatic resources. OMA can be freely accessed at https://omabrowser.org.

INTRODUCTION

Orthology, the formalization of the intuitive notion of 'corresponding genes in different species', is a cornerstone of genomics (reviewed in 1). Two genes are defined as orthologs if they diverged from a common ancestral gene through speciation (2). Orthologs can have conserved biological functions over long evolutionary ranges (e.g. 3) and are thus key to transferring knowledge of biological processes across species. Furthermore, orthologs are used as phylogenetic markers and as anchors to align chromosomes or genomes from different species. Because orthologs are so important, a large number of methods and resources for their inference have been developed over the years, such as the COGs database (4), Inparanoid (5), OrthoMCL (6), Ensembl Compara (7), KEGG Orthology (8), PhylomeDb (9), OrthoDB (10), EggNOG (11), MBGD (12), PLAZA (13) or OMA (14). An overview of general developments in orthology resources are provided in recent reports of the *Quest for Orthologs* consortium (15,16).

OMA ('Orthologous Matrix') distinguishes itself through high-quality orthology inferences, a broad coverage of all three domains of life, feature-rich web interface, availability of data in a wide range of formats and interfaces, and a frequent update schedule of two releases per year (14,17).

Here, we present key recent developments of OMA. We first review the improvements in species coverage and in the inference pipeline. Then, we review some of the major new functionalities, including a viewer for hierarchical orthologous groups, domain annotations, a dotplot synteny viewer and improved programmatic accesses. We conclude with a case study of OMA's use in the industry and with future perspectives.

SPECIES COVERAGE AND RELEASE SCHEDULE

We strive to release an updated OMA browser two times per year. Since our last update paper (14), there have been five new releases. The newest one covers ∼2100 species with over eleven million protein sequences from all three domains of

*To whom correspondence should be addressed. Tel: +41 21 692 4155; Fax: +41 21 692 54 55; Email: Christophe.Dessimoz@unil.ch

life (1617 Bacteria, 141 Archaea, 327 Eukaryota; Figure 1). Contrary to most other orthology resources, we also infer orthology across domain boundaries, which makes it possible to identify orthologs shared among e.g. bacteria, archaea, plants, fungi and animals.

In OMA, we update the genomes of the most important model organisms at every release (the 10 genomes with most experimentally backed gene ontology annotations). For other genomes, we only update them if they have been substantially re-annotated. New genomes are generally added to the browser based on user requests, our own needs or that of our collaborators. As a result, we focused our recent efforts on increasing the number of plants, early-branching eukaryotes, drosophila flies and ants. For example, we now cover three allopolyploid plant genomes (bread wheat, rapeseed and upland cotton) and provide homoeology predictions among them (18). OMA users can request new or updated genomes through a web-based form at https://omabrowser.org/suggest. Alternatively, they can still perform their own computations using the OMA standalone software, possibly reusing some of the genomes already analyzed in OMA through the all-against-all export function (14).

ALGORITHMIC IMPROVEMENTS

From the March 2017 release onward, the OMA Browser uses the updated 2.0 version of the OMA algorithm, which we recently described and benchmarked in a separate publication (19). This new algorithm improves both pairwise orthology and hierarchical orthologous group (HOG) inference. First, it is relatively common, following a gene duplication, for the two copies ('in-paralogs') to evolve at different rates. If the duplication occurred within one of two lineages of interest, this induces one-to-many orthologs between them. But because of the asymmetry in the evolutionary rate, one pair may appear to be significantly closer than the other, leading the original OMA algorithm (and other graph-based methods) to only infer the closer one as ortholog—thus missing the other pair. The new version attempts to address this issue by considering the evolutionary distances between in-paralogs, which results in a much higher recall.

Second, we also improved the scalability of HOG inference. We detail the definition and usefulness of HOGs in the next section, but for now it suffices to know that a HOG is a set of genes that have descended from a common ancestral gene in a clade of interest. There is a correspondence between HOGs, gene trees and pairwise orthologs (20). In OMA, we infer HOGs from the pairwise orthologs. The original algorithm, which worked in a 'top-down' fashion (from the root of the species tree to the leaves), was too slow to process very large gene families. In OMA 2.0, we introduced a 'bottom-up' variant of the algorithm which is several orders of magnitude faster with no negative impact on the performance (19).

IMPROVED SUPPORT OF HIERARCHICAL ORTHOLOGOUS GROUPS (HOGs)

When simultaneously considering many genomes across all of life, gene families can become huge. This results in complex evolutionary histories consisting of multiple nested evolutionary events. As a result, the traditional approach of considering pairwise relationships or gene trees becomes prohibitively complex to infer and to interpret.

To make sense of gene evolution in a more scalable framework, OMA adopts the concept of Hierarchical Orthologous Groups (HOGs). HOGs are sets of genes all descendant from a single common ancestral gene within a specific taxonomic range (Figure 2). For instance, the NADPH oxidase (NOX) family in vertebrates contains several paralogs which result from gene duplications, mostly ancestral to the vertebrates (21,22). Although their general sequence, structure, and function is relatively well conserved, the paralogous copies are associated with different diseases, indicating subtle but important differences among the copies (23). At the vertebrate taxonomic level, NOX1, NOX2 and NOX3 genes are clustered by OMA into distinct HOGs, consistent with the accepted notion that these were already distinct copies in the last common ancestor of the vertebrates. By contrast, at the Deuterostome taxonomic level, the three copies are clustered in the same HOG, indicating that they descended from a single ancestral gene in the last common ancestor of the Deuterostomes. Thus duplication of these genes is likely to have occurred in between the deuterostomes and vertebrate branches in the tree of life—perhaps as part of the 2R whole genome duplication at the basis of the vertebrates (24).

We now provide a HOG viewer in OMA, which takes advantage of the interactive and dynamic nature of modern web widgets. The viewer is composed of a familiar species tree, which lets the user select the taxonomic range of interest by clicking on the corresponding ancestral node, highlighted in red (Figure 2). Right of the tree, the viewer displays extant genes as squares, horizontally aligned with the species to which they belong. Crucially, genes are partitioned in HOGs according to the taxonomic level of reference, where HOG boundaries are denoted by vertical bars. It is possible to color the genes according to the corresponding protein lengths or GC content. Furthermore, it is also possible to remove HOGs that only contain a low proportion of genes across the taxonomic range of interest, because many of these are likely to be spurious. The viewer is implemented using the flexible TNT javascript framework (25).

We have also improved HOGs data structure retrieval for user-side analysis. HOG pages now feature dynamic tables with a domain architecture viewer. Individual HOG datums, such as the HOG structure in OrthoXML format (26), or a fasta file of the sequences for all the genes contained in that HOG, are now available for download directly from the OMA Browser (see also section below on programmatic access). In addition, we have recently developed a standalone python package ('pyham') which can be used to retrieve either single HOGs, or patterns of gene duplications and losses for multiple HOGs. Pyham can be installed by the standard Python package manager 'pip'.

DOMAIN ANNOTATIONS AND EXPLORATION

OMA now integrates domain annotations from Gene3D for individual protein entries (27). Currently, 78.3% of all entries in OMA have a domain annotation, resulting in an

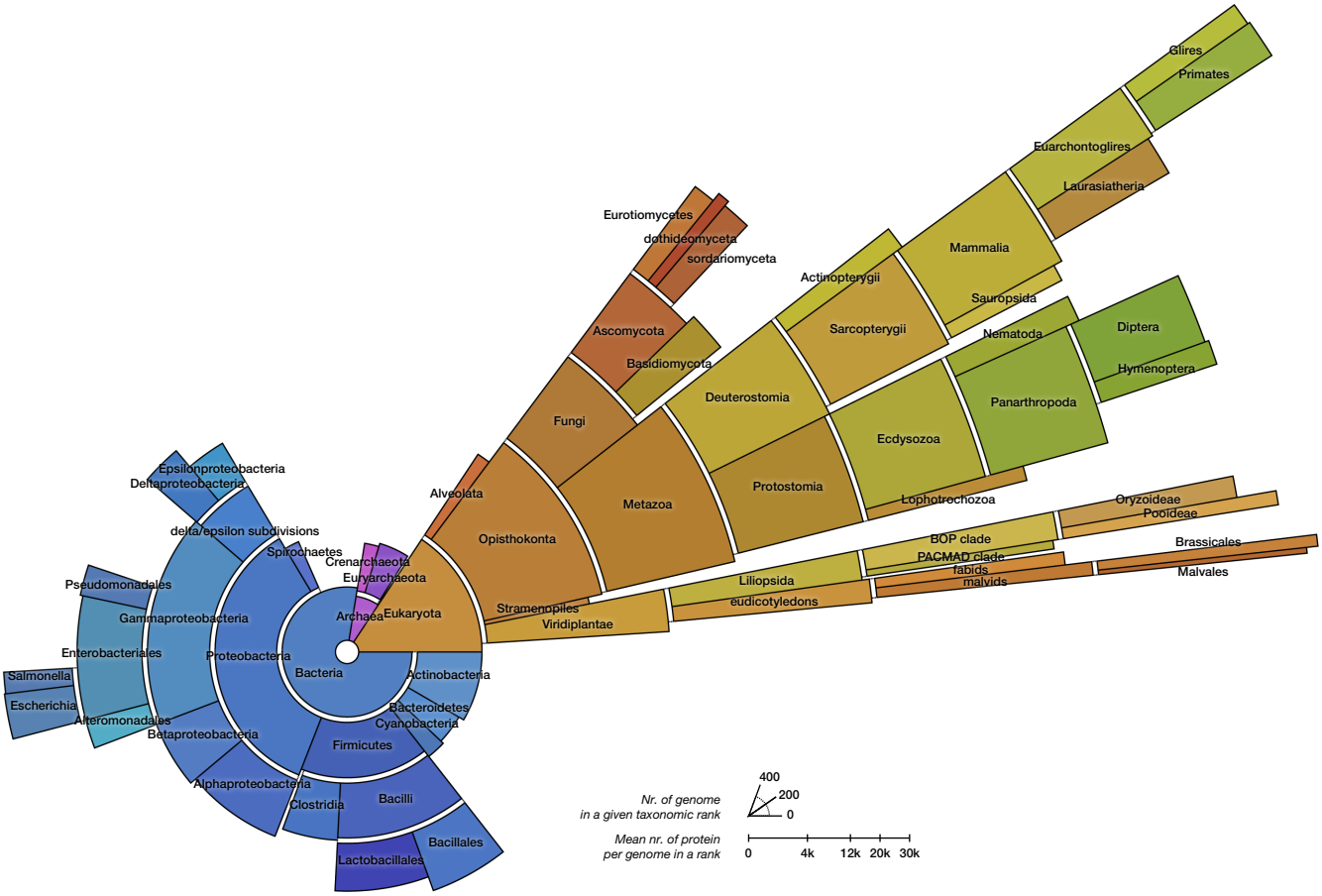

Figure 1. Distribution of the 2085 species contained in the October 2017 OMA release. The number of genomes in each taxonomic rank is conveyed as the angle of the relevant sector, and the average number of proteins is conveyed as its height in a square-root scale. Colors are automatically selected to contrast the different domains of life, and within them the different sister clades.

overall proportion of 55.1% amino-acid residues annotated as part of a domain. For each protein, the sequence of annotated domains is depicted using the conventional 'colored-boxes-on-a-line' representation, which we include in most protein lists. This makes it possible to easily check whether the domain architecture of a protein is conserved among orthologs, or to identify entries which are likely to be truncated or otherwise problematic. CATH domains (28) are depicted in colors specific to their first and second level classification. We assign the most prevalent domain architecture to the HOG itself.

Domains can also be used to establish links between HOGs. Given an initial HOG, a user can retrieve a table of the most similar HOGs based on conserved domain architecture. The similarity is computed by counting the number of domains in common between two HOGs. Genes that belong to distinct but similar HOGs can be paralogs separated by a very deep duplication, orthologs misclassified by OMA in separate groups or genes that are homologous for only part of their sequence (e.g. genes spanning over a domain fusion or fission event, artefactual fragments, etc.). This domain architecture view allows users to estimate how specific or widespread the domains that make up a protein family are, and allows them to make hypotheses about the origin of a protein family.

For example, Figure 3 depicts a ligase family specific to Bacteria (HOG:0564376) that could have originated from a fusion of a ubiquitous ligase family (HOG:0585097) with Carboxynorspermidine decarboxylase enzyme family (HOG:0580230). The domain-based search also identifies the Bacteria-specific family of UDP-N-acetylmuramyl-tripeptide synthetase (HOG:0560737), which is likely to have originated from a tandem duplication of a member of the ubiquitous ligase family.

PHYLOGENETIC MARKER GENE EXPORT

To infer a phylogenetic species tree, it is first necessary to identify sets of orthologous genes among the genomes of interest. One of the outputs of the OMA database are 'OMA Groups,' or sets of genes which are all orthologous to each other. Since genes in OMA Groups are related exclusively by speciation events, there is at most one sequence per species in each OMA group. In contrast to most other phylogenetic methods, OMA makes no assumption about species relationships when inferring OMA groups. This makes OMA Groups particularly useful for phylogenetic species tree inference.

The OMA groups are computed at each release over all species. Since many users are only interested in a small sub-

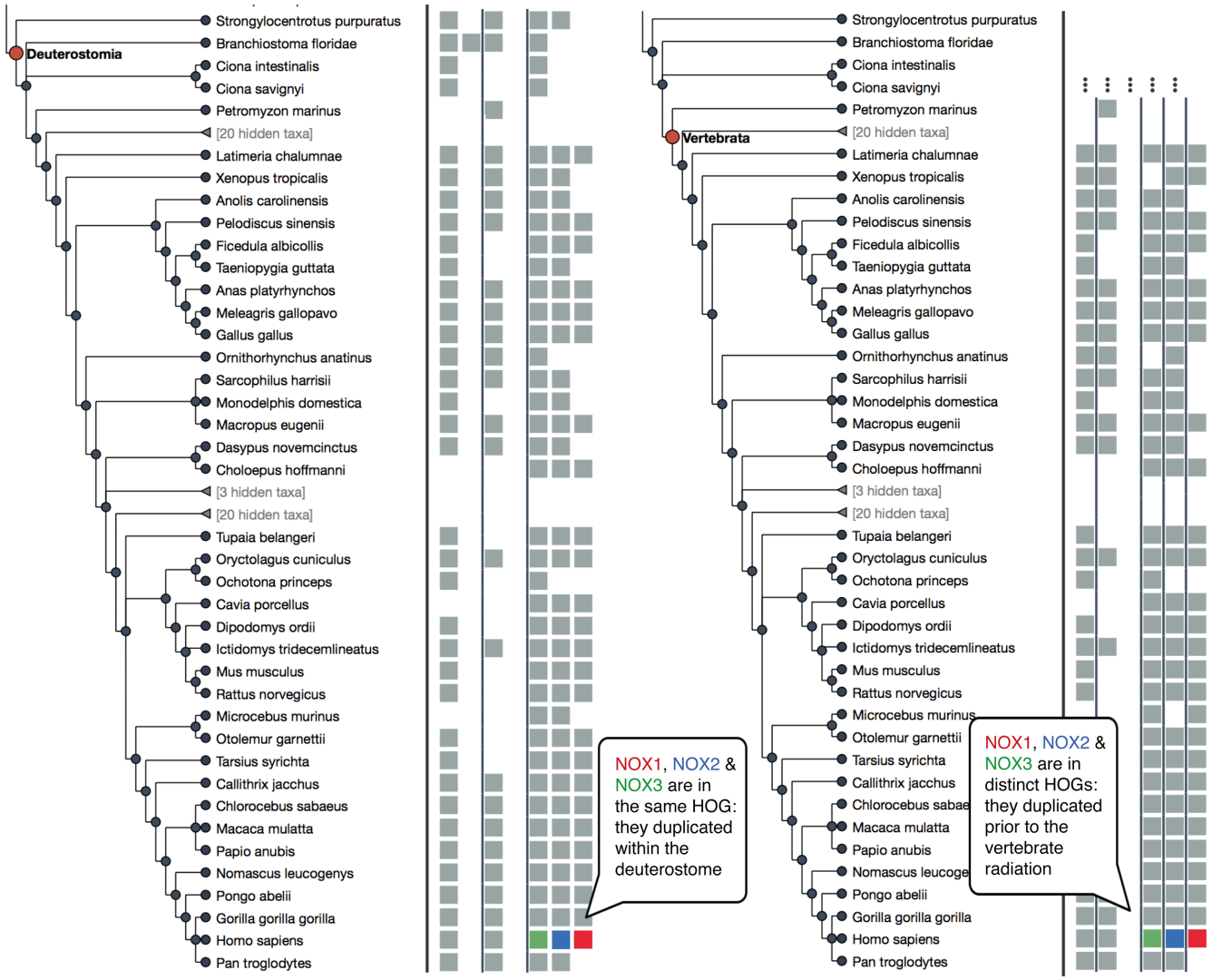

Figure 2. New interactive HOG viewer. An excerpt of the NOX family at the deuterostome level (left) and at the vertebrate level (right). The tree depicts relationships between species, squares depict genes (human NOX1, NOX2 and NOX3 genes are highlighted in color) and HOGs are delineated by vertical black lines.

set of genomes, we now provide a function to retrieve, for a given subset of species, the most complete OMA groups. The new functionality, entitled 'Export marker genes', is accessible under the 'Compute' menu. Users can optionally choose a minimum proportion of species present in each group ('occupancy'), and a maximum number of groups to export. From the choice of species and parameters, the OMA server identifies the most complete groups and produces a compressed archive file containing one fasta file per marker gene (i.e. per OMA group).

To illustrate this functionality, we exported marker genes for all 88 Fungi in the March 2017 release, requesting 100 markers with at least 50% occupancy. We independently aligned each group using Mafft (29), concatenated the resulting alignments without filtering (30) and inferred trees using FastTree (31)—using default parameters of each software tool. The entire procedure took 40 minutes on a single CPU, mostly spent aligning sequences. The resulting tree, highly resolved, is congruent with the NCBI taxonomy, with the sole exception of the placement of *Fomitopsis pinicola* (the disagreeing branch has however a lower support of 0.84; Supplementary Figure S1).

SYNTENY DOTPLOT

When comparing two related species, the position of orthologous genes is often conserved. Positional conservation can be at the chromosomal level—e.g. when there are entire chromosomes or chromosomal segments that are orthologous between species; or it can be more local—e.g. neighboring genes in one genome are orthologous to neighboring genes in the other genome. In OMA, we refer to global synteny for the former, and local synteny for the latter (local synteny is sometimes also referred to as 'colinearity').

The breakdown of synteny can be caused by gene movement via transposition/translocation, as well as large chromosomal or segmental rearrangements. Conservation of synteny, or lack thereof, can have several uses and impli-

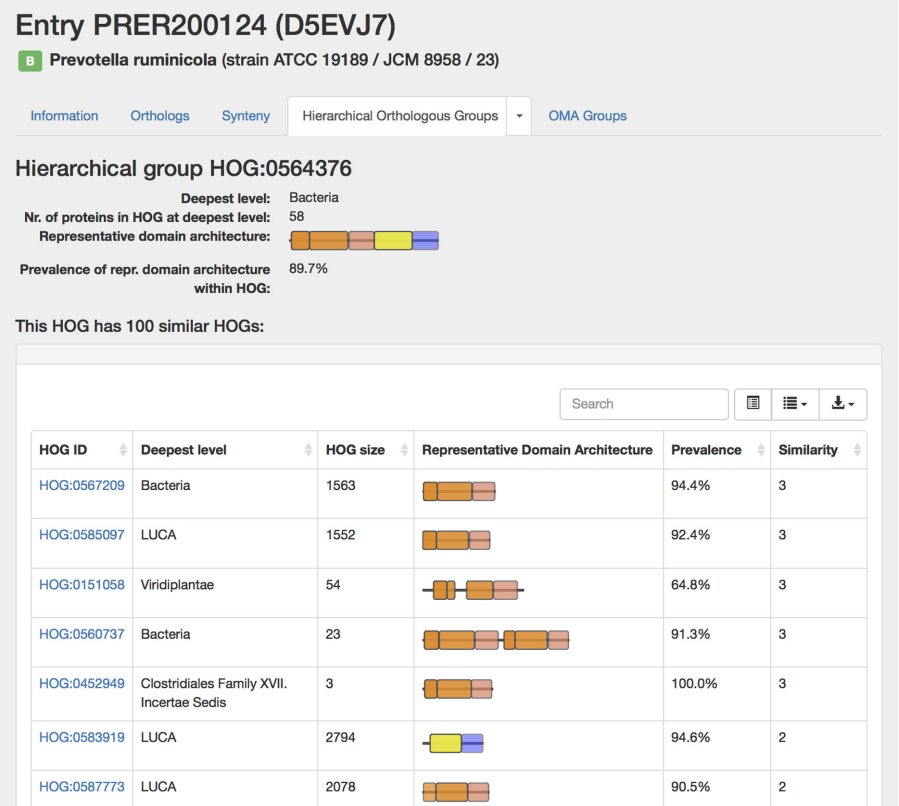

Figure 3. The domain architecture view of a HOG. Information about the HOG (on the top) is followed by the table containing information about other HOGs that share at least one domain in common with the HOG of interest. Deepest level: the last common ancestor of the species represented in a HOG; HOG size: the number of genes in a HOG; Representative Domain Architecture: the architecture that is characteristic of most of the proteins in a HOG; Prevalence: the percentage of the proteins in a HOG that have this domain architecture; Similarity: the number of the domains shared between this HOG and the HOG of interest (including duplicated domains). The table can be sorted by any of the attributes.

cations in evolutionary and comparative genomics: for example, synteny can be used to gauge how closely related genomes are, to identify genomic rearrangements, to reconstruct ancestral genomes and to aid genome assembly.

A few years ago, we introduced a local synteny viewer in OMA, which enables users to see orthology of neighboring genes across many species (14). This functionality has proven useful, particularly if we consider that many gene duplications are tandem duplication, and thus one-to-many and many-to-many orthology relationships can often be depicted even if one focuses on a narrow genomic window in each species. However, to identify larger events, such as large duplications and inversions, or to identify non-syntenic orthologs between an otherwise largely syntenic pair of genomes, a more global view is necessary.

Here, we introduce a synteny dotplot viewer in the OMA Browser. For any pair of chromosomes (in different species if we consider orthologs, or different subgenomes if we consider homoeologs), the plot draws orthologs as dots on a two-dimensional plot, where the axes are absolute physical location of the genes along the chromosome. Diagonals in the plot can thus be interpreted as syntenic regions, and one can easily identify genomic rearrangements such as inversions, duplications, insertions, deletions and highly repetitive regions (Figure 4). Users can zoom on particular regions of interest and obtain more details on orthologs of interest by selecting them. Each dot is colored based on a color scale reflecting the evolutionary distance in point accepted mutation (PAM) units. Furthermore, one can filter the orthologs to a specific distance range by clicking on the filtering icon and selecting the desired range on a histogram. Other features include panning and exporting the view as a high-resolution vector graphic. Thus, the new synteny dotplot complements the existing local synteny viewer by providing a more global and interactive view of positional conservation.

GO FUNCTION ANNOTATIONS

An important application of orthology is the ability to transfer gene function annotations from the few well-studied model organisms to the large number of poorly studied genomes. We previously described our approach to predict Gene Ontology (GO) annotations from OMA Groups (14). The approach was found to perform well in the Critical Assessment of Function Annotation 2 (CAFA2) experiment (32), where it scored highly under several criteria. Note however that large-scale benchmarking of functional prediction is notoriously difficult (33), so these results should be interpreted with caution.

In the same spirit as the mapping tool of the EggNOG database (34), we now provide a feature to annotate custom protein sequences through a fast approximate search with

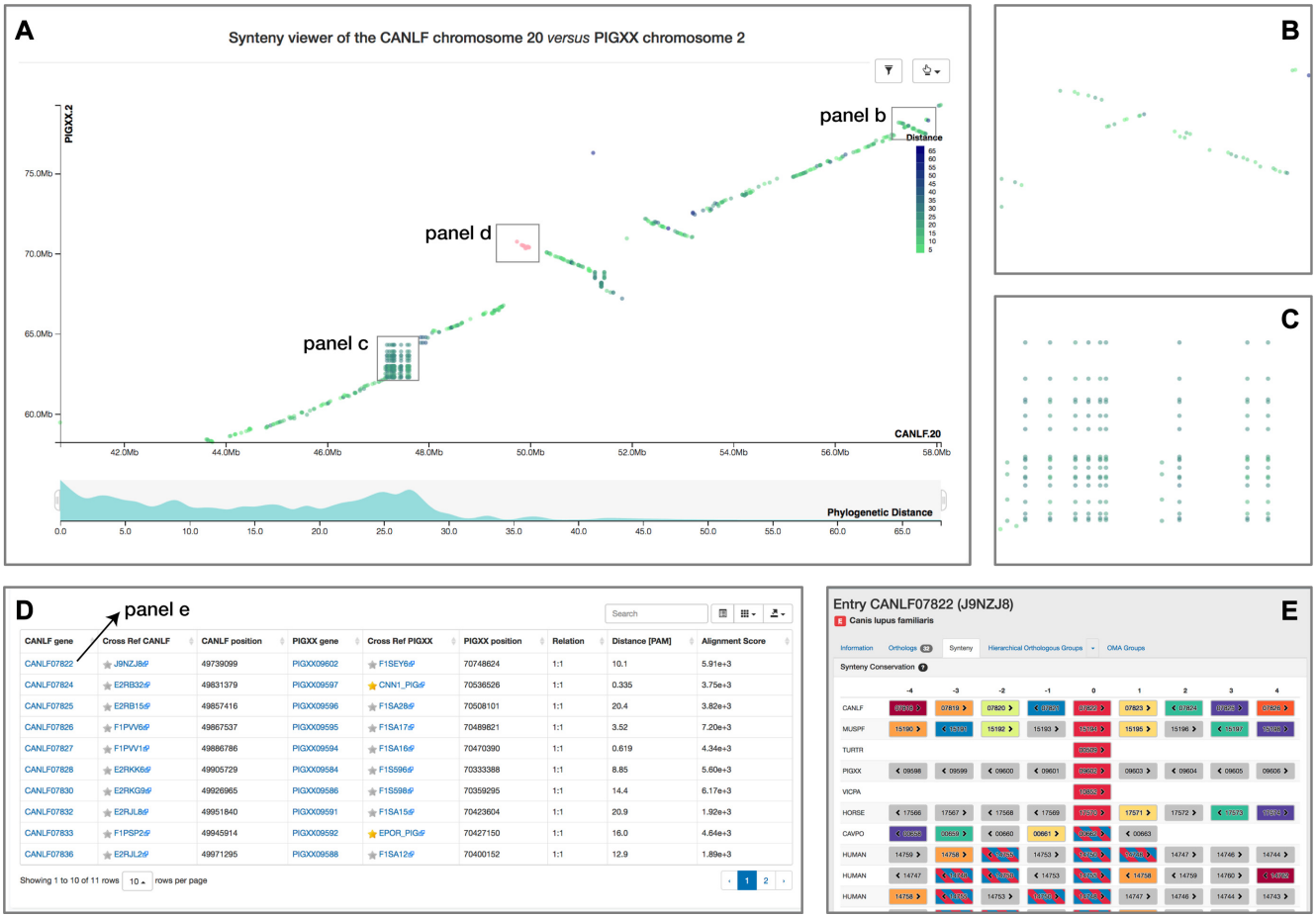

Figure 4. New dotplot synteny viewer, which enables users to identify gene order conservation between chromosomes as diagonal segments (main view in panel **A**). Inversions are visible as diagonal flips, which can be nested (panel **B**). Tandem duplications on one or the other chromosome are visible as vertical or horizontal lines—and, if both are present, as blocks (panel **C**). To focus on a subset of the data according to sequence divergence, the user can restrict the desired range of the distribution of the evolutionary distance of each point. Points can be selected by the user, in which case more details are provided in a table (panel **D**), including links to the local synteny viewer (panel **E**).

all the sequences in OMA. The user can upload a fasta formatted file and will receive the GO annotations (GAF 2.1 format) based on the closest sequence in OMA. These results can directly be further analyzed using other tools, e.g. to perform a gene enrichment analysis (reviewed in 35). This functionality is accessible under the 'Compute' menu in the OMA browser.

MODERN PROGRAMMATIC ACCESS: REST AND SPARQL

Allowing users to programmatically query the OMA data has been a goal early on: in 2007 we introduced Simple Object Access Protocol (SOAP) API and Distributed Annotation Service (DAS) endpoints. Since then, both technologies have however fallen out of favor by many users or developers. We are thus discontinuing support for SOAP and DAS, and replacing them with new Representational State Transfer (REST) and SPARQL Protocol and RDF Query Language (SPARQL) APIs.

The new REST API provides programmatic access to a comprehensive set of features provided through the web server. This API can be used to automate almost any analysis that a user could do on the website. On the REST API documentation page, which is accessible under https://omabrowser.org/api, all the endpoints and their parameters are described. Each endpoint includes also a live example. In addition, for R and python users, we provide native libraries wrapping around the REST API that further facilitates querying the OMA database in these languages.

Ontologies provide a way to describe and organize concepts used in biological databases, and thereby facilitate data interoperability across multiple resources. An Orthology Ontology (ORTH) was recently introduced (36), and we adapted and extended the ORTH ontology to fully support OMA. To enhance interoperability among resources, this updated ontology uses whenever possible terms compatible with other resources, such as the Microbial Genome Database (MBGD) (12) and Universal Protein Resource (UniProt) (37) ontologies. This version also describes additional orthology data such as OMA groups, domain architecture, nucleotide sequences and cross-references. Moreover, one of the major interoperability issues of orthology and life science databases is the heterogeneity of gene and protein identifiers used in these databases. To solve this is-

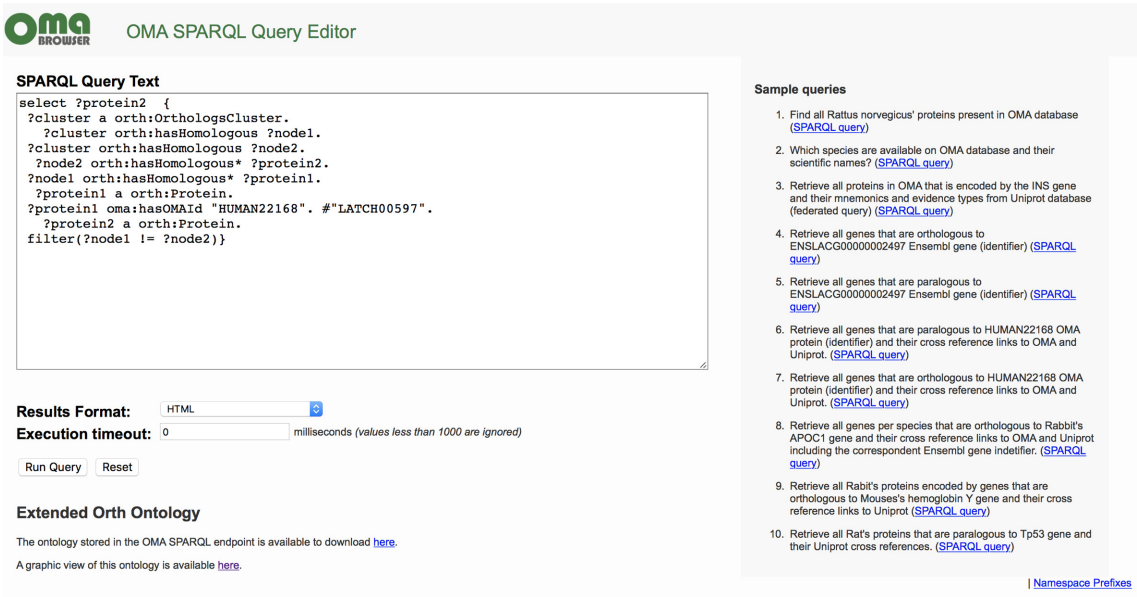

Figure 5. Example of a SPARQL query to programmatically retrieve pairwise orthologs involving the sequence LATCH00597. Sample queries are provided in the right column of the page, accessible at http://sparql.omabrowser.org.

sue, we extended the ORTH ontology by defining terms to explicitly represent multiple gene and protein identifiers such as the OWL property *identifier* and its sub-properties *ensemblGeneId*, *uniProtId*, *entrezGeneId* and *hasOMAId*. Therefore, these terms can be used by other data providers to avoid ambiguity among different identifiers. Furthermore, based on this extended version of the ORTH ontology, we released a SPARQL endpoint that is available on https://sparql.omabrowser.org to compose complex and federated queries over orthology and life science data (Figure 5).

OTHER NOTEWORTHY IMPROVEMENTS TO THE WEB INTERFACE

In addition to the above, we have implemented a number of smaller refinements that are worth mentioning here.

We now use dynamic tables for most lists in OMA. This enables users to sort according to the various table columns and to search rows using keywords. Responsiveness is also improved, with asynchronous loading of the table content and flexible pagination of the results. Finally, the new interface makes it easier to export the table contents in a variety of formats (e.g. JSON, XML, CSV, etc.).

The search function in OMA now supports autocompletion of identifiers and gene names. Whenever available, we use the gene name established by the HUGO gene nomenclature committee (38).

To display multiple sequence alignments, which we compute both for HOGs and OMA groups using Mafft (29), we now use the native web viewer MSAviewer (39).

We have also streamlined communication with users. OMA users can follow our latest updates on Twitter (@omabrowser), following the OMA blog (http://omabrowser.blogspot.com) or by signing up to our low frequency mailing list oma@lists.dessimoz.org. If they have questions, the preferred way to reach us is by asking questions on the BioStars Q&A platform (40) using the tag 'oma'.

The species selection in the all-against-all export functionality now uses the phylo.io tree viewer (41). All basic features of manipulation of a phylogenetic tree are included, such as label searching, re-rooting or branch swapping. Selected species are now automatically highlighted, making it easier to keep an overview on the tree of what is selected for export. Finally, once the final list of exported species is selected, phylo.io allows users to trim residual branches and display the tree of selected species only.

USE OF OMA IN THE INDUSTRY: THE EXAMPLE OF BAYER CROP SCIENCE

The access to accurate orthology relationships across all relevant species provides added value for applied research in industry applications, particularly at plant biotechnology companies. OMA collaborates with Bayer Crop Science (BCS) to accelerate the process of discovering and validating genes associated with crop traits related to yield potential, maintenance and tolerance to biotic and abiotic stresses by enabling the efficient mapping of gene functional information across model and crop species.

Through the five-year collaboration, we have deployed a private, scalable and extensible OMA instance combining proprietary and publicly available genomes from plant, insect, fungal and microbial species. Together with PLAZA (13), it constitutes the comparative genomics framework enabling BCS scientists to query orthologous pairs, to visualize the diversity in genomic content, to study the phylogenetic profiles of gene families of interest and to perform computational functional annotation based on orthology relationships.

The OMA@BCS resource is also updated twice a year, in line with the public OMA. The build code is merged to BCS code repositories on a regular basis and the publicly integrated data can be reused by BCS without repeating the computationally intensive all-against-all alignments thanks to the permissive licensing policy of the OMA project (Creative Common BY-SA 2.5 for the web browser, and open source MLP 2.0 license for code).

FUTURE PERSPECTIVES

This paper surveys a substantial number of improvements to the algorithm, coverage and interfaces of the OMA database. Just as importantly, OMA continues to be maintained and regularly updated.

As the cost of sequencing continues to drop, genomic data is shifting from consortium-led, general purpose, sequencing efforts to one-off user-generated data. OMA is adapting accordingly. We will continue to provide up-to-date, high-quality and user-friendly orthology relationships among many genomes across all of life in the public OMA database; in so doing, we will prioritize general-purpose, high-quality genomes, with a special effort toward better sampling life's diversity. At the same time, through web services operating on user-submitted data (e.g. the new function prediction tool introduced above), more flexible programmatic access, and OMA standalone, we aim to facilitate orthology analyses on custom data. And as our collaboration with Bayer demonstrates, it is already possible to deploy custom OMA Browser instances within organizations or individual laboratories interested in relating in-house data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Miguel Pignatelli (EMBL-European Bioinformatics Institute and Sanger Institute) and Matthieu Muffato (EMBL-European Bioinformatics Institute) for helpful discussions on the new hierarchical orthologous group viewer. We thank Ed Chalstrey and Jon Lees (University College London) for their help toward integrating domains in OMA. Finally, we thank all OMA users for making our efforts worthwhile (please keep sending us features and genome inclusion requests and bug reports). Computations were performed on the Computer Science cluster at University College London, the Vital-IT cluster at the University of Lausanne and the Euler cluster at ETH Zurich.

FUNDING

Swiss Institute of Bioinformatics, Service and Infrastructure grant (to G.H.G., C.D.); UK Biotechnology and Biological Sciences Research Council [BB/L018241/1 to C.D., BB/M009513/1 to K.Z.]; University College London, UCL Impact Award (to C.D.); Bayer Crop Science NV. Funding for open access charge: University College Library open access fund.

Conflict of interest statement. None declared.

REFERENCES

1. Gabaldón,T. and Koonin,E.V. (2013) Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.*, **14**, 360–366.
2. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
3. Kachroo,A.H., Laurent,J.M., Yellman,C.M., Meyer,A.G., Wilke,C.O. and Marcotte,E.M. (2015) Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science*, **348**, 921–925.
4. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
5. Sonnhammer,E.L.L. and Ostlund,G. (2014) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, **43**, D234–D239.
6. Chen,F., Mackey,A.J., Stoeckert,C.J. and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
7. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2008) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
8. Mao,X., Cai,T., Olyarchuk,J.G. and Wei,L. (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, **21**, 3787–3793.
9. Huerta-Cepas,J., Capella-Gutiérrez,S., Pryszcz,L.P., Marcet-Houben,M. and Gabaldón,T. (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.*, **42**, D897–D902.
10. Kriventseva,E.V., Tegenfeldt,F., Petty,T.J., Waterhouse,R.M., Simão,F.A., Pozdnyakov,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.
11. Huerta-Cepas,J., Szklarczyk,D., Forslund,K., Cook,H., Heller,D., Walter,M.C., Rattei,T., Mende,D.R., Sunagawa,S., Kuhn,M. *et al.* (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.*, **44**, D286–D293.
12. Uchiyama,I., Mihara,M., Nishide,H. and Chiba,H. (2012) MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. *Nucleic Acids Res.*, **41**, D631–D635.
13. Proost,S., Van Bel,M., Vaneechoutte,D., Van de Peer,Y., Inzé,D., Mueller-Roeber,B. and Vandepoele,K. (2015) PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Res.*, **43**, D974–D981.
14. Altenhoff,A.M., Škunca,N., Glover,N., Train,C.-M., Sueki,A., Piližota,I., Gori,K., Tomiczek,B., Müller,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.
15. Sonnhammer,E.L.L., Gabaldón,T., Sousa da Silva,A.W., Martin,M., Robinson-Rechavi,M., Boeckmann,B., Thomas,P.D., Dessimoz,C. and Quest,for Orthologs consortium (2014) Big data and other challenges in the quest for orthologs. *Bioinformatics*, **30**, 2993–2998.
16. Forslund,K., Pereira,C., Capella-Gutierrez,S., Sousa da Silva,A., Altenhoff,A., Huerta-Cepas,J., Muffato,M., Patricio,M., Vandepoele,K., Ebersberger,I. *et al.* (2017) Gearing up to handle the mosaic nature of life in the quest for orthologs. *Bioinformatics*, doi:10.1093/bioinformatics/btx542.
17. Altenhoff,A.M., Schneider,A., Gonnet,G.H. and Dessimoz,C. (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.*, **39**, D289–D294.
18. Glover,N.M., Redestig,H. and Dessimoz,C. (2016) Homoeologs: what are they and how do we infer them? *Trends Plant Sci.*, **21**, 609–621.
19. Train,C.-M., Glover,N.M., Gonnet,G.H., Altenhoff,A.M. and Dessimoz,C. (2017) Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. *Bioinformatics*, **33**, i75–i82.

20. Altenhoff,A.M., Gil,M., Gonnet,G.H. and Dessimoz,C. (2013) Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS One*, **8**, e53786.
21. Bedard,K. and Krause,K.-H. (2007) The NOX family of ROS-generating NADPH oxidases: physiology and pathophysiology. *Physiol. Rev.*, **87**, 245–313.
22. Boeckmann,B., Robinson-Rechavi,M., Xenarios,I. and Dessimoz,C. (2011) Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief. Bioinform.*, **12**, 423–435.
23. Katsuyama,M., Matsuno,K. and Yabe-Nishimura,C. (2012) Physiological roles of NOX/NADPH oxidase, the superoxide-generating enzyme. *J. Clin. Biochem. Nutr.*, **50**, 9–22.
24. Dehal,P. and Boore,J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.*, **3**, e314.
25. Pignatelli,M. (2016) TnT: a set of libraries for visualizing trees and track-based annotations for the web. *Bioinformatics*, **32**, 2524–2525.
26. Schmitt,T., Messina,D.N., Schreiber,F. and Sonnhammer,E.L.L. (2011) Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief. Bioinform.*, **12**, 485–488.
27. Lam,S.D., Dawson,N.L., Das,S., Sillitoe,I., Ashford,P., Lee,D., Lehtinen,S., Orengo,C.A. and Lees,J.G. (2016) Gene3D: expanding the utility of domain assignments. *Nucleic Acids Res.*, **44**, D404–D409.
28. Sillitoe,I., Lewis,T.E., Cuff,A., Das,S., Ashford,P., Dawson,N.L., Furnham,N., Laskowski,R.A., Lee,D., Lees,J.G. *et al.* (2015) CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.*, **43**,D376–D381.
29. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
30. Tan,G., Muffato,M., Ledergerber,C., Herrero,J., Goldman,N., Gil,M. and Dessimoz,C. (2015) Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Syst. Biol.*, **64**, 778–791.
31. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
32. Jiang,Y., Oron,T.R., Clark,W.T., Bankapur,A.R., D'Andrea,D., Lepore,R., Funk,C.S., Kahanda,I., Verspoor,K.M., Ben-Hur,A. *et al.* (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.*, **17**, 184.
33. Dessimoz,C., Skunca,N. and Thomas,P.D. (2013) CAFA and the open world of protein function predictions. *Trends Genet.*, **29**, 609–610.
34. Huerta-Cepas,J., Forslund,K., Pedro Coelho,L., Szklarczyk,D., Juhl Jensen,L., von Mering,C. and Bork,P. (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.*, **34**, 2115–2122.
35. Bauer,S. (2017) Gene-category analysis. *Methods Mol. Biol.*, **1446**, 175–188.
36. Fernández-Breis,J.T., Chiba,H., Legaz-García,M.D.C. and Uchiyama,I. (2016) The Orthology Ontology: development and applications. *J. Biomed. Semantics*, **7**, 34.
37. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
38. Yates,B., Braschi,B., Gray,K.A., Seal,R.L., Tweedie,S. and Bruford,E.A. (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625.
39. Yachdav,G., Wilzbach,S., Rauscher,B., Sheridan,R., Sillitoe,I., Procter,J., Lewis,S.E., Rost,B. and Goldberg,T. (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, **32**, 3501–3503.
40. Parnell,L.D., Lindenbaum,P., Shameer,K., Dall'Olio,G.M., Swan,D.C., Jensen,L.J., Cockell,S.J., Pedersen,B.S., Mangan,M.E., Miller,C.A. *et al.* (2011) BioStar: an online question & answer resource for the bioinformatics community. *PLoS Comput. Biol.*, **7**, e1002216.
41. Robinson,O., Dylus,D. and Dessimoz,C. (2016) Phylo.io: interactive viewing and comparison of large phylogenetic trees on the web. *Mol. Biol. Evol.*, **33**, 2163–2166.