OrthoMaM v12: a database of curated single-copy ortholog alignments and trees to study mammalian evolutionary genomics

Rémi Allio, Frédéric Delsuc, Khalid Belkhir, Emmanuel J P Douzery, Vincent Ranwez, Céline Scornavacca

Nucleic Acids Research, 2023, 1–7
https://doi.org/10.1093/nar/gkad834
Database issue

OrthoMaM v12: a database of curated single-copy ortholog alignments and trees to study mammalian evolutionary genomics

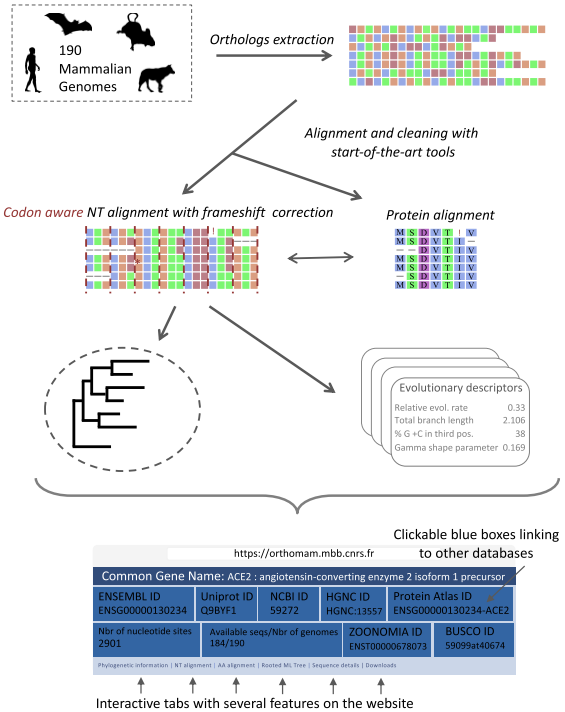
Rémi Allio [1,2], Frédéric Delsuc [2], Khalid Belkhir [2], Emmanuel J.P. Douzery [2], Vincent Ranwez [3] and Céline Scornavacca [2,*]

[1]CBGP, INRAE, CIRAD, IRD, Institut Agro, Univ. Montpellier, Montpellier, 34988, France
[2]ISEM, Univ. Montpellier, CNRS, IRD, Montpellier, 34095, France
[3]AGAP, Univ. Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, 34398, France
[*]To whom correspondence should be addressed. Tel: +33 467143697; Fax: +33 467145610; Email: celine.scornavacca@umontpellier.fr

Abstract

To date, the databases built to gather information on gene orthology do not provide end-users with descriptors of the molecular evolution information and phylogenetic pattern of these orthologues. In this context, we developed OrthoMaM, a database of ORTHOlogous MAmmalian Markers describing the evolutionary dynamics of coding sequences in mammalian genomes. OrthoMaM version 12 includes 15,868 alignments of orthologous coding sequences (CDS) from the 190 complete mammalian genomes currently available. All annotations and 1-to-1 orthology assignments are based on NCBI. Orthologous CDS can be mined for potential informative markers at the different taxonomic levels of the mammalian tree. To this end, several evolutionary descriptors of DNA sequences are provided for querying purposes (e.g. base composition and relative substitution rate). The graphical web interface allows the user to easily browse and sort the results of combined queries. The corresponding multiple sequence alignments and ML trees, inferred using state-of-the art approaches, are available for download both at the nucleotide and amino acid levels. OrthoMaM v12 can be used by researchers interested either in reconstructing the phylogenetic relationships of mammalian taxa or in understanding the evolutionary dynamics of coding sequences in their genomes. OrthoMaM is available for browsing, querying and complete or filtered download at https://orthomam.mbb.cnrs.fr/.

Graphical abstract



Received: August 14, 2023. **Revised:** September 19, 2023. **Editorial Decision:** September 20, 2023. **Accepted:** September 26, 2023

Introduction

OrthoMaM is a comprehensive curated database that relies on an expert phylogenetic framework to describe the evolutionary dynamics of orthologous genes in mammalian genomes. Since its first release (1), OrthoMaM has regularly evolved to include newly available genomes and incorporate up-to-date software in its automated analytic pipeline. The initial release contained a set of nucleotide exon alignments of 3,170 single-copy orthologous genes for the 12 mammalian genomes available in Ensembl (2) at the time. Since then, each new OrthoMaM version sequentially incorporated more mammalian genomes as they became available, while implementing new features (3). From the v10 release (4), we drastically changed our pipeline to include the ever-increasing number of mammalian genomes released in NCBI (5). In the present version, sequence alignments and phylogenies were computed with state-of-the-art tools. Nucleotide and amino acid alignments were obtained using our codon-aware multiple sequence alignment tool MACSE (6) together with efficient filtering methods such as HMMCleaner (7) and PhylteR (8). These high-quality alignments were then used as input for maximum likelihood phylogenetic inference performed with IQ-TREE (9). Importantly, the web interface has also been extensively redesigned to improve user experience.

Previous versions of our database have been widely used in the evolutionary and comparative genomics community. In particular, the inclusion of NCBI genomes generated much interest from users. Our database has been used, for instance, in studies aimed at reconstructing the evolutionary history of functionally important genes (10,11), reconstructing the phylogeny of numerous mammalian clades (12,13), studying the process governing the evolution of genome-wide base composition (14,15), inferring patterns of natural selection acting on protein-coding genes (16,17), and as a benchmark dataset for evaluating various bioinformatic methods (18–20), and testing the impact of different potential sources of phylogenetic incongruence (21). This usage panel will be widened by the many improvements provided with this updated version that gives access to the latest released NCBI genomes.

Materials and methods

Database content

In contrast to orthology databases such as OrthoDB (22), InParanoiDB (23), PhylomeDB (24) or EggNOG (25), which are solely based on protein sequences, the purpose of our OrthoMaM database is not to infer orthology relationships, as we rely on NCBI orthology predictions. The uniqueness of our database lies in providing a comprehensive, curated set of high-quality codon and amino acid alignments, together with corresponding phylogenetic trees, for all single-copy protein-coding genes annotated in mammalian genomes. The search facility permits querying the database either by sequence, gene ID, species, taxonomic level or evolutionary parameter values to download subsets of genes of interest. This allows a range of evolutionary genomic analyses to be performed at both the nucleotide and amino acid levels. The database can also be searched by sequence similarity using user-provided sequences. To our knowledge, it is the only DNA, protein and tree database to provide such a valuable resource for mammals, a group of organisms of biomedical, agronomic, and ecological importance.

To date, the OrthoMaM database includes 15,868 nucleotide and amino acid alignments for 190 species. The length of the orthologous genes ranges from ∼100 to ∼56k nucleotides (19,417 sites ± 2102), leading to a total of ∼31M of nucleotides (with more than 18M variable sites). The alignments are very complete with 177/190 species on average. Gene trees and statistics are provided for every marker and a supertree inferred from all the individual gene trees is also available for download.

Bioinformatic pipeline

The aim of the OrthoMaM database is to provide ready-to-use curated mammalian orthologous gene alignments to anyone interested in mammal evolution and focusing either on a specific gene or larger sets of orthologs. The two main efforts done to construct such a database have been: (i) to develop a pipeline comprising state-of-the-art approaches for pre-filtering, aligning, post-filtering and analysing gene alignments and (ii) to create a user-friendly website interface allowing users to easily collect genes and trees, but also precomputed associated statistics.

The OrthoMaM (OMM) database takes advantage of the publicly available NCBI database, which provides access to a large amount of raw genetic data, provided by scientists from all over the world. As of January 2023, a subset of 190 annotated assemblies (one assembly per mammalian species) was selected based on several assembly statistics (number of scaffolds, N50, etc.). The orthogroup annotation computed by the NCBI pipelines was used to evaluate gene orthology among the coding sequences (CDSs) of these assemblies. Orthogroups including at least four species and a single copy CDS for *Homo sapiens* (GCF_009914755.1), *Mus musculus* (GCF_000001635.27) and *Canis lupus dingo* (GCF_003254725.2) were selected and the corresponding single copy CDSs of the 190 considered assemblies were downloaded to be included in OrthoMaM (Figure 1, step 1). This process led to the selection of 15,879 CDSs including 181.1 out of 190 species on average (ranging from 4 to 190 species).

The next step was to perform an initial filtering of these orthologous CDSs to eliminate potentially erroneous sequences that had been included, due either to close paralogs resulting from recent gene duplications, or to annotation errors (Figure 1, steps 2–6). For the filtering step, we use PhylteR (8). This tool allows the detection and removal of outlier sequences in a set of gene alignments by iteratively removing taxa from the gene trees (inferred from these alignments) to optimise a score of concordance between all gene trees. PhylteR is particularly well adapted to remove potential erroneous sequences in CDS alignments since erroneous sequences will lead to abnormal phylogenetic placements or unexpectedly long phylogenetic branches. Since PhylteR requires gene trees, steps 2–5 of the OMM pipeline (Figure 1) were designed to quickly produce a first set of accurate gene trees for the CDSs downloaded from NCBI. First, we used the MACSE subprogram 'trimNonHomologousFragments' to eliminate potential contamination in every single-copy orthologous gene (Figure 1, step 2). This subprogram detects non-homologous fragments that often result from annotation errors and correspond to remaining intron fragments or untranslated regions

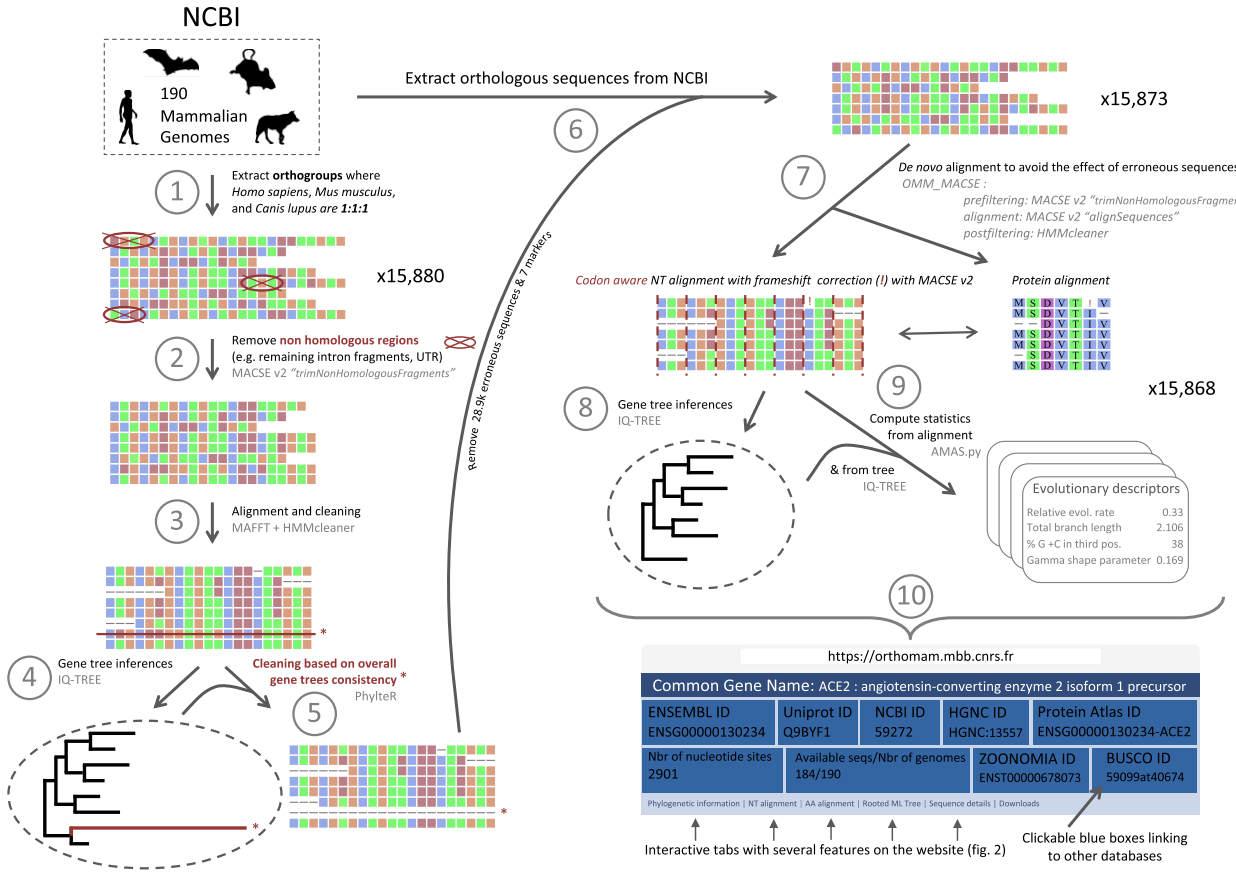


Figure 1. The OrthoMaM pipeline in short. Step 1 was performed for each genome assembly independently. Steps 2–4, 6 and 7–10 were performed independently for every marker. Step 5 was performed using five batches of ∼3180 markers. One additional step, not shown in the figure, consisted in inferring a Mammal supertree including the 190 species using all gene trees produced in step 8.

(UTRs). Then, the sequences of each orthologous gene were aligned using MAFFT (26) and a soft alignment cleaning was performed using HMMcleanNucl.pl (7) (Figure 1, step 3). The best evolutionary model for each gene was then chosen using ModelFinder as implemented in IQTREE v2.1.3 (-m MFP) and followed by gene tree reconstruction. Node supports were evaluated using ultrafast bootstraps estimated by IQ-TREE (-bb 1000; (9)). Finally, using the phylogenetic gene trees inferred from all CDSs as reference (combined in five batches of ∼3k markers to save computation time), PhylteR was able to detect 28,886 (1.01%) problematic sequences leading to strange-behaving phylogenetic branches. In most cases, PhylteR just removed a few sequences of a gene dataset to fix it, most likely corresponding to paralogous sequences. In seven extreme cases, where too many problematic sequences were detected in an alignment, PhylteR discarded the corresponding gene entirely (Figure 1, step 5: 15,880 versus 15,873 at step 1).

Because of the potential impact of the presence of erroneous sequences in previous alignment and cleaning steps (Figure 1, steps 2 and 3), we decided to remove the outlier sequences detected by PhylteR from the raw CDSs extracted from NCBI and restart alignments and cleaning from scratch (Figure 1, step 6). The new alignment step was performed using the OMM_MACSE v12.01 pipeline implemented in a Singularity container (27); the '–MACSE_min_MEM_length 8' option was used to save RAM when needed; Figure 1, step 7). This pipeline is designed to provide the best possible nucleotide and amino acid alignments for each marker by combining MAFFT (26) with state-of-the-art alignment cleaning methods (MACSE cleaning subprograms: trimNonHomologousFragments, (6); and HMMcleaner, (7)), and by detecting and correcting potential frameshifts in coding sequences (using MACSE v2 'alignSequence' subprogram, (6)). The highly accurate nucleotide alignments were then used to infer gene trees using IQ-TREE v2.1.3 with three partitions per CDS, corresponding to the three codon positions (Figure 1, step 8). The best evolutionary model was selected for each partition using ModelFinder implemented in IQ-TREE and merged if necessary (-m MFP+MERGE). Node supports were evaluated using ultrafast bootstraps estimated by IQ-TREE (-bb 1000). A second phylogenetic analysis was then performed using the resulting gene tree as constraint (-te and –blfix IQ-TREE options) and the model GTR + Γ (-m GTR+G) to infer the α shape parameter of the gamma distribution (9).

Finally, for each CDS marker, several evolutionary indicators (see below) were evaluated using AMAS.py (28), IQ-TREE inferences (de novo and with constrained models), and ERaBLE (29) (Figure 1, step 9). Gene level information (gene name, GO annotation), full sequence traceability information (sequence identifier in Ensembl/NCBI, filtering details), nucleotide and amino acid alignments, phylogenetic trees, as well as these evolutionary indicators are reported in the OrthoMaM website interface (Figure 1, step 10; see Figure 2).

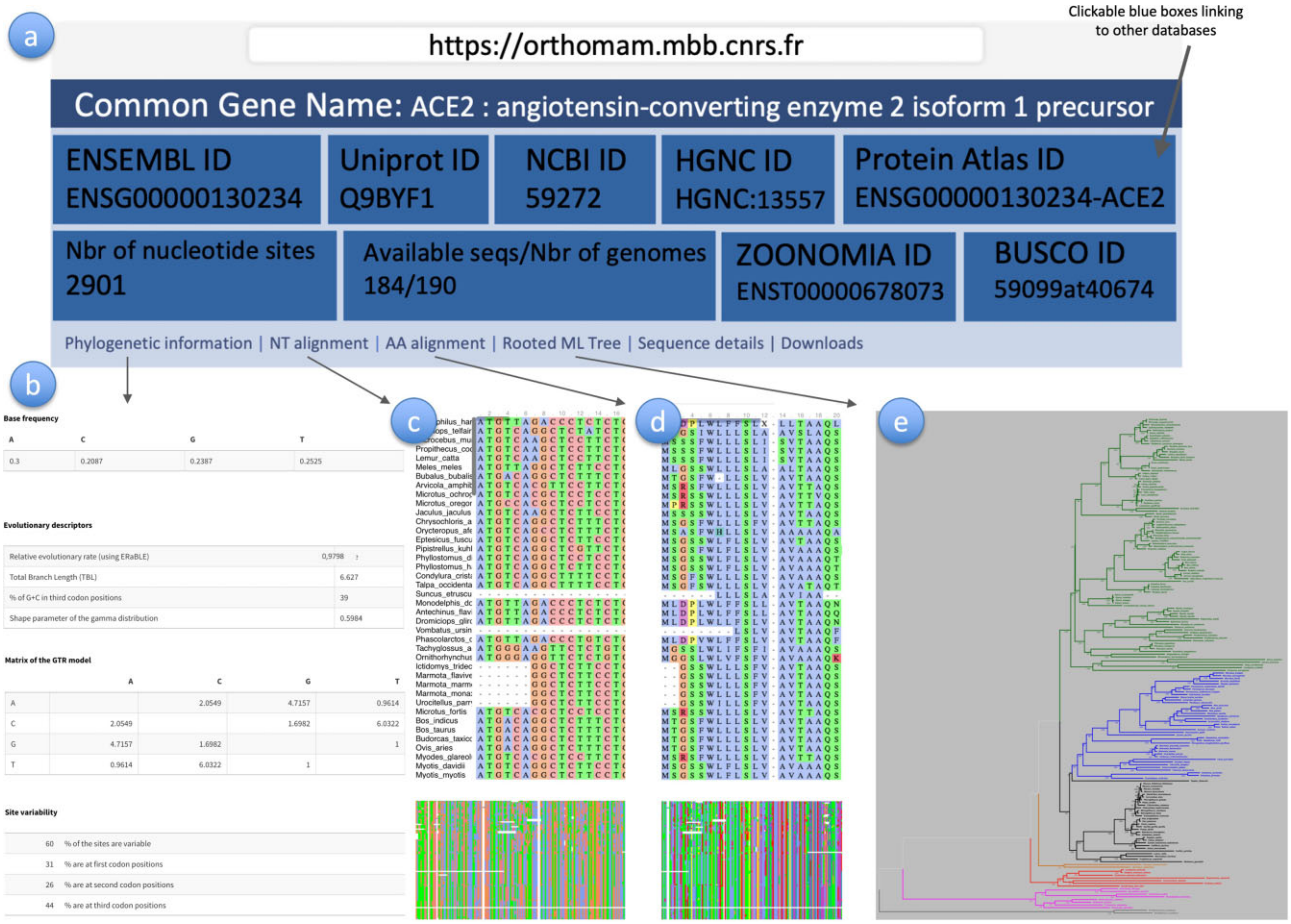


Figure 2. Website interface summary. This figure shows an example of the information available for every marker: General information and links to other databases (**a**), 'Phylogenetic information' (**b**), 'NT alignment' (**c**), 'AA alignment' (**d**), 'Rooted ML tree' (**e**). To navigate through very big alignments, the user can either click and drag directly on the alignment (c, d), or click on the zone of interest in the zoomed-out version of the alignment provided at the bottom of the page. To zoom in the phylogenetic tree (**e**), the user can click left on the tree, or click right and open a new tab where to zoom in or save the image. No illustration is provided for the 'Sequence details' tab since it consists of a simple table, and the 'Download' tab, which consists of a list of files for download.

Website interface

With the development of the latest version of OrthoMaM, we have incorporated some improvements to the website interface using the R Shiny framework (30), which should make the website more appealing.

Database access and query

While the full database can be downloaded from the website, to facilitate user experience, the OMM website allows users to search for markers in two different ways (Figure 2a), either by searching for one specific gene or by filtering the database using several parameters to download subsets.

On one hand, specific gene alignments can be access through the search of their NCBI gene ID ('search marker' page) or by using a BLAST research ('blast' page, 31), based on nucleotide (blastn or tblastx) or amino acid sequences (tblastn).

On the other hand, from the main page, one can access the 'query' page in which the database can be filtered by selecting specific ranges of evolutionary indicator values (relative evolutionary rate, percentage of G+C at third codon position, α shape parameter [Γ distribution], and alignment length) or specific conditions (*e.g.* genes present in all species, in a list of species, or in species belonging to a specific taxonomic group; markers part of the mammalian BUSCO genes; markers located on a given Human chromosome). For example, for the default query values, we have 339 CDSs located on Human chromosome 1, but this number goes up to 811 when the maximum GC3 percentage is fixed to 60.

For each CDS, OrthoMaM_v12 provides immediate access to many evolutionary features and information with a dedicated page (Figure 2). To present all information provided by the OMM website interface for each single-copy protein-coding gene included in the database, we used the page dedicated to the *ACE2* gene as an example. This gene is of particular medical significance as the ACE2 receptor is used by the SARS-CoV-2 and other coronaviruses to enter mammalian cells; several evolutionary analyses have been conducted for this gene to predict functionally important sites and to identify other potentially susceptible mammalian species besides humans (32,33).

General information and links to other databases

First, at the top of the marker page (Figure 2a), the common name of the gene is provided (ACE2: angiotensin-converting enzyme 2 isoform 1 precursor). Then, nine blue boxes

provide global information about the gene. The first five blocks (first line) give access to Ensembl, Uniprot, NCBI, HGNC and Protein Atlas databases in which information about each gene (through different identifiers: ENSG00000130234, Q9BYF1, 59272, HGNC:13557, ENSG00000130234-ACE2, respectively) is present. The four next boxes (second line) show the length of the alignment (2901 nucleotides), the number of species for which the gene is available (184 out of 190), a link to multi-codon alignment for the given gene as provided by the Zoonomia project, and a link to orthoDB details if the gene is included in the BUSCO list for mammals (mammalia_odb10, 34). Note that Zoonomia genomes have been/will be progressively integrated in NCBI. Since we strongly rely on the NCBI orthology, we prefer to wait until the genomes are annotated by the NCBI consortium to fully integrate them in OrthoMaM. In the meantime, we provide in the help page, a script to merge the OrthoMaM and Zoonomia raw files and run the OrthoMaM pipeline on the merged file.

Then, six tabs provide additional information and resources associated with the gene: 'Phylogenetic information', 'NT alignment', 'AA alignment', 'Rooted ML tree', 'Sequence details', 'Download'.

Evolutionary descriptors

The first tab, called 'Phylogenetic Information' (Figure 2b), displays the base frequencies, the substitution matrix under a GTR model, the site variability (percentage of variable sites at each codon position), and additional evolutionary descriptors estimated through a phylogenetic approach. These evolutionary descriptors are incorporated in the database, and some of them can be used to query CDSs (see 'Query' section below).

First, the 'Relative evolutionary rate' (RER) is estimated with ERaBLE (29). The relative evolutionary rate of a CDS is important to evaluate its usefulness in analyses of phylogenetics and molecular evolution. Faster-evolving genes will be more suitable for genomic comparisons at smaller taxonomic scales, while slower-evolving genes will be more suitable at deeper taxonomic scales. It is correlated with the second evolutionary descriptor of the marker, the total branch length (TBL), corresponding to the sum of all internal and terminal branches of the ML phylogram inferred from the corresponding alignment. Interestingly, the RER of OrthoMaM markers range from 0.127 to 5.92 corresponding to a 46.6-fold contrast between the slowest and fastest evolving genes.

The third evolutionary descriptor concerns the G+C content of the markers. G+C percentage varies among markers, and the variability is exacerbated at neutrally evolving synonymous third codon positions (GC3). Interestingly, because GC3 is correlated with G+C ($r^2 = 0.83$), it can be used as a proxy of gene variations in the composition of G+C.

Finally, the last evolutionary descriptor is the α shape parameter of the gamma distribution. This value quantifies the substitution rate variation among sites along the alignment. In genes undergoing strong constraint at the amino-acid level, α is low while α increases when the constraint is weaker. Interestingly, genes with $\alpha > 1$ have a strong phylogenetic potential as they accumulate variability quite evenly along their sequences, thus lessening the probability of multiple substitutions at the same sites. For example, BRCA1, with its high α value (1.642), has become a famous marker for phylogeny (35) and molecular evolution (36) in mammals.

Cleaned NT and AA alignments

For each CDS, the nucleotide alignment is provided in a dedicated tab called 'NT alignment' (Figure 2c). The corresponding amino acid alignment, which exactly matches the nucleotide alignment, is also provided in a second tab, called 'AA alignment' (Figure 2d). Both alignments are available for download. To navigate through very big alignments, the user can either click and drag directly on the alignment, or click on the zone of interest in the zoomed-out version of the alignment provided at the bottom of the page. Thanks to the zoomed-out view, we can easily see the long insertion present in *Lemur catta* for ACE2, which corresponds to a presumed exon duplication.

Rooted ML tree

For each OrthoMaM marker, a ML phylogenetic tree with branch lengths (i.e. phylogram) is provided (Figure 2e). It represents a synthesis of the base composition, substitution pattern, and among-site substitution rate heterogeneity of the corresponding CDS alignment. Phylogenetic trees were rooted through a sequential re-rooting procedure implemented in bio++ (37), using either monotremes, or marsupials, afrotherians, xenarthrans, laurasiatheria and euarchontoglires, in this order until one species to use as outgroup is present. To improve tree readability, the major clades have been coloured using the APE R-package (38, see Figure 2). To zoom in, the user can click left on the tree, or click right and open a new tab where to zoom in or save the image. The rooted gene trees in Newick format are available for download.

Sequence details

In the fifth and last tab describing the CDS, called 'Sequence details', information about each sequence of the alignment is provided. This information allows the sequence to be traced back to NCBI through a gene identifier and a transcript identifier. Chromosomal coordinates, strands and original sequence lengths are also provided for each species.

Download

For each marker, the evolutionary descriptors, sequence details, cleaned alignments and rooted ML tree in Newick format are provided for download in this tab.

For full sequence traceability, we also provide for download (i) alignments before the HMMcleaner postfiltering (step 7 of the pipeline) and (ii) annotated NBCI raw sequences.

The first corresponds to unfiltered alignments of sequences that are considered as orthologs by NCBI and our paralogy checks (PhylteR and the 'trimNonHomologousFragments'). The latter are the NCBI raw sequences where nucleotides that are not present in the final alignments because of all different steps of our pipeline are shown in lowercase.

Results and discussion

The earliest versions of OrthoMaM based on Ensembl also included individual orthologous exon alignments in addition to CDSs. This feature was temporarily abandoned from v10 because of the incorporation of NCBI data. In future versions, we plan on reinstating the exon database since powerful programs such as Minimap2 (39) and Miniprot (40) now allow efficiently extracting exons from genomic data (41). Future developments also include the implementation of a command line API allowing to interrogate the database

interactively and to download subsets of interest. In the near future, we aim at continuing to renew the database thanks to our automated pipeline allowing even more frequent updates while still adding new functionalities. These high-quality alignments and phylogenetic trees, along with relevant evolutionary information and the blast query functionality will be useful well beyond the evolutionary and comparative genomics community and will permit researchers that are not experts of the phylogenetic framework to access alignments and phylogenies computed with state-of-the-art tools.

Data availability

The data underlying this article are available at https://orthomam.mbb.cnrs.fr.

Acknowledgements

The authors would like to thank the Genotoul Bioinformatics Platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing computing resources and Rémy Dernat (ISEM) for the web server maintenance. This is the contribution ISEM 2023-223 of the Institut des Sciences de l'Evolution de Montpellier.

Author contributions: All authors have contributed to the design of the pipeline and the OMM website interface. R.A., E.J.P.D., V.R. and C.S. extracted the data, developed the pipeline, and produced the database content. K.B. developed the website interface. R.A. produced the figures. R.A., F.D. and C.S. drafted the paper with substantial input from all authors.

Funding

Agence Nationale de la Recherche [CEBA: ANR-10-LABX-25-01, CEMEB: ANR-10-LABX-0004, CoCoAlSeq: ANR-19-CE45-0012 to C.S.]; European Research Council [ConvergeAnt: ERC-2015-CoG-683257 to FD]. Funding for open access charge: Agence Nationale de la Recherche [CEBA: ANR-10-LABX-25-01].

Conflict of interest statement

None declared.

References

1. Ranwez,V., Delsuc,F., Ranwez,S., Belkhir,K., Tilak,M.K. and Douzery,E.J. (2007) OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol. Biol.*, **7**, 241.
2. Martin,F.J., Amode,M.R., Aneja,A., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Becker,A., Bennett,R., Berry,A., Bhai,J., *et al.* (2023) Ensembl 2023. *Nucleic Acids Res.*, **51**, D933–D941.
3. Douzery,E.J., Scornavacca,C., Romiguier,J., Belkhir,K., Galtier,N., Delsuc,F. and Ranwez,V. (2014) OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. *Mol. Biol. Evol.*, **31**, 1923–1928.
4. Scornavacca,C., Belkhir,K., Lopez,J., Dernat,R., Delsuc,F., Douzery,E.J. and Ranwez,V. (2019) OrthoMaM v10: scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Mol. Biol. Evol.*, **36**, 861–862.
5. Sayers,E.W., Bolton,E.E., Brister,J.R., Canese,K., Chan,J., Comeau,D.C., Farell,C.M., Feldgarden,M., Fine,A.M., Funk,K., *et al.* (2023) Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res.*, **51**, D29–D38.
6. Ranwez,V., Douzery,E.J., Cambon,C., Chantret,N. and Delsuc,F. (2018) MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.*, **35**, 2582–2584.
7. Di Franco,A., Poujol,R., Baurain,D. and Philippe,H. (2019) Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol. Biol.*, **19**, 21.
8. Comte,A., Tricou,T., Tannier,E., Joseph,J., Siberchicot,A., Penel,S., Allio,R., Delsuc,F., Dray,S. and de Vienne,D.M. (2023) PhylteR: efficient identification of outlier sequences in phylogenomic datasets. bioRxiv doi: https://doi.org/10.1101/2023.02.02.526888, 03 February 2023, preprint: not peer reviewed.
9. Minh,B.Q., Schmidt,H.A., Chernomor,O., Schrempf,D., Woodhams,M.D., Von Haeseler,A. and Lanfear,R. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, **37**, 1530–1534.
10. Mu,Y., Huang,X., Liu,R., Gai,Y., Liang,N., Yin,D., Shan,L., Xu,S. and Yang,G. (2021) ACPT gene is inactivated in mammalian lineages that lack enamel or teeth. *PeerJ*, **9**, e10219.
11. D'Oliviera,A., Dai,X., Mottaghinia,S., Geissler,E.P., Etienne,L., Zhang,Y. and Mugridge,J.S. (2023) Recognition and cleavage of human tRNA methyltransferase TRMT1 by the SARS-CoV-2 main protease. bioRxiv doi: https://doi.org/10.1101/2023.02.20.529306, 09 September 2023, preprint: not peer reviewed.
12. Mason,V.C., Helgen,K.M. and Murphy,W.J. (2019) Comparative phylogeography of forest-dependent mammals reveals Paleo-forest corridors throughout Sundaland. *J. Hered.*, **110**, 158–172.
13. Roycroft,E.J., Moussalli,A. and Rowe,K.C. (2020) Phylogenomics uncovers confidence and conflict in the rapid radiation of Australo-Papuan rodents. *Syst. Biol.*, **69**, 431–444.
14. Rousselle,M., Laverré,A., Figuet,E., Nabholz,B. and Galtier,N. (2019) Influence of recombination and GC-biased gene conversion on the adaptive and nonadaptive substitution rate in mammals versus birds. *Mol. Biol. Evol.*, **36**, 458–471.
15. Galtier,N. (2021) Fine-scale quantification of GC-biased gene conversion intensity in mammals. *Peer Commun. J.*, **1**, e17.
16. He,K., Liu,Q., Xu,D.M., Qi,F.Y., Bai,J., He,S.W., Chen,P., Zhou,X., Cai,W.-Z., Chen,Z.-Z., *et al.* (2021) Echolocation in soft-furred tree mice. *Science*, **372**, eaay1513.
17. Latrille,T., Rodrigue,N. and Lartillot,N. (2023) Genes and sites under adaptation at the phylogenetic scale also exhibit adaptation at the population-genetic scale. *Proc. Natl. Acad. Sci. U.S.A.*, **120**, e2214977120.
18. Abadi,S., Avram,O., Rosset,S., Pupko,T. and Mayrose,I. (2020) ModelTeller: model selection for optimal phylogenetic reconstruction using machine learning. *Mol. Biol. Evol.*, **37**, 3338–3352.
19. Islam,M., Sarker,K., Das,T., Reaz,R. and Bayzid,M.S. (2020) STELAR: a statistically consistent coalescent-based species tree estimation method by maximizing triplet consistency. *BMC Genomics*, **21**, 136.
20. Duchemin,L., Lanore,V., Veber,P. and Boussau,B. (2023) Evaluation of methods to detect shifts in directional selection at the genome scale. *Mol. Biol. Evol.*, **40**, msac247.
21. Scornavacca,C. and Galtier,N. (2017) Incomplete lineage sorting in mammalian phylogenomics. *Syst. Biol.*, **66**, 112–120.
22. Kuznetsov,D., Tegenfeldt,F., Manni,M., Seppey,M., Berkeley,M., Kriventseva,E.V. and Zdobnov,E.M. (2023) OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.*, **51**, D445–D451.
23. Persson,E. and Sonnhammer,E.L. (2023) InParanoiDB 9: ortholog groups for protein domains and full-length proteins. *J. Mol. Biol.*, **435**, 168001.
24. Fuentes,D., Molina,M., Chorostecki,U., Capella-Gutiérrez,S., Marcet-Houben,M. and Gabaldon,T. (2022) PhylomeDB V5: an

expanding repository for genome-wide catalogues of annotated gene phylogenies. *Nucleic Acids Res*., **50**, D1062–D1068.
25. Hernández-Plaza,A., Szklarczyk,D., Botas,J., Cantalapiedra,C.P., Giner-Lamia,J., Mende,D.R., Kirsch,R., Rattei,T., Letunic,I., Jensen,L., *et al.* (2023) eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res*., **51**, D389–D394.
26. 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.
27. Ranwez,V., Chantret,N. and Delsuc,F. (2021) Aligning Protein-Coding nucleotide sequences with MACSE. *Methods Mol Biol*., **2231**, 51–70.
28. Borowiec,M.L. (2016) AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, **4**, e1660.
29. Binet,M., Gascuel,O., Scornavacca,C., Douzery,E.J. and Pardi,F. (2016) Fast and accurate branch lengths estimation for phylogenomic trees. *BMC Bioinf*., **17**, 23.
30. Chang,W., Cheng,J., Allaire,J., Xie,Y. and McPherson,J. (2017) Shiny: web application framework for R. *R Package Version*, **1**, 2017.
31. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinf*., **10**, 421.
32. Damas,J., Hughes,G.M., Keough,K.C., Painter,C.A., Persky,N.S., Corbo,M., Hiller,M., Koepfli,K.-P., Pfenning,A.R., Zhao,H., *et al.* (2020) Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. *Proc. Natl. Acad. Sci*., **117**, 22311–22322.
33. Melin,A.D., Janiak,M.C., Marrone III,F., Arora,P.S. and Higham,J.P. (2020) Comparative ACE2 variation and primate COVID-19 risk. *Commun. Biol*., **3**, 641.
34. Manni,M., Berkeley,M.R., Seppey,M., Simão,F.A. and Zdobnov,E.M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol*., **38**, 4647–4654.
35. Madsen,O., Scally,M., Douady,C.J., Kao,D.J., DeBry,R.W., Adkins,R., Amrine,H.M., Stanhope,M.J., de Jong,W. and Springer,M.S. (2001) Parallel adaptive radiations in two major clades of placental mammals. *Nature*, **409**, 610–614.
36. Burk-Herrick,A., Scally,M., Amrine-Madsen,H., Stanhope,M.J. and Springer,M.S. (2006) Natural selection and mammalian BRCA1 sequences: elucidating functionally important sites relevant to breast cancer susceptibility in humans. *Mamm. Genome*, **17**, 257–270.
37. Dutheil,J., Gaillard,S., Bazin,E., Glémin,S., Ranwez,V., Galtier,N. and Belkhir,K. (2006) Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinf*., **7**, 188.
38. Paradis,E., Claude,J. and Strimmer,K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.
39. Li,H. (2021) New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, **37**, 4572–4574.
40. Li,H. (2023) Protein-to-genome alignment with miniprot. *Bioinformatics*, **39**, btad014.
41. Huang,N. and Li,H. (2023) miniBUSCO: a faster and more accurate reimplementation of BUSCO. bioRxiv doi: https://doi.org/10.1101/2023.06.03.543588, 06 June 2023, preprint: not peer reviewed.

Received: August 14, 2023. **Revised:** September 19, 2023. **Editorial Decision:** September 20, 2023. **Accepted:** September 26, 2023