OrthoMaM v8: A Database of Orthologous Exons and Coding Sequences for Comparative Genomics in Mammals

Emmanuel J. P. Douzery,*[,1] Celine Scornavacca,[1] Jonathan Romiguier,[1] Khalid Belkhir,[1] Nicolas Galtier,[1] Frédéric Delsuc,[1] and Vincent Ranwez[2]

[1]Institut des Sciences de l'Evolution de Montpellier (ISE-M), UMR 5554 CNRS IRD, Université Montpellier 2, Montpellier, France

[2]Montpellier SupAgro, UMR AGAP, Montpellier, France

***Corresponding author:** E-mail: emmanuel.douzery@univ-montp2.fr.

Associate editor: Xun Gu

Abstract

Comparative genomic studies extensively rely on alignments of orthologous sequences. Yet, selecting, gathering, and aligning orthologous exons and protein-coding sequences (CDS) that are relevant for a given evolutionary analysis can be a difficult and time-consuming task. In this context, we developed OrthoMaM, a database of ORTHOlogous MAmmalian Markers describing the evolutionary dynamics of orthologous genes in mammalian genomes using a phylogenetic framework. Since its first release in 2007, OrthoMaM has regularly evolved, not only to include newly available genomes but also to incorporate up-to-date software in its analytic pipeline. This eighth release integrates the 40 complete mammalian genomes available in Ensembl v73 and provides alignments, phylogenies, evolutionary descriptor information, and functional annotations for 13,404 single-copy orthologous CDS and 6,953 long exons. The graphical interface allows to easily explore OrthoMaM to identify markers with specific characteristics (e.g., taxa availability, alignment size, %G + C, evolutionary rate, chromosome location). It hence provides an efficient solution to sample preprocessed markers adapted to user-specific needs. OrthoMaM has proven to be a valuable resource for researchers interested in mammalian phylogenomics, evolutionary genomics, and has served as a source of benchmark empirical data sets in several methodological studies. OrthoMaM is available for browsing, query and complete or filtered downloads at http://www.orthomam.univ-montp2.fr/.

Key words: **orthologous sequences, mammals, coding sequences, phylogenomics, comparative genomics.**

Introduction

Orthologous protein-coding sequences (CDS) are of great interest to study patterns of organismal evolution (species phylogenies) and genomic processes (molecular evolution). The wide use of exons and CDS in phylogenomics and comparative genomics is facilitated by the existence of several independent databases of orthologs (Alexeyenko et al. 2006), each with their pros and cons. Some are generalist—for example, COG/KOG (Tatusov et al. 2003), HOGENOM (Dufayard et al. 2005), and InParanoid (Östlund et al. 2010), some are taxonomically specialized—for example, OPTIC (Heger and Ponting 2008) for vertebrates, INVHOGEN (Paulsen and von Haeseler 2006) for nonvertebrates, EvolMarkers (Li et al. 2012) for metazoans, FUNYBASE for fungi (Marthey et al. 2008), GreenPhylDB (Conte et al. 2008) for plants, HOBACGEN (Perriere et al. 2000) for bacteria, and some are built on functional information, such as OrthoDisease (O'Brien et al. 2004). In particular taxonomic groups, researchers have identified potentially useful phylogenetic DNA markers from complete genomes and have validated their use in nonmodel species such as primates (Horvath et al. 2008), actinopterygian fishes (Li et al. 2007), or rosids (Duarte et al. 2010). However, these databases generally do not provide end-users with key parameters describing the evolutionary pattern of orthologs, and orientating the choice of the molecular markers to be studied from the viewpoint of phylogenomic and molecular evolution. Also, few of them provide high-quality nucleotide and amino acid alignments preserving the key underlying codon structure.

OrthoMaM (Ranwez et al. 2007) is a database of ORTHOlogous MAmmalian coding sequence Markers, which helps filling these gaps. It provides high-quality codon alignments of exon and CDS markers associated with a detailed characterization of their evolutionary dynamics in terms of phylogenetic signal, base composition, substitution rate, and chromosome location. Moreover, OrthoMaM focuses only on one-to-one orthologs identified by Ensembl (Flicek et al. 2014), that is, sequences for which no duplication is detected since the last common ancestor of the corresponding species. Indeed, as one-to-one orthologs are unaffected by complex intragenomic processes such as gene duplication or gene loss, the differences in their sequences are ensured to have occurred through common descent and therefore reflect the divergence between species.

Database Overview and Improvements

Mammalia is among the first animal taxa with many complete genomes available and has been extensively used to define most of the gold-standard methods in phylogenomic and molecular evolution studies. Based on the 12 mammalian genomes available in Ensembl v41, the first version of

Article

OrthoMaM was released in July 2007 and contained 3,170 exons (Ranwez et al. 2007).

Several major improvements have been made since then. In the current version (OrthoMaM version 8, October 2013, based on Ensembl v73), the database includes 6,953 exons and covers 40 mammalian species. In addition to exons, full orthologous CDS are now available. Queries have been made more flexible and can be performed taxonomically. Results can be dynamically sorted according to key descriptors, for example, number of orthologs, alignment length, α parameter of the among-site substitution rate heterogeneity, and G + C nucleotide composition on third codon positions (%GC3). The latter statistics has recently been connected to the performance of CDS as phylogenetic markers (Romiguier et al. 2013). Nucleotide and amino acid alignments, maximum likelihood (ML) gene trees, and detailed marker information can be downloaded for all exons and CDS. To improve readability, the phylogenetic tree of each marker is colored according to the major mammalian clades using the APE package (Paradis 2006). We also enriched the information associated with each marker by linking exons to their corresponding CDS and including functional annotations (gene ontology concepts) graphically displayed thanks to OntoFocus (Ranwez et al. 2012) and Graphviz (Ellson et al. 2002). Figure 1 displays screenshots associated with a given query on the OrthoMaM website.

The current OrthoMaM release contains a total of 13,404 CDS markers covering half of the known mammalian genes and providing a uniform representativity along chromosomes (fig. 2*a*). However, the number of available CDS widely varies among species, mainly because of the uneven sequencing coverage of the corresponding genomes. Figure 2*b* provides the phylogeny of the 40 species represented in OrthoMaM together with the number of CDS available for the different species and clades. For example, 973 CDS markers share the full set of 36 placental mammals of OrthoMaM, and 5,806 CDS markers share the full subset of 10 primates.

The OrthoMaM Pipeline

Identification of Orthologous Sequences

We start by using Ensembl annotations (Flicek et al. 2014) to identify one-to-one orthologous genes among pairs of three high-coverage reference species (*Homo–Mus*, *Homo–Canis*, and *Mus–Canis*). We then enrich each of those clusters of one-to-one orthologs by adding sequences of additional mammals that are annotated as one-to-one orthologs to the human gene (Ranwez et al. 2007). Note that the chromosomal distribution of OrthoMaM human genes basically mirrors the distribution of the full set of Ensembl human genes (fig. 2*a*), which is to be expected from an unbiased database.

Those clusters of one-to-one orthologous genes are turned into clusters of one-to-one orthologous CDS by selecting the longest transcript of each gene. We choose to consider the longest sequence as this is the one used by Ensembl to define the orthology relationships among genes, and this will maximize the evolutionary information to be analyzed.

The one-to-one orthologous exon clusters are not provided by Ensembl. Their identification is complicated by alternative splicing and by the variability in number and length of exons of a given gene across species. We tackle those problems by relying on the alignments of the one-to-one orthologous CDS to infer one-to-one orthology among their exons. Each human exon annotated by Ensembl initiates a one-to-one orthologous exon cluster. Exons from additional

Fig. 1. Screenshots from the OrthoMaM website. Here, we searched for CDS with 15–40 mammals, a relative evolutionary rate between 0.5 and 3, an α parameter of the Γ distribution ranging from 1 to 1.5, and a GC3 between 22% and 35%. We got 23 target CDS and focused on the LRRC63 marker. We then visualized the evolutionary dynamics parameters, the first 80 sites of the DNA alignment, and the corresponding phylogenetic tree.

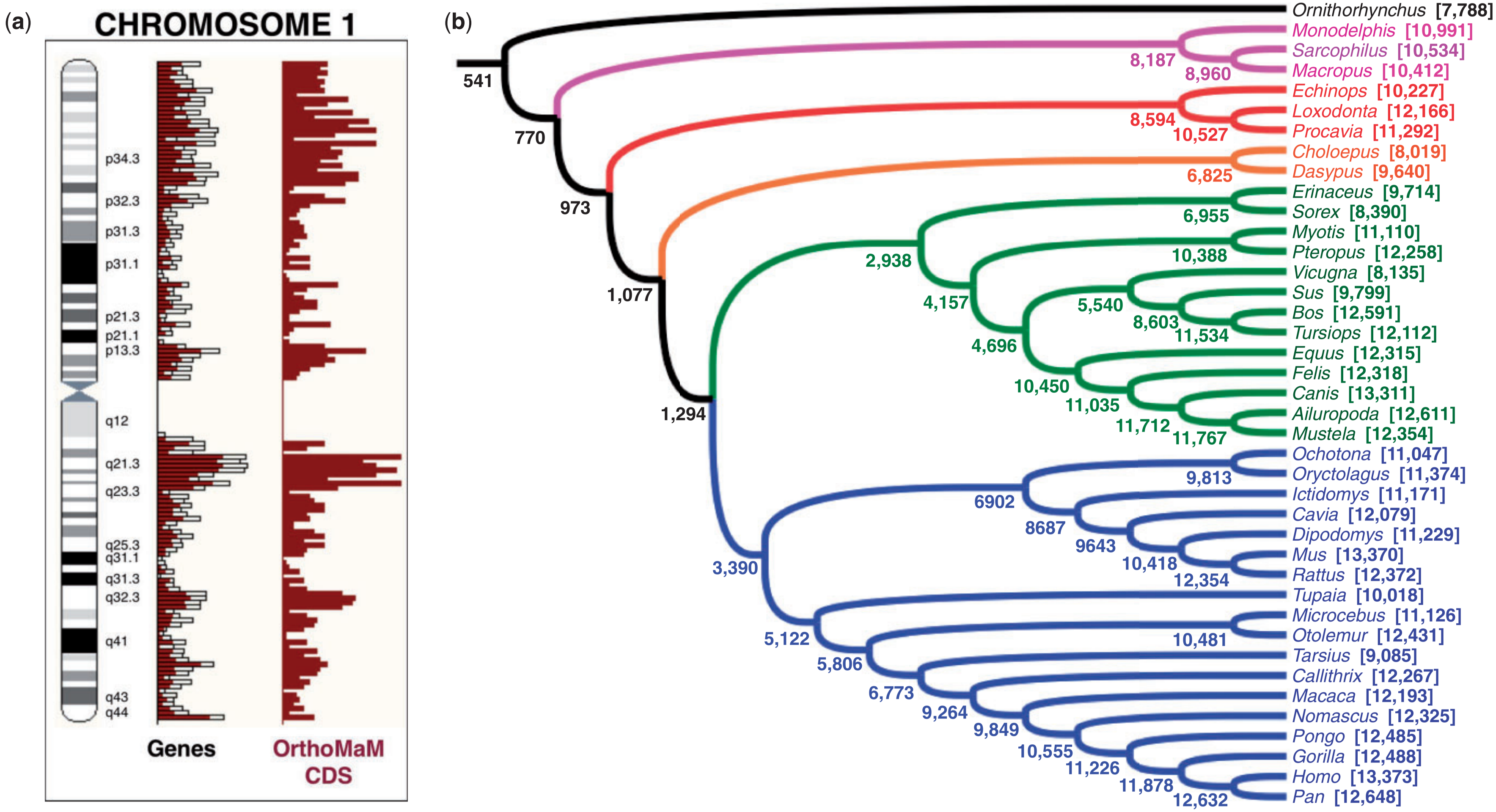

Fig. 2. (*a*) Genomic distribution of OrthoMaM CDS along human chromosome 1. The ideogram for human chromosome 1 is provided together with the distribution of OrthoMaM CDS (dark red bars to the right). The distributions of Ensembl predicted genes (white bars) and database-known genes (red bars) are also indicated (centre). (*b*) The phylogeny of the 40 species present in OrthoMaM. For each species, we provide the number of available CDS. For each node, we also indicate the number of CDS markers containing all species of the corresponding clade.

species are added to this cluster if they share a number of identical amino acids greater than half the length of the CDS alignment restricted to the candidate exon and the human one. This similarity threshold ensures that no more than one exon from a given species will be included in the predicted set of orthologs. Initial exon alignments longer than 400 sites are selected as our evolutionary marker descriptors are not meaningful enough for shorter sequences. Clusters with less than four sequences are discarded for the same reason.

Alignments and Trees

CDS and exon sequences are aligned at the codon level in two steps. First, the translated amino acids are aligned using MAFFT (Katoh et al. 2005) and gaps are reported onto the nucleotide sequences. This alignment is then refined using MACSE (Ranwez et al. 2011) to obtain a final codon alignment unaffected by frameshifts, misassemblies, and sequencing errors. Nucleotide and amino acid alignments are then filtered to remove spurious sequences and/or codons using trimAl (Capella-Gutiérrez et al. 2009). The filtering is conducted under the "automated1" option, which has been specifically designed to clean alignments before conducting ML phylogenetic inference. This step can yield final alignments shorter than 400 sites though the average length is far higher for both exons (956 sites) and CDS (1,850 sites). To ensure data traceability, each sequence is linked to the corresponding Ensembl CDS/exon. Moreover, each OrthoMaM alignment is available for download before and after filtering. All previous releases of OrthoMaM also remain available through the website.

The ML tree is identified for each marker by analyzing codon alignments with RAxML (Stamatakis 2006) under the general time reversible (GTR) + Γ model (Yang 1996). We acknowledge that using the proper model of sequence evolution is vital in probabilistic inference. However, we here used the same model for all CDS and exons because 1) it warrants a fair comparison among all markers of the database, 2) it is the one that best fits the majority of the markers (Ranwez et al. 2007), 3) the GTR exchangeability matrix is the only one available at the nucleotide level in RAxML, and 4) the parameter-rich GTR + Γ model is more likely to introduce increased variance rather than bias in the estimates (Lemmon and Moriarty 2004).

All parameters describing the evolutionary dynamics of exons and CDS are gathered by running PAUP* (Swofford 2003) on the ML tree inferred by RAxML. Branch lengths of ML phylograms are also examined, and if some exceed the unrealistic value of two substitutions per site, the corresponding alignment is excluded from OrthoMaM. This phylogenetic-based filter enables to detect and remove markers that likely contain misaligned sequences, misspecified open reading frames, or misannotated paralogs.

Database Updates and Scalability

OrthoMaM is regularly updated and its pipeline is constantly optimized to keep pace with the ever increasing number of available genomes and software developments in the field. Orthology annotation and sequences are now retrieved using the BioMart facilities, which allow massive retrieval of Ensembl data (Flicek et al. 2014). Those data are processed by home made Java tools to identify clusters of one-to-one orthologous CDS/exons. Phylogenetic analyses rely on shell scripts to chain up-to-date software. The website is based on a php/mysql database for query facilities and on XML/XSLT for exchange and graphic representation of marker details. All analyses are performed on the computing cluster of the Montpellier Bioinformatics Biodiversity (MBB) platform.

Query Options

There are three entry points in OrthoMaM. First, exon and CDS sections can be graphically browsed using a clickable phylogeny and ideograms of human chromosomes. Second, markers can be queried according to several of their key characteristics, including: minimal alignment length, number of sequences, mandatory species, base composition (%GC3), relative evolutionary rate of the marker, Ensembl gene identifier or HUGO gene symbol (see fig. 1). Third, a BLAST (Altschul et al. 1990) similarity search can be run to find OrthoMaM markers related to a given request.

Examples of Contributions

OrthoMaM has proven its usefulness in several phylogenomic and comparative genomic studies. We briefly list some of them to illustrate the broad spectrum of analyses facilitated by OrthoMaM. Our database has been used for developing new markers in multigene phylogenetic studies (Zhou et al. 2011; Hassanin et al. 2013) and also as a source of large-scale molecular data in phylogenomic (Parker et al. 2013; Romiguier et al. 2013), molecular dating (Schrago and Voloch 2013), and evolutionary genomic (Galtier et al. 2009; Romiguier et al. 2010; Rorick and Wagner 2011; Lartillot 2013) analyses. The high-quality codon alignments have also been utilized as benchmark empirical data sets for testing new analytical methods (Egan et al. 2008; López-Giráldez and Townsend 2011; Li and Drummond 2012; Wu et al. 2013) and for detecting footprints of purifying or positive selection (Jobson et al. 2010; Laguette et al. 2012). Finally, the inferred ML gene trees have served for assessing the performance of supertree methods (Scornavacca et al. 2008; Ranwez et al. 2010). With the ongoing pace of mammalian genome sequencing, we envision an enhanced potential for the uses of OrthoMaM in comparative genomic studies aiming at understanding the evolutionary dynamics of protein-coding genes.

Future Prospects

The primary aim of OrthoMaM is to provide high-quality genome scale alignments and phylogenetic analysis for one-to-one orthologous exons and CDS among mammals. Its analysis pipeline strategy has been adapted to cope with the increasing number of mammalian genomes that will be released in the upcoming years. This bioinformatic pipeline is constantly improved and we are currently testing the possibility of relying on codon-based phylogenetic inference using codon-phyML (Gil et al. 2013) and including in future releases per branch dN/dS estimations using mapNH (Romiguier et al. 2012). Moreover, we are considering possible solutions to filter only parts of a sequence in order to further improve the quality of our codon alignments with respect to potential exon annotation errors in CDS. We are also evaluating the relevance of expanding the database toward noncoding markers, such as intronic, untranslated, and regulatory regions.

Acknowledgments

This work was supported by the Montpellier Bioinformatics Biodiversity platform, the Agence Nationale de la Recherche "Investissements d'avenir/Bioinformatique" (ANR-10-BINF-01-02 "Ancestrome"), and the European Research Council ("PopPhyl": Population phylogenomics). The authors thank two reviewers for their comments on the manuscript. This publication is contribution 2014-022 of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554—CNRS-IRD-UM2).

References

Alexeyenko A, Lindberg J, Pérez-Bercoff A, Sonnhammer E. 2006. Overview and comparison of ortholog databases. *Drug Discov Today Technol.* 3:137–143.

Altschul S, Gish W, Miller W, Myers E, Lipman D. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.

Conte MG, Gaillard S, Lanau N, Rouard M, Périn C. 2008. GreenPhylDB: a database for plant comparative genomics. *Nucleic Acids Res.* 36: D991–D998.

Duarte J, Wall P, Edger P, Landherr L, Ma H, Pires J, Leebens-Mack J, dePamphilis C. 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol Biol.* 10:61.

Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perriere G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 21:2596–2603.

Egan A, Mahurkar A, Crabtree J, Badger JH, Carlton JM, Silva JC. 2008. IDEA: interactive display for evolutionary analyses. *BMC Bioinformatics* 9:524.

Ellson J, Gansner E, Koutsofios L, North SC, Woodhull G. 2002. Graphviz—open source graph drawing tools. *Graph Drawing* 2265:483–484.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res.* 42(Database issue):D749–D755.

Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.* 25:1–5.

Gil M, Zanetti MS, Zoller S, Anisimova M. 2013. Codonphyml: fast maximum likelihood phylogeny estimation under codon substitution models. *Mol Biol Evol.* 30(6):1270–1280.

Hassanin A, An J, Ropiquet A, Nguyen TT, Couloux A. 2013. Combining multiple autosomal introns for studying shallow phylogeny and taxonomy of Laurasiatherian mammals: application to the tribe Bovini (Cetartiodactyla, Bovidae). *Mol Phylogenet Evol.* 66(3):766–775.

Heger A, Ponting CP. 2008. OPTIC: orthologous and paralogous transcripts in clades. *Nucleic Acids Res.* 36:D267–D270.

Horvath JE, Weisrock DW, Embry SL, Fiorentino I, Balhoff JP, Kappeler P, Wray GA, Willard HF, Yoder AD. 2008. Development and application of a phylogenomic toolkit: resolving the evolutionary history of Madagascar's lemurs. *Genome Res.* 18:489–499.

Jobson R, Nabholz B, Galtier N. 2010. An evolutionary genome scan for longevity-related natural selection in mammals. *Mol Biol Evol.* 27: 840–847.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33:511–518.

Laguette N, Rahm N, Sobhian B, Chable-Bessia C, Munch J, Snoeck J, Sauter D, Switzer WM, Heneine W, Kirchhoff F, et al. 2012. Evolutionary and functional analyses of the interaction between the myeloid restriction factor SAMHD1 and the lentiviral Vpx protein. *Cell Host Microbe* 11:205–217.

Lartillot N. 2013. Phylogenetic patterns of gc-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Mol Biol Evol.* 30(3):489–502.

Lemmon A, Moriarty E. 2004. The importance of proper model assumption in Bayesian phylogenetics. *Syst Biol.* 53:265–277.

Li C, Riethoven J-J, Naylor G. 2012. Evolmarkers: a database for mining exon and intron markers for evolution, ecology and conservation studies. *Mol Ecol Res.* 12:967–971.

Li CH, Orti G, Zhang G, Lu GQ. 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evol Biol.* 7:44.

Li WLS, Drummond AJ. 2012. Model averaging and bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. *Mol Biol Evol.* 29(2):751–761.

López-Giráldez F, Townsend JP. 2011. Phydesign: an online application for profiling phylogenetic informativeness. *BMC Evol Biol.* 11(1):152.

Marthey S, Aguileta G, Rodolphe F, Gendrault A, Giraud T, Fournier E, Lopez-Villavicencio M, Gautier A, Lebrun M-H, Chiapello H. 2008. FUNYBASE: a FUNgal phYlogenomic dataBASE. *BMC Bioinformatics* 9:456.

O'Brien K, Westerlund I, Sonnhammer E. 2004. OrthoDisease: a database of human disease orthologs. *Hum Mutat.* 24:112–119.

Östlund G, Schmitt T, Forslund K, Köstler T, Messina D, Roopra S, Frings O, Sonnhammer E. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38:D196–D203.

Paradis E. 2006. Analysis of phylogenetics and evolution with R. Use R! series. New York: Springer Science + Business Media.

Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* 502(7470):228–231.

Paulsen I, von Haeseler A. 2006. INVHOGEN: a database of homologous invertebrate genes. *Nucleic Acids Res.* 34(Database issue):D349–D353.

Perriere G, Duret L, Gouy M. 2000. HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.* 10:379–385.

Ranwez V, Criscuolo A, Douzery EJ. 2010. Supertriplets: a triplet-based supertree approach to phylogenomics. *Bioinformatics* 26(12):i115–i123.

Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak M, Douzery EJP. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol Biol.* 7:241.

Ranwez V, Harispe S, Delsuc F, Douzery E. 2011. MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS One* 6(9):e22594.

Ranwez V, Ranwez S, Janaqi S. 2012. Subontology extraction using hyponym and hypernym closure on is-a directed acyclic graphs. *IEEE Trans Knowl Data Eng.* 24(12):2288–2300.

Romiguier J, Figuet E, Galtier N, Douzery E, Boussau B, Dutheil J, Ranwez V. 2012. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. *PLoS One* 7(3):e33852.

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJ. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol Biol Evol.* 30(9):2134–2144.

Romiguier J, Ranwez V, Douzery E, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res.* 20:1001–1009.
Rorick MM, Wagner GP. 2011. Protein structural modularity and robustness are associated with evolvability. *Genome Biol Evol.* 3:456.
Schrago C, Voloch C. 2013. The precision of the hominid timescale estimated by relaxed clock methods. *J Evol Biol.* 26(4):746–755.
Scornavacca C, Berry V, Lefort V, Douzery EJ, Ranwez V. 2008. PhySIC_IST: cleaning source trees to infer more informative supertrees. *BMC Bioinformatics* 9(1):413.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.
Swofford DL. 2003. PAUP*: phylogenetic analysis using parsimony (* and other methods). Version 4.0b10. Sunderland (MA): Sinauer Associates.
Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, Koonin E, Krylov D, Mazumder R, Mekhedov S, Nikolskaya A, et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41.
Wu C-H, Suchard MA, Drummond AJ. 2013. Bayesian selection of nucleotide substitution models and their site assignments. *Mol Biol Evol.* 30(3):669–688.
Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol.* 11(9):367–372.
Zhou X, Xu S, Zhang P, Yang G. 2011. Developing a series of conservative anchor markers and their application to phylogenomics of laurasiatherian mammals. *Mol Ecol Res.* 11(1):134–140.