Published online 3 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D787–D792***
doi:10.1093/nar/gkm878

Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees

Akihiro Matsuya[1,2], Ryuichi Sakate[1,3,*], Yoshihiro Kawahara[1,3], Kanako O. Koyanagi[4], Yoshiharu Sato[1,3], Yasuyuki Fujii[1,3], Chisato Yamasaki[1,3], Takuya Habara[1,3], Hajime Nakaoka[5], Fusano Todokoro[1,6], Kaori Yamaguchi[1,3], Toshinori Endo[4], Satoshi OOta[7], Wojciech Makalowski[8], Kazuho Ikeo[9], Yoshiyuki Suzuki[9], Kousuke Hanada[9], Katsuyuki Hashimoto[10], Momoki Hirai[11], Hisakazu Iwama[12], Naruya Saitou[13], Aiko T. Hiraki[1,3], Lihua Jin[9], Yayoi Kaneko[1,3], Masako Kanno[1,3], Katsuhiko Murakami[1,3], Akiko Ogura Noda[1,3], Naomi Saichi[1,3], Ryoko Sanbonmatsu[1,3], Mami Suzuki[1,3], Jun-ichi Takeda[1,3], Masayuki Tanaka[1,3], Takashi Gojobori[3,9], Tadashi Imanishi[3] and Takeshi Itoh[3,14]

[1]Integrated Database Group, Japan Biological Information Research Center, Japan Biological Informatics Consortium, [2]Government & Public Corporation Information Systems, Hitachi, Co., Ltd., [3]Integrated Database Group, Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, [4]Graduate School of Information Science and Technology, Hokkaido University, Hokkaido, [5]C's Lab Co., Ltd., Hokkaido, [6]DYNACOM Co., Ltd, Chiba, [7]BioResource Center, RIKEN, Ibaraki, Japan, [8]Institute of Bioinformatics, University of Muenster, Muenster, Germany, [9]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Shizuoka, [10]Department of Biomedical Resources, National Institute of Biomedical Innovation, Osaka, [11]International Research and Educational Institute for Integrated Medical Sciences, Tokyo Women's Medical University, Tokyo, [12]Kagawa University, Kagawa, [13]Department of Population Genetics, National Institute of Genetics, Shizuoka and [14]Division of Genome and Biodiversity Research, National Institute of Agrobiological Sciences, Ibaraki, Japan

Received August 15, 2007; Revised September 27, 2007; Accepted October 1, 2007

ABSTRACT

Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Currently, with the rapid growth of transcriptome data of various species, more reliable orthology information is prerequisite for further studies. However, detection of orthologs could be erroneous if pairwise distance-based methods, such as reciprocal BLAST searches, are utilized. Thus, as a sub-database of H-InvDB, an integrated database of annotated human genes (http://h-invitational.jp/), we constructed a fully curated database of evolutionary features of human genes, called 'Evola'. In the process of the ortholog detection, computational analysis based on conserved genome synteny and transcript sequence similarity was followed by manual curation by researchers examining phylogenetic trees. In total, 18 968 human genes have orthologs among 11 vertebrates (chimpanzee, mouse, cow, chicken, zebrafish, etc.), either computationally detected or manually curated orthologs. Evola provides amino acid sequence alignments and phylogenetic trees of orthologs and homologs. In 'd_N/d_S view', natural selection on genes can be analyzed between human and other species. In 'Locus maps', all transcript variants and their exon/intron structures can be compared among orthologous gene loci. We expect the Evola to serve as a comprehensive and reliable database to be utilized in comparative analyses for obtaining new knowledge about human genes. Evola is available at http://www.h-invitational.jp/evola/.

INTRODUCTION

A large number of genome and transcript sequences accumulated in the last decade give us an opportunity

*To whom correspondence should be addressed. Tel: +81 3 3599 8800; Fax: +81 3 3599 8801; Email: rsakate@ni.aist.go.jp

for large-scale comparative analyses. In particular, detection of orthologs, groups of genes in different species that evolved by speciation, accelerates functional and evolutionary studies. Despite the past efforts to develop bioinformatics methods for analyzing a large number of sequences, it is still a challenge to comprehensively identify orthologs between species. A number of automated pairwise distance-based methods for ortholog detection have been proposed, as represented by the reciprocal best BLAST hits (RBH) method (1) and the reciprocal smallest distance (RSD) method (2). However, as genes might have frequently undergone duplications and losses in evolutionary lineages leading to human (3), pairwise distance-based methods might lead to erroneous inferences of phylogenetic relationships and thus of orthologs. Thus, phylogenetic tree-based detection can be the most plausible solution to provide more reliable orthologs.

Here this database 'Evola', a sub-database complementary to the H-Invitational database (H-InvDB), was developed to provide orthology information for the originally annotated human genes in H-InvDB. Evola features its ortholog detection in which genome synteny-based computational analysis was followed by manual curation of molecular phylogenetic trees. Evola differs in this way from other ortholog databases such as Inparanoid (4), Ensembl-Compara (5), Homologene (6), HOGENOM (7) and TreeFam (8). These databases are based on BLAST hits (Inparanoid), BLAST hits and synteny (Ensembl-Compara and Homologene) and phylogenetic trees (HOGENOM and TreeFam). The concept of Evola is that genomic region (gene locus) is a unit of genes that are duplicated or lost. In collaboration with H-InvDB, Evola enables users to compare gene structure, transcript variants, upstream/downstream region of the genome among species.

H-InvDB is an integrated database of annotated human genes providing annotation of human full-length enriched cDNAs (9,10,11). At the meetings of the Human Full-Length cDNA Annotation Invitational held in Japan (2002 and 2003), Evola started with H-InvDB to annotate evolutionary features of the human genes. With several updates afterwards and a subsequent All Human Genes Evolutionary Annotation (AHG-EV) meeting in 2006, the current strategy of evolutionary annotation (computational analysis and manual curation) in Evola has been established. Orthology information for human and other 11 vertebrates is currently included in the Evola: human, chimpanzee, macaque, mouse, rat, dog, cow, opossum, chicken, zebrafish, Tetraodon and Fugu. Several visualization tools are incorporated into the database, including sequence alignment viewer, natural selection plot and graphical representation of orthologous gene loci among different species. Evola is now one of the databases listed in the Comparison of Orthology Predictions project of the HUGO Gene Nomenclature Committee (HGNC, http://www.genenames.org/).

ORTHOLOG DETECTION

Computational analysis: Ortholog detection based on conserved genomic synteny and pairwise distance

Species for ortholog detection were selected with consideration of completeness of their genome assemblies (chromosome level), abundance of transcript sequences (~20 000) and importance in biology (intensively studied or a representative of a phylogenetic clade). Whole genome sequence assemblies of human (hg18), chimpanzee (panTro2), macaque (rheMac2), mouse (mm8), rat (rn4), dog (canFam2), cow (rn4), opossum (monDom4), chicken (galGal3), zebrafish (danRer4), Tetraodon (tetNig1) and Fugu (fr1) were downloaded from UCSC (http://genome.ucsc.edu/). Conserved syntenic regions were detected by a modified pairwise genome alignment method (12) using BLASTZ (13) with the options of C = 2, T = 4, Y = 3400 between human and other primates (between more similar genome sequences), and C = 2 between human and non-primate vertebrates (between less similar genome sequences).

For human transcripts, H-InvDB representative transcripts (HITs) were used. Other vertebrates' transcripts (mRNAs) were downloaded from DDBJ (http://www.ddbj.nig.ac.jp/) release 66, Ensembl (http://www.ensembl.org/) release 38 and RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) release 17, and their genomic locations (one location per transcript) were detected on cognate genomes by a hybrid method using BLAT (14), BLAST (15) and est2genome (16) as they were used to detect genomic locations of human transcripts in H-InvDB. Representative transcripts (one transcript per gene locus) were determined in consideration of percent identity and coverage to the genome, number of exons, etc. of all transcripts in each locus (9,10,11). Thus, in Evola, representative transcripts were defined as genes.

Lengths of overlapping exons of each gene pair between human and other species were calculated in the genome alignment. A gene pair with the maximum length was selected as the best assignment (not a minimum length was defined). Every gene in a species was assigned to a gene in the other species. If two human genes were assigned to one mouse gene, this was defined as a two-to-one ortholog. As a result, Evola contains not only one-to-one orthologs but also many-to-many orthologs. For all the assignment pairs, coding sequences (CDSs) and amino acid (a.a.) sequences of other species were predicted by FASTY (17). They were predicted by comparing with the amino acid sequences of the corresponding human genes. Finally, if the length of the alignable region between human and other species ortholog candidates was $\geq$80 a.a., they were defined as computationally detected orthologs.

Manual curation: Examination of phylogenetic trees by experts

Homologs of human genes (amino acid sequences) were obtained from UniProt (http://www.uniprot.org/) and human RefSeq (NP) by FASTY similarity searches with the option of E-value of $<$1e$-$5. For each human gene, a

sequence set consisting of both the computationally detected orthologs and the homologs was prepared. For these sequence sets, phylogenetic trees were constructed by the neighbor-joining (NJ) method (18). In detail, multiple amino acid sequence alignments and phylogenetic trees were constructed by ClustalW (19) with the options of bootstrap = 1000, seed = 1, kimura, tossgaps, bootlabels = node.

Phylogenetic trees were examined by experts in the field of molecular evolution, who attended the evolutionary annotation meetings described in the introduction. The trees were drawn by NJplot (http://pbil.univ-lyon1.fr/software/njplot.html) and the default rooting was used. Discarding or re-rooting the tree was judged by the experts if necessary. All the ortholog pairs of human and other species detected by the computational analysis were examined (Figure 1). The primary principles of manual curation in Evola to be checked were as follows. [1] Phylogenetic topology between gene tree and species tree is consistent. As a gene tree, the minimum sub-clade including the pair (a part of the tree) was examined. As a species tree of reference, a phylogenetic tree indicating the trifurcation among primates, rodents and Laurasiatherian (dog, cow, etc.) species (20) was used, because the phylogenetic relationship has been controversial among them (21). In fact, we found that ((human–mouse)–dog) clades for some genes and ((human–dog)–mouse) clades for other genes. [2] Outgroup includes either two or more species that are phylogenetically distant from all the species in the sub-clade, or human and other species. In the latter case, human duplicate genes might exist. [3] Available bootstrap values of the corresponding three branches (one between the sub-clade and outgroup, and its two descendants) are all $\geq$900. The gene pairs consistent with all the principles were defined as 'manually curated orthologs', otherwise their annotation status remained to be 'computationally detected orthologs'.

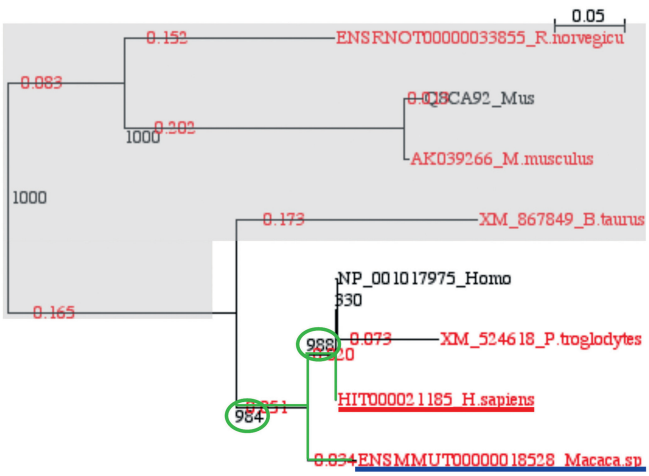

Figure 1. An example of manually curated gene pair from H.sapiens (red underline) and Macaca.sp (blue underline). In this case, conditions of phylogenetic topologies, outgroup species (light gray background) and bootstrap values (two circles) are sufficient (refer to the text). Thus, the pair was defined as a manually curated ortholog.

Table 1. Number of orthologs provided in Evola (release 4.1, June 2007)

Species	Genes	Human genes
Homo sapiens (Human)	18 968	–
Pan troglodytes (Chimpanzee)	16 368	15 615
Macaca sp. (Macaque)[a]	12 037	12 352
Mus musculus (Mouse)	15 570	14 574
Rattus norvegicus (Rat)	15 632	14 302
Canis familiaris (Dog)	14 730	13 916
Bos taurus (Cow)	9375	10 181
Monodelphis domestica (Opossum)	13 201	13 588
Gallus gallus (Chicken)	9266	10 738
Danio rerio (Zebrafish)	12 334	10 468
Tetraodon nigroviridis (Tetraodon)	11 505	9820
Takifugu rubripes (Fugu)	9738	9459

Numbers of genes of both human and other species are listed. Owing to lineage-specific duplication or loss, the numbers are usually different (for example, 15 570 mouse genes are orthologous to 14 574 human genes). 18 968 human genes have at least one ortholog among other 11 species.
[a]*Macaca mulatta*, *Macaca fascicularis*, *Macaca fuscata*, etc. are included.

DATABASE CONTENTS

Evola contains two ortholog datasets: (1) more comprehensive set of orthologs (computational analysis); and (2) more reliable orthologs (computational analysis supported by manual curation). In the current Evola (release 4.1), orthology information for 18 968 human genes is available among 11 vertebrates: chimpanzee, macaque, mouse, rat, dog, cow, opossum, chicken, zebrafish, Tetraodon and Fugu (Table 1). Manually curated orthologs occupied 25.4% of all computationally detected ortholog pairs (24 122/94 935) (release 4.1, 2007).

Evola is a sub-database of H-InvDB (9,10,11), and orthology information in Evola is, as 'Evolutionary annotation', a part of the comprehensive human gene annotations in H-InvDB. Thus, orthology information can be utilized with close reference to other annotation in H-InvDB. For example, 2090 human genes with orthology information belonged to H-Inv protein similarity categories of 'hypothetical proteins' (similarity category IV–VI). Molecular functions of these hypothetical proteins can be analyzed using model species. Moreover, cross references between Evola and other annotations in H-InvDB (protein–protein interaction (PPI), expression, polymorphism, disease, etc.) can produce valuable information contributing to the comprehensive understanding of the human genes.

We aimed to develop user-friendly interfaces that provide easy access to a variety of orthology information in Evola. Users can search orthologs in the top page of Evola as well as in the search systems of H-InvDB [simple search, advanced search and navigation system (Navi)]. Users can download data for each human gene on the main page as well as all the data of Evola in the download page. On the main page of Evola (Figure 2),

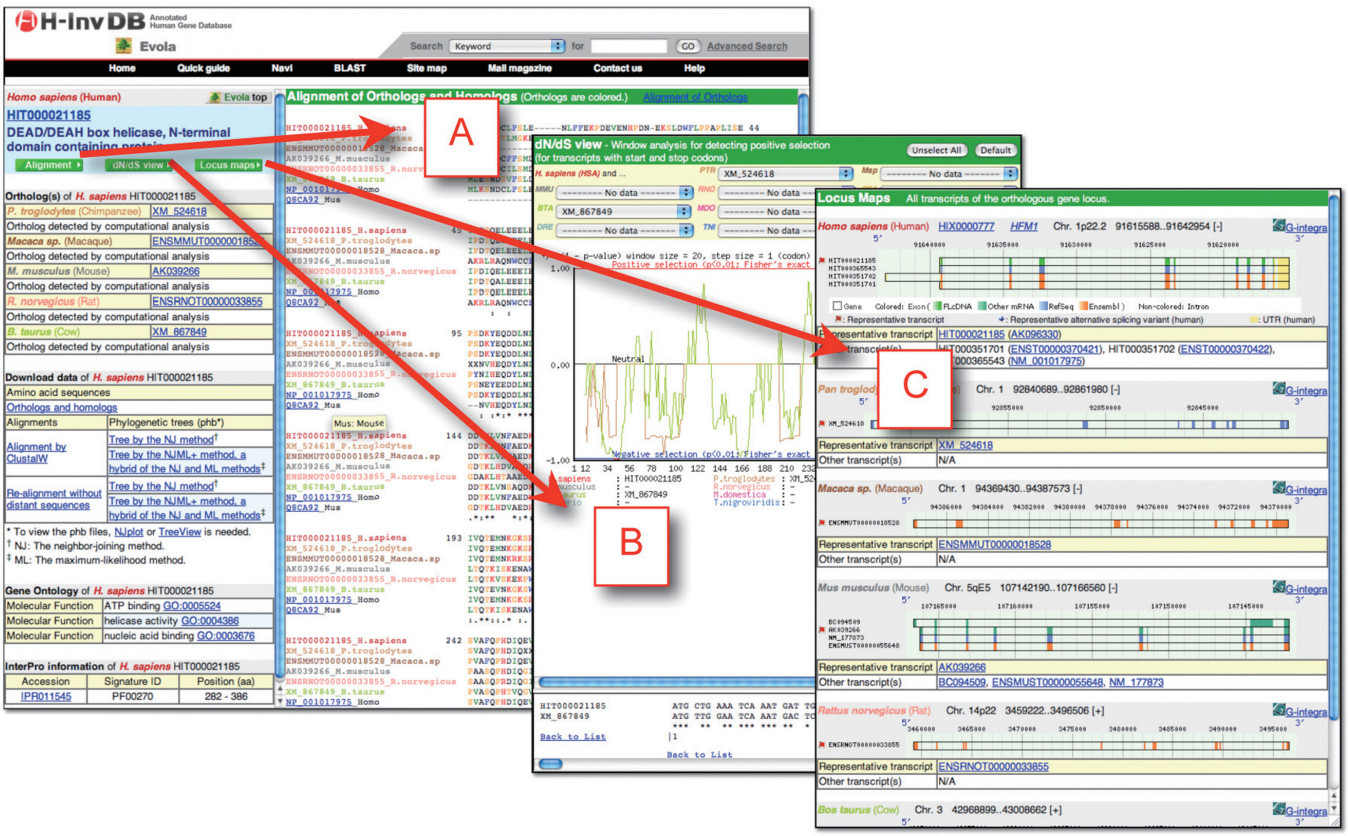

Figure 2. Evola main page. This page is divided into left and right frames. In the left frame, tables of orthologs, download data, Gene ontology, and InterPro are listed. Three green buttons are links to show 'Alignment' (A), 'd_N/d_S view' (B) and 'Locus maps' (C) in the right frame.

the following information for a human gene is available in the left frame: gene name, ortholog list with annotation status, download of sequences, alignments and phylogenetic trees, Gene ontology (22) and InterPro (23). In addition to the set of original ClustalW alignments, another set of alignments, including properly aligned sequences only (24), was also constructed and provided. In the latter sets, sequences with distinctively low identity to other sequences in an alignment were excluded. Based on both alignment sets, phylogenetic trees were constructed by the neighbor-joining method (18) and the NJML+ method (25).

In the right frame of the main page, Evola features the three views described below. Users can switch among the views.

Alignment: Multiple alignments of orthologs and homologs (Figure 2A)

Amino acid sequence alignments of orthologs and homologs are displayed. Users can switch from 'Alignment of Orthologs' (default) to 'Alignment of Orthologs and Homologs', or vice versa. Each amino acid residue is color coded as defined in ClustalX (19). Accession numbers and species names of orthologs (human and other species) are colored in their species colors defined in Evola (human in red, mouse in gray, etc.). Accession numbers of homologs are linked to the original data sources of UniProt or RefSeq. While species are labeled by their scientific names (Homo, Mus, etc.), users can activate a popup window giving a species common name by placing the mouse cursor over homolog accession numbers (for example, 'Q5R508_Pongo'). InterPro data in the left frame include positional information on a human gene, and they can be utilized to detect conserved domains in the proteins.

d_N/d_S view: Window analysis detecting regions under positive or negative selection (Figure 2B)

Users can select one or more species for which to show the plots in the graph. In the lower frame under the graph, the pairwise nucleotide sequence alignment of CDSs is shown. The sequence positions (a.a. or codon) appearing in the graph and alignment are those of human genes.

The nonsynonymous to synonymous substitution rate ratio (d_N/d_S) is a commonly used measure of natural selection. In order to visualize positively and negatively selected regions, sliding window analysis was conducted (a 20 codon window with 1 codon stepping; result for the first window appears as a plot at 11th codon of the human gene). The statistical significance (P-value) of the difference between the number of nonsynonymous substitution (n) per synonymous substitutions (s): n/s, and the number of nonsynonymous sites (N) per synonymous sites (S): N/S was calculated by Fisher's exact test.

d_S, d_N, s and n values were estimated by the modified Nei–Gojobori method (26,27). If $d_N/d_S > 1$, the score ($= 1 - P$-value) was plotted above the zero line (neutral), and if $d_N/d_S < 1$, the score [$= -(1 - P$-value)] was plotted below the zero line. The regions plotted above the red line indicate that the sites might be under positive selection ($d_N/d_S > 1$ and $P < 0.01$). Conversely, the regions plotted below the blue line indicate that the sites might be under negative (purifying) selection ($d_N/d_S < 1$ and $P < 0.01$).

Locus maps: Comparative maps of orthologous gene loci (Figure 2C)

Orthologs were detected for representative transcripts (one transcript per gene locus) in Evola. However, there could be transcript variants in gene loci that have different exon–intron structures leading to produce different protein isoforms. Thus, information on other transcripts besides the representative transcript among orthologous gene loci are shown in Locus maps. In the figures, exon/intron structure, coding sequence (CDS) and untranslated regions (UTR) for each transcript are visualized. H-Inv cluster ID (HIX, an identifier of gene locus), Gene symbol, genomic location and a link to 'G-integra', an integrated genome browser of H-InvDB, are available. The flag icon denotes the representative transcript. The blue diamond icon denotes the Representative Alternative Splicing Variant (RASV) that is another representative per transcript group consisting of the same alternative splicing pattern (28). Representative transcripts are also RASVs, and blue diamonds do not appear if there is only one splicing isoform. In the tables, the H-Inv transcript ID (HIX) and original accession numbers (DDBJ/EMBL/GenBank, Ensembl and RefSeq) of the representative transcript and other transcripts are listed.

FUTURE DIRECTIONS

As our update policy, orthology information in Evola is updated when H-InvDB annotation is updated. One major update and three minor updates per year are scheduled. At the next major update on December 2007, a new duplicate gene family view is planned to be integrated within Evola. Human duplicate gene family data was originally constructed based on both amino acid sequence similarity (29) and orthology information. In the current Evola (release 4.1), parts of human duplicate gene annotation have been already implemented. The human duplicate genes are included in the alignments and phylogenetic trees of orthologs and homologs. Finally, we expect Evola to serve as a new database for evolutionary annotation of human genes. We sincerely welcome any requests and feedback from users.

ACKNOWLEDGEMENTS

We thank the members of Integrated Database Group, Japan Biological Information Research Center for their helpful suggestions. We are also grateful to Craig Gough for critical reading of the manuscript. This work was supported by the Ministry of Economy, Trade and Industry of Japan (METI), and the Japan Biological Informatics Consortium (JBIC). Funding to pay the Open Access publication charges for this article was provided by JBIC.

Conflict of interest statement. None declared.

REFERENCES

1. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631.
2. Wall,D.P., Fraser,H.B. and Hirsh,A.E. (2003) Detecting putative orthologs. *Bioinformatics*, **19**, 1710–1711.
3. Fortna,A., Kim,Y., MacLaren,E., Marshall,K., Hahn,G., Meltesen,L., Brenton,M., Hink,R., Burgers,S., Hernandez-Boussard,T. *et al.* (2004) Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS. Biol.*, **2**, E207.
4. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
5. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, D556–D561.
6. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
7. Dufayard,J.F., Duret,L., Penel,S., Gouy,M., Rechenmann,F. and Perriere,G. (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, **21**, 2596–2603.
8. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Heriche,J.K., Osmotherly,L., Li,R., Liu,T., Zhang,Z., Bolund,L. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, **34**, D572–D580.
9. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS. Biol.*, **2**, e162.
10. Yamasaki,C., Koyanagi,K.O., Fujii,Y., Itoh,T., Barrero,R., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M., Takeda,J., Fukuchi,S. *et al.* (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). *Gene*, **364**, 99–107.
11. Yamasaki,C., Murakami,K., Fujii,Y., Sato,Y., Harada,E., Takeda,J., Taniya,T., Sakate,R., Kikugawa,S., Shimada,M. *et al.* (2008) The H-Invitational Database (H-InvDB), A comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res.*, **36**, in press.
12. Fujii,Y., Itoh,T., Sakate,R., Koyanagi,K.O., Matsuya,A., Habara,T., Yamaguchi,K., Kaneko,Y., Gojobori,T., Imanishi,T. *et al.* (2005) A web tool for comparative genomics: G-compass. *Gene*, **364**, 45–52.
13. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
14. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome. Res.*, **12**, 656–664.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
16. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
17. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.

18. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
19. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
20. Hedges,S.B. and Kumar,S. (2002) Genomics. Vertebrate genomes compared. *Science*, **297**, 1283–1285.
21. Huttley,G.A., Wakefield,M.J. and Easteal,S. (2007) Rates of genome evolution and branching order from whole genome analysis. *Mol. Biol. Evol.*, **24**, 1722–1730.
22. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
23. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L., Copley,R. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.
24. Endo,T., Ogishima,S. and Tanaka,H. (2002) ETools: Tools to Handle Biological Sequences and Alignments for Evolutionary Studies. *Genome Inform.*, **13**, 543–544.
25. Ota,S. and Li,W.H. (2001) NJML+: an extension of the NJML method to handle protein sequence data and computer software implementation. *Mol. Biol. Evol.*, **18**, 1983–1992.
26. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
27. Zhang,J., Rosenberg,H.F. and Nei,M. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. USA*, **95**, 3708–3713.
28. Takeda,J., Suzuki,Y., Nakao,M., Barrero,R.A., Koyanagi,K.O., Jin,L., Motono,C., Hata,H., Isogai,T., Nagai,K. *et al.* (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Res.*, **34**, 3917–3928.
29. Gu,Z., Cavalcanti,A., Chen,F.C., Bouman,P. and Li,W.H. (2002) Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Mol. Biol. Evol.*, **19**, 256–262.