Nucleic Acids Research, 2003, Vol. 31, No. 1 63–67
DOI: 10.1093/nar/gkg106

ParaDB: a tool for paralogy mapping in vertebrate genomes

Magalie Leveugle[1,3,*], Karine Prat[1,4], Nadine Perrier[1], Daniel Birnbaum[1,2,5] and François Coulier[1,2]

[1]Laboratoire d'Oncologie Moléculaire and [2]Atelier de Bioinformatique, Unité 119 INSERM, Marseille, France, [3]Université de la Méditerranée, Marseille, France, [4]Laboratoire de Minéralogie-Cristallographie LMCP, CNRS-UMR 7590C, Paris, France and [5]Laboratoire de Biologie des Tumeurs, Institut Paoli Calmettes, Marseille, France

Received August 21, 2002; Revised and Accepted October 30, 2002

ABSTRACT

We present ParaDB (http://abi.marseille.inserm.fr/paradb/), a new database for large-scale paralogy studies in vertebrate genomes. We intended to collect all information (sequence, mapping and phylogenetic data) needed to map and detect new paralogous regions, previously defined as Paralogons. The AceDB database software was used to generate graphical objects and to organize data. General data were automatically collated from public sources (Ensembl, GadFly and RefSeq). ParaDB provides access to data derived from whole genome sequences (*Homo sapiens*, *Mus musculus* and *Drosophila melanogaster*): cDNA and protein sequences, positional information, bibliographical links. In addition, we provide BLAST results for each protein sequence, InParanoid orthologs and 'In-Paralogs' data, previously established paralogy data, and, to compare vertebrates and *Drosophila*, orthology data.

INTRODUCTION

Large-scale duplications, including polyploïdizations, are thought to have molded early vertebrate evolution (1). It has been hypothesized that two rounds of such large-scale duplications occurred after the divergence of the vertebrates from the cephalochordates (1–6). This hypothesis is sometimes known as the '2R hypothesis' (7,8).

We have defined a paralogon as a series of paralogous regions, within the same species, that could be recognized as deriving from a common ancestor region. Paralogons are thought to be the result of genome-wide duplications that took place early in vertebrate evolution (9) after the separation of chordates non-vertebrates from vertebrates.

The analysis of the human genome sequence confirmed the presence of duplicated genes on large chromosomal segments and the high number of paralogons (8,10,11), but did not evidence a peak of gene families with four members (10–12), probably because the process of duplication is followed by extensive loss of newly duplicated genes.

To precise the term 'paralog', we will call 'ohnologs' (from S. Ohno) the genes derived from the 2R duplications.

By increasing the number of genes, gene duplication provides a fertile ground to functional diversification and acquisition of new characters; duplicated copies undergo sequence modifications and acquire functional specificity, either by taking on a complete new function (neofunctionalization) or by each insuring part of the previous function (subfunctionalization) (13). Depending on the extent in paralog evolution, a certain degree of redundancy may be retained or not. We would learn much on gene evolution and gene function, if we were able to correctly recognize ohnologs and to trace paralog formation and evolution. This may be the only way to build a comprehensive classification of gene families and may have consequences on our understanding of several biological aspects, from gene regulation to phenotypes derived from gene knock-outs.

Correct identification of paralogs and paralogons requires easy access and integration of data of different types (primary sequence, chromosomal localization, sequence similarity, gene phylogeny), and from different sources. The goal of ParaDB is to combine these data under a unique interface.

ParaDB DESIGN AND IMPLEMENTATION

We used AceDB (J. Thierry-Mieg and R. Durbin, 1996. http://www.acedb.org) version 4.9c on Linux Mandrake 8.0 as database software and CGI.pm Perl package and AceBrowser version 2.1 from Lincoln Stein (http://stein.cshl.org/AcePerl/AceBrowser) to build the ParaDB web interface.

ParaDB was constructed in three distinct but parallel steps: model design, data selection and data formatting.

Model design

AceDB is an object-oriented system; the object structure is defined by the object classes in the models.wrm file. This file

*To whom correspondence should be addressed at INSERM Unité 119, 27 bd Leï Roure, 13009 Marseille, France. Tel: +33 491758423; Fax: +33 491260364; Email: leveugle@marseille.inserm.fr

provides exact structure allowed for each type of object and nature of data displayed (Text, Float, Links). Graphical objects such as map and tree were taken and modified from the AceDB model file. Text-based objects (tree objects) were created for ParaDB. All object types have interactive links with each other's. The model file as well as Perl scripts and Ace files are available from ParaDB web page at: http://abi.marseille.inserm.fr/paradb/.

The central object class is Gene_name, which contains general data about the gene of interest. For example there are definition, bibliography, links to other databases through the web (OMIM, LocusLink and MEDLINE). From this object class, users can access all other data type such as sequences, maps, similarity results and paralogy mapping.

Data selection

To build the March 2002 ParaDB release, we used proteome, transcriptome and bibliography data from four main source databases: (i) Ensembl (14) human 3–26 and mouse 4.1 sequences, (ii) GadFly 2.1 (15) *Drosophila melanogaster* sequences, (iii) HUGO (16), LocusLink and RefSeq (17), and (iv) SWISS-PROT (18) human and mouse complementary data. cDNA and protein sequences were downloaded as fasta files from which we extracted sequences and positional information. RefSeq human and mouse files (GenBank format) provided bibliographical information, links to other databases (OMIM, LocusLink and MEDLINE), and definition for the genetic element described. Known human genes were named according to HUGO or alternatively LocusLink/RefSeq databases, and mouse genes according to SWISS-PROT and LocusLink. A species suffix was added to fly and mouse gene names (_Dme for *D. melanogaster* and _Mmu for *Mus musculus*).

Data formatting. Data files were retrieved from these databases and formatted with Perl 5.6.1 scripts. We used Perl regular expressions to extract and to format data from various file formats (GenBank, fasta and spreadsheets) and to rewrite them automatically in .ace file format, according to the ParaDB models. These files were automatically integrated in the database using the tace AceDB interface in a csh script from http://greengenes.cit.cornell.edu/acedoc/aceloading.html.

Bibliographical references were given a unique 14-character-long key based on the journal and title lines of the corresponding RefSeq entry, which contain the publication year in the last four characters.

DATA

All paralogy and orthology data were added from previous studies done in our laboratory (9,19–21).

Similarity data

We provided BLAST results for each protein in the database. All proteins from each species (human, mouse and fly) were queried chromosome by chromosome against sequences of all species (with the exclusion of the pairs mouse/fly and fly/fly).

Table 1. Blast results for paralogous proteins in the database

Result number	Protein >4 paralogs groups (16)		4 paralogs groups (63)		3 paralogs groups (244)		2 paralogs groups (454)	
0	27	21.95%	84	23.01%	238	26.10%	312	30.03%
1	6	4.88%	31	8.49%	131	14.36%	323	31.09%
2	8	6.50%	47	12.88%	197	21.60%	125	12.03%
3	8	6.50%	59	16.16%	90	9.87%	69	6.64%
4	10	8.13%	33	9.04%	97	10.64%	50	4.81%
>4	64	52.03%	133	36.44%	159	17.43%	160	15.40%
Total	96	78.05%	281	76.99%	674	73.90%	727	69.97%
Total protein	123	100%	365	100%	912	100%	1039	100%

For each type of group (2, 3, 4 or more paralogs), this table gives the repartition of proteins based on result numbers after BLAST-filtering by the I′ criterion. Families with more than 4 paralogous genes (ohnologs) are due to *cis*-duplications that have occurred after the 2R duplications. For example, DUSP paralogy group contains 5 genes but there are two recently duplicated genes on chromosome 2 (DUSP2 and DUSP11).

We used pgp BLAST (22) with default parameters (Matrix Blosum 62 and $T = 11$). BLAST outputs were collected and formatted after filtering by Li *et al.* (12) I′ criterion. The Li criterion is based on length of alignment and protein and on percentage of identities in this alignment. It allows selection of alignments with a sufficient length and a sufficient score to avoid local similarity between the compared proteins. We choose stringent I′ (30 for human/human and mouse/mouse BLAST, 25 for human/mouse, and 20 for human/*Drosophila*) to select sequences corresponding to potential paralog/ortholog. Self-matching of query proteins were eliminated from the given results.

Table 1 shows the repartition of the results per protein following the size of the paralogy group. For example, a protein belonging to a 4-member group was expected to match three similar proteins or more. In fact, only 61.64% of these proteins obtained 3 results or more and 23.01% obtained no result. These numbers were found for all types of group, with an average percentage of proteins with no result around 27.

Only half of human proteins encoded by a gene belonging to a paralogy group matches at least the same number of proteins as its number of known paralogs. This result is explained by the high I′ threshold we chose to select results. We wanted to avoid similarity due to protein domains, which is not representative of homologous proteins and can be obtained by convergent evolution and domain shuffling between genes. High selection threshold allowed eliminating most of the false positives but certainly increased the number of false negatives. Two homologous genes can have diverged because of lower selective pressure on one of the duplicate. In this way, a high I′ could not allow the selection of alignment between the two sequences.

The BLAST results were added as complementary information to help identify potential paralog/ortholog in the Ensembl predicted sequences.

An integrated BLAST interface from the National Center for Biotechnology Information (ftp://ftp.ncbi.nih.gov/blast/server/current_release/readme.html) was also added, to allow users to

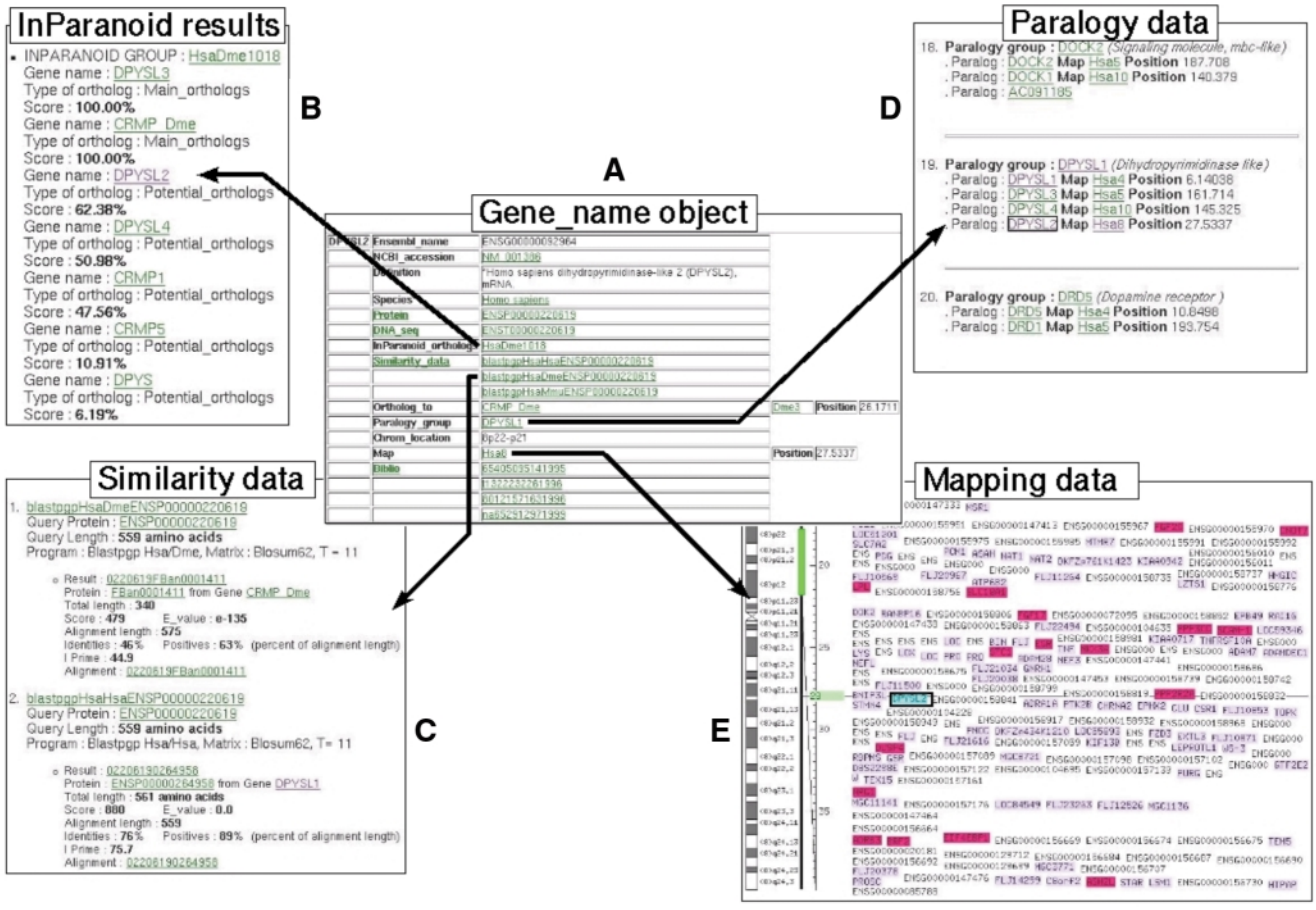

Figure 1. Example of a ParaDB query Organization of data through the ParaDB web interface. The Gene_name window (**A**) displays general data about the DPYSL2 gene and gives links to other types of data by clicking on the names of these objects. For example: InParanoid results window (**B**) shows InParanoid orthologs and In-Paralogs detected by local InParanoid performed on all the species of ParaDB. The Similarity data (**C**) shows results of BLAST pgp of DPYSL2 protein against all the others proteins in the database. Resulting alignments are displayed in another window by clicking on the link. Paralogy data window (**D**) shows sample of organization of paralogous genes in the MetaHOX paralogon. Mapping data (**E**) give an overview of the genetic neighborhood of DPYSL2 in human chromosome 8: genes with purple background are known genes with RefSeq identifier and genes with pink background belong to a paralogy group.

BLAST a cDNA or protein sequence from the database against a collection of sequence databases and filter their results with various I′ values.

InParanoid data

The three species in ParaDB were compared pairwise using the 2.3 version of InParanoid (23) program which automatically detects orthologs (or groups of orthologs) from two species. A Perl script was written to: (i) automatically download the protein dataset (Ensembl 3–26 human, Ensembl 4.1 mouse and GadFly 2.1), (ii) run InParanoid, (iii) extract data from the html resulting files, and (iv) write .ace files containing each group of orthologs linked to the corresponding Gene_name objects under the tag InParanoid_orthologs.

In addition, a local InParanoid interface was created, including data from four species (*Homo sapiens*, *M. musculus*, *D. melanogaster* and *Danio rerio*), to allow users to search orthologs of a protein by gene name, species, query word or accession number.

ParaDB INTERFACES

Browsing example

Figure 1 shows an example of ParaDB session with the web Browser interface. The DPYSL2 (dihydropyrimidinase-like 2) gene_name object, a collapsin mediator response protein (see Definition line), illustrates this session. DPYSL2 was predicted by the Ensembl genome project 3–26 release under the ENSG00000092964 gene identifier. According to the correspondence file provided by the Ensembl Helpdesk, we used the corresponding HUGO name (DPYSL2) and RefSeq accession number (NM_001386). The Gene_name window displays general data about this gene: definition, protein and cDNA identifiers, species, paralogy_group, chromosomal location, similarity_data and InParanoid group of orthologs. All underlined fields are linked to other objects.

Mapping data The map object shows a zoomed view of the human chromosome 8 region that contains DPYSL2. Maps provide a general view of paralogs (ohnologs) distribution

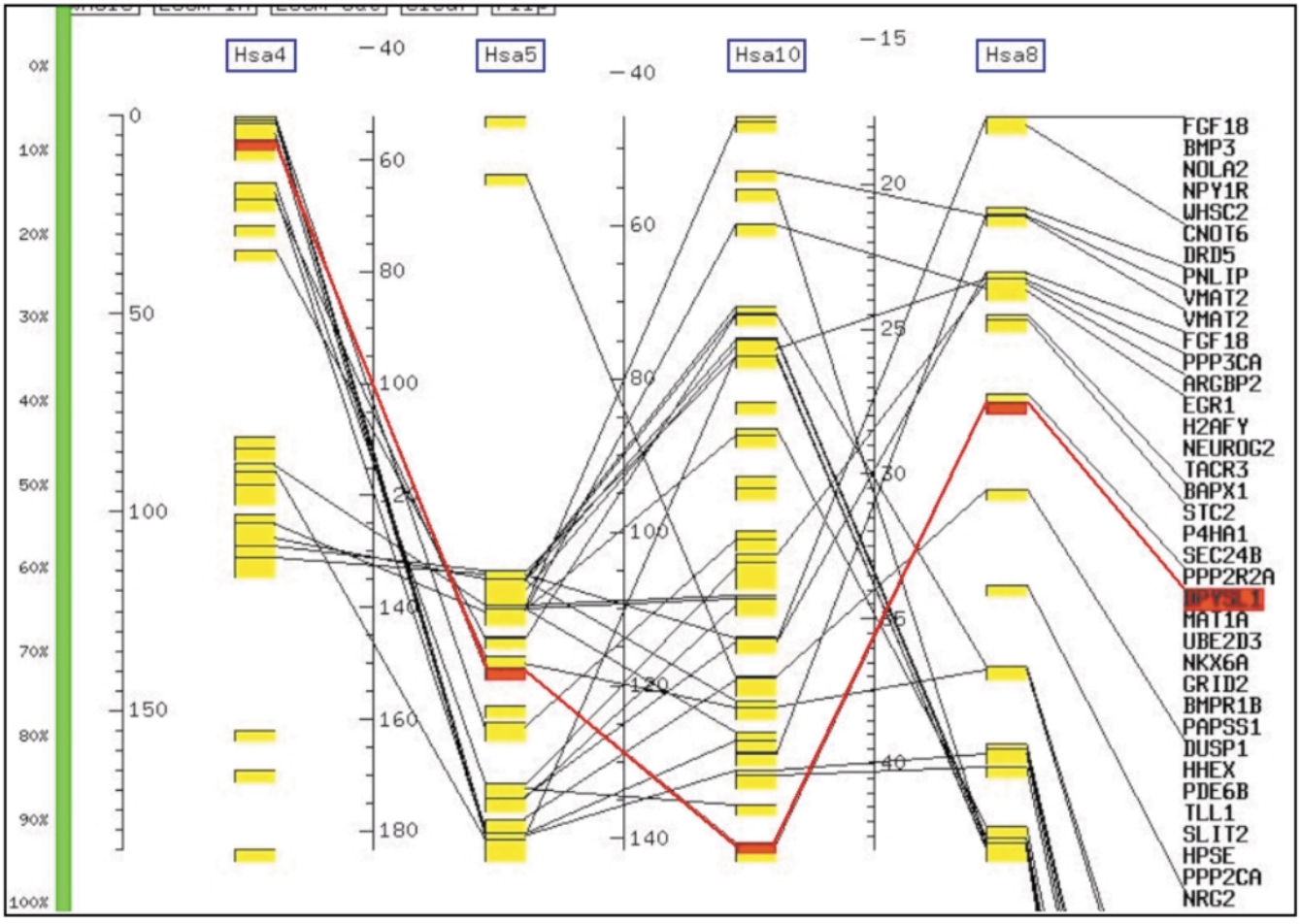

Figure 2. 4/5q/10q Paralogon MultiMap. Paralogous gene organisation in the 4/5q/10q Paralogon which includes the DPYSL2 gene. Section of chromosomes 4, 5, 8 and 10 are figured by graduated scales, paralogs by linked yellow boxes and the active paralogy group, DPYSL1, by red boxes. This figure was obtained with the xace Table Maker and only screen captures are available through the web interface. Interactive MultiMaps should be available in a future release of ParaDB.

and thus of organization of the paralogon on the chromosome. Every gene in the chromosome can be accessed by clicking on its name. Maps are also available for fly and mouse genes (not shown).

Similarity data. Similarity data give results of BLAST pgp for the protein encoded by DPYSL2 against all the human, mouse and fly proteins. In this example, we obtained 14 proteins: CRMP, the *Drosophila* ortholog of DPYSL2; DPY1, 2, 3 and 4, DPYS and CRMP5 mouse proteins (not shown); DPYSL1, 3 and 4 (not shown), encoded by human ohnologs of this gene; CRMP3 and 5 human proteins, and DPYS, the human dihydropyrimidinase, whose gene is also located on chromosome 8 (not shown).

InParanoid orthologs. This window displays InParanoid predicted orthologs in fly and In-Paralogs for DPYSL2. The two main orthologs have a 100% score and the In-Paralogs, such as DPYSL2, are scored according to their similarity with the main ortholog from the same species.

Paralogy and orthology data. DPYSL2 and its paralogous genes belong to the 4/5q/10q (MetaHOX) paralogon (12), as shown in the paralogy data window. This window displays all the 85 paralogy groups in the paralogon, with mapping data for each paralog. Here only 3 paralogy groups are shown. From this object, users have access to a graphical and static view of the paralogon organization, called MultiMap (Fig. 2).

The ortholog_to line (in Gene_name window) gives a link to CRMP, the *Drosophila* DPYSL2 ortholog and indicates the location of this gene on *Drosophila* chromosome 3.

ParaDB query builder

The ParaDB web interface allows five query types; four are included in the original AceDB software (Simple search, Text search, Class Browser and Ace Query), but the fifth (Query Builder) has been rewritten to work with Ace Browser. The first three types are keyword-based searches, which are convenient for simple queries. The Ace Query Language was created to formulate complex queries based on several criteria, but users need to learn a specific syntax and to know the structure of each object model. The Query builder is a step-by-step graphic

interface to formulate Ace queries. It was available in a previous release of the web interface (webace) and in the xace (X-Window) interface, but not in the new Ace Browser web interface. Therefore, we rewrote this interface for Ace Browser, using Lincoln Stein AcePerl and CGI.pm Perl packages.

CONCLUSION

Few databases are exclusively dedicated to the description of paralogy relationships, such as Ken Wolfe's interactive view of human paralogons, based on Ensembl sequences (8). ParaDB includes, in addition to known paralogy information, gene and map data from three species, orthology relationships between human and fly, and in a future release, orthology and paralogy data from mouse sequences will be included. One of the major advantage of ParaDB is the availability, for each gene, of similarity results obtained from two different methods, i.e. filtered BLAST and InParanoid, chromosome localization in clickable maps, possibility to run filtered BLAST and to ask five different types of query (i.e. Simple search, Text search, Class Browser, Ace Query and Query Builder).

Identification and mapping of paralogous genes and regions is an important step in understanding the evolution of the vertebrate genomes. This mapping depends on accessing to various types of data. Ideally, all these data should be integrated into a convenient interactive tool, such as ParaDB. In ParaDB, users can navigate through graphical representations of relationships between genes and all centralized needed data.

Future enhancements of ParaDB should include mouse orthology data, as the Mouse Genome Sequencing Consortium is completing the annotation of the mouse genome sequence, dynamic MultiMap and inclusion of phylogenetic trees. Data from other species will also be included, as ParaDB models make it simple to add them as they become available.

ACKNOWLEDGEMENTS

We thank Françoise Birg and Claude Mawas for enthusiastic support, Esther Schmidt and Pascal Hingamp for their help with Ensembl data, and Jean Thierry-Mieg and Ed Griffith for their assistance in setting up our AceDB server. This work was supported by INSERM and by the Ligue Nationale Contre le Cancer (Équipe Labellisée). M.L. was supported by a fellowship from the Ministère de la Recherche. This work constitutes part of a thesis to be submitted by M.L. to the Université de la Méditerranée to obtain a PhD degree.

REFERENCES

1. Ohno,S. (1970) *Evolution by Gene Duplication*. Heidelberg, Springer Verlag, New York, NY.
2. Schughart,K., Kappen,C. and Ruddle,F.H. (1989) Duplication of large genomic regions during the evolution of vertebrate homeobox genes. *Proc. Natl Acad. Sci. USA*, **86**, 7067–7071.
3. Lundin,L.G. (1993) Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics*, **16**, 1–19.
4. Holland,P.W., Garcia-Fernandez,J., Williams,N.A. and Sidow,A. (1994) Gene duplications and the origins of vertebrate development. *Dev. Suppl.*, 125–133.
5. Spring,J. (1997) Vertebrate evolution by interspecific hybridisation—are we polyploid? *FEBS Lett.*, **400**, 2–8.
6. Robinson-Rechavi,M., Marchand,O., Escriva,H., Bardet,P.L., Zelus,D., Hughes,S. and Laudet,V. (2001) Euteleost fish genomes are characterized by expansion of gene families. *Genome Res.*, **11**, 781–788.
7. Wolfe,K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nature Rev. Genet.*, **2**, 333–341.
8. McLysaght,A., Hokamp,K. and Wolfe,K.H. (2002) Extensive genomic duplication during early chordate evolution. *Nature Genet.*, **31**, 200–204.
9. Popovici,C., Leveugle,M., Birnbaum,D. and Coulier,F. (2001) Homeobox gene clusters and the human paralogy map. *FEBS Lett.*, **491**, 237–242.
10. Lander,E.S. and the International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
11. Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
12. Li,W.H., Gu,Z., Wang,H. and Nekrutenko,A. (2001) Evolutionary analysis of the human genome. *Nature*, **409**, 847–849.
13. Lynch,M. and Force,A. (1999) The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**, 459–473.
14. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
15. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D. *et al.* (2000) The Genome Sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
16. Povey,S. (2002) Guidelines for Human Gene Nomenclature. *Genomics*, **79**, 463–463.
17. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
18. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL. *Nucleic Acids Res.*, **28**, 45–48.
19. Coulier,F., Burtey,S., Chaffanet,M., Birg,F. and Birnbaum,D. (2000) Ancestrally-duplicated paraHOX gene clusters in humans. *Int. J. Oncol.*, **17**, 439–444.
20. Coulier,F., Popovici,C., Villet,R. and Birnbaum,D. (2000) MetaHox gene clusters. *J. Exp. Zool.*, **288**, 345–351.
21. Popovici,C., Leveugle,M., Birnbaum,D. and Coulier,F. (2001) Coparalogy: physical and functional clusterings in the human genome. *Biochem. Biophys. Res. Commun.*, **288**, 362–370.
22. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
23. Remm,M., Storm,C.E.V. and Sonnhammer,E.L.L. (2001) Automatic clustering of Orthologs and In-Paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.