HUMAN MUTATION 24:112–119 (2004)

DATABASES

OrthoDisease: A Database of Human Disease Orthologs

Kevin P. O'Brien,* Isabelle Westerlund, and Erik L.L. Sonnhammer

Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden

Communicated by A. Jamie Cuticchia

One of the greatest promises of genome sequencing projects is to further the understanding of human diseases and to develop new therapies. Model organism genomes have been sequenced in parallel to human genomes to provide effective tools for the investigation of human gene function. Many of their genes share a common ancestry and function with human genes, and this is particularly true for orthologous genes. Here we present OrthoDisease, a comprehensive database of model organism genes that are orthologous to human disease genes. OrthoDisease was constructed by applying the Inparanoid ortholog detection algorithm to disease genes derived from the Online Mendelian Inheritance in Man database (OMIM). Pairwise whole genome/proteome comparisons between *Homo sapiens* and six other organisms were performed to identify ortholog clusters. OMIM numbers were extracted from the OMIM Morbid Map and were converted to gene sequences using the Locuslink mim2loc and loc2acc tables. These were mapped to Inparanoid ortholog clusters using Blast. The number of ortholog clusters in OrthoDisease with each respective species is currently: *M. musculus*, 1,354; *D. melanogaster*, 724; *C. elegans*, 533; *A. thaliana*, 398; *S. cerevisiae*, 290; and *E. coli*, 153. The database is accessible online at http://orthodisease.cgb.ki.se, and can be searched with disease or protein names. The web interface presents all ortholog clusters that include a selected disease gene. A capability to download the entire dataset is also provided. Hum Mutat 24:112–119, 2004. © 2004 Wiley-Liss, Inc.

KEY WORDS: comparative genomics; protein; Mendelian; evolution; genetic disease

DATABASES:
http://orthodisease.cgb.ki.se (OrthoDisease); http://inparanoid.cgb.ki.se (Inparanoid)
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM (OMIM)

INTRODUCTION

All genomes present in the modern biosphere have evolved from common ancestral genomes. Comparative genomics has enabled us to identify and study commonly derived processes which occur both in humans and model organisms. Modeling human genetic diseases in these organisms requires the correct assignment of orthology. Orthologs are described as genes in divergent species that have originated from a single gene in the last shared ancestor, i.e., homology following speciation [Fitch, 1970]. Homologs that arise from gene duplications are termed paralogs, a term often applied to homologs within a genome. However, paralogy can exist between genes in different species, since gene duplication events occur before and after speciation. To distinguish between these types of paralogs, the terms "inparalogs" and "outparalogs" have been adopted; inparalogs indicate paralogs that arise through a gene duplication event following speciation, while outparalogs arise following a gene duplication preceding speciation (Fig. 1) [Remm et al., 2001; Sonnhammer and Koonin, 2002]. In a simple scenario, paralogs that have separate orthologs in another species are outparalogs; however, if they shared the same ortholog in the other species they would be inparalogs (co-orthologs). Therefore, identification of inparalog clusters (expected to contain functionally similar proteins) by excluding outparalogs (likely to be more functionally diverged) becomes the primary objective of orthology assignment.

Ortholog identification is a process that ought to be performed taking the entire genomic content of each organism into consideration. The vast computational resources required to construct phylogenetic trees from whole genome sequence alignment renders this approach unfeasible for even a modestly sized genome. Using the reciprocally best hits between species from pairwise gene comparisons (for example using Blast) has been done in the past, but this is often not sufficient to determine all orthologs and may, in fact, be misleading [Chervitz et al.,

Received 2 June 2003; accepted revised manuscript 26 March 2004.

*Correspondence to: Kevin P. O'Brien; Center for Genomics and Bioinformatics, Karolinska Institutet, S-171 77 Stockholm, Sweden. E-mail: kevobr@mbox.ki.se

Grant sponsors: Swedish Research Council; Karolinska Institutet; Pfizer Corporation.

DOI 10.1002/humu.20068
Published online in Wiley InterScience (www.interscience.wiley.com).

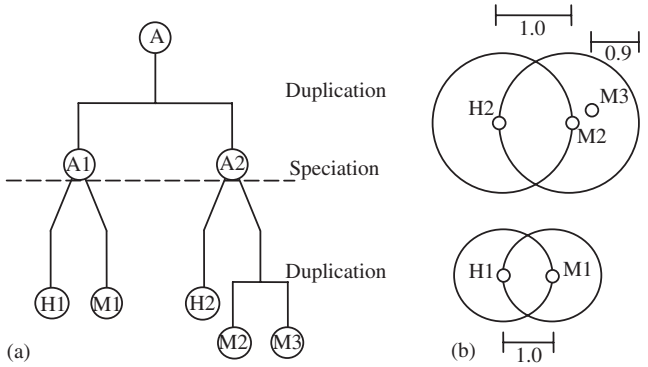

FIGURE 1. **A hypothetical gene tree to illustrate the relationships leading to inparalog (co-ortholog) and outparalog assignments. a: A protein "A" in an ancestral species is duplicated prior to the speciation event that forms the mouse (M) and Human (H) lineages. In the mouse genome, M2 and M3 are inparalogs, since their duplication occurred postspeciation; they are thus co-orthologs to the human H2 (one common ancestral protein upon speciation). M1 is an outparalog of M2 and M3, as H1 is of H2 (duplication and divergence prior to speciation). b: A graphical representation of Inparanoid clustering and hypothetical scoring of the tree shown in (a). *H2* and *M2* are the original seed-ortholog pair, thus both received an inparalog score of 1.0. Other inparalogs (in this case *M3*) are scored according to their relative similarity to the seed-inparalog (here *M2*). The value is calculated for *M3* as follows: inparalog score = (Blast[*M2*:*M3*] – Blast[*M2*:*H2*])/(Blast[*M2*:*M2*] – Blast[*M2*:*H2*]), where Blast[X:Y] is the averaged similarity score between X and Y in bits after Blast comparison. In this case, *M2* is relatively more similar to *H2* than *M3* is, and thus *M3* receives a lower inparalog score (0.9). *M1* and *H1* are orthologous to each other, but are outparalogs of the other cluster, and thus form a cluster of their own.**

1998; Mushegian et al., 1998; Remm et al., 2001; Rubin et al., 2000]. The true ortholog may not be a single gene, but instead a cluster of inparalogs [Nembaware et al., 2002; Storm and Sonnhammer, 2002; Xie and Ding, 2000]. The Cluster of Orthologous Groups database (COG) is built upon all-against-all Blast approach, which is performed on completed genomes [Tatusov et al., 2000]. Orthologous protein clusters are assembled through a triangular homology to other evolutionary distant species. The Eukaryotic Gene Orthologs database (EGO) is an application of this approach using The Institute for Genomic Research (TIGR) protein dataset (www.tigr.org/tdb/tgi/ego). The COG approach appears to function quite well when an obvious ortholog is present in a completed genome, but can be confounded by the presence of multiple outparalogs and report these as orthologous proteins [Li et al., 2003]. A further limitation of this dataset is its application to incomplete genomes, in which detected orthologs should be interpreted with extreme caution. OrthoMCL is a tool which uses a Markov clustering approach following all-against-all Blast step to cluster proteins in orthologous groups while differentiating "recent" and "ancient" paralogs, which can be interpreted as inparalogs and outparalogs, respectively [Li et al., 2003]. The Inparanoid program was developed to address the need to identify orthologs while differentiating between inparalogs and outparalogs [Remm et al., 2001]. Inparanoid is freely available (http://inparanoid.cgb.ki.se), uses NCBI Blast, and can be run locally in a relatively short time, even for complete eukaryotic genomes. Therefore, it was chosen as the ortholog identification tool for the current study.

The Online Inheritance In Man database (OMIM; www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) is a manually curated catalog of human diseases and phenotypes [Hamosh et al., 2002]. It is also a resource of genes that have been implicated or shown to be mutated in these heritable diseases and phenotypes. The generation of disease models in other organisms is one of the most valuable resources in medical research, permitting the testing of novel therapeutic strategies, and aiding the understanding of disease pathogenesis. One of the critical steps in this regard is the proper identification of orthologs. Most previous attempts to collect disease gene orthologs have been limited to manually-created disease information with simple best-best Blast matches [Aboobaker and Blaxter, 2000; Ahringer, 1997], e.g., the analysis of 289 human disease gene homologs in fly, worm, and yeast [Rubin et al., 2000]. A larger scale effort was the Homophila database (http://homophila.sdsc.edu), which was generated by obtaining 911 disease genes from the OMIM Morbid Map and searching for *Drosophila melanogaster* homologs using Blast [Chien et al., 2002; Reiter et al., 2001]. The first multigenome ortholog database of disease genes was the EGO database (www.tigr.org/tdb/tgi/ego/human_dis_gene.shtml), which derived its disease information from the OMIM Gene Map. This approach uses a COG-based approach on complete and incomplete genomes, which, as mentioned above, may result in reporting outparalogs as orthologs.

Here we announce the release of a searchable online database called OrthoDisease. This database was constructed by first extracting disease/gene relationships from the OMIM Morbid Map (a list of diseases with associated genome locations), and then associating these genes with Inparanoid clusters calculated from complete genomic protein sets. The dataset can be searched using a gene or disease as a starting point or can be downloaded in its entirety as a flat file.

DATABASE CONSTRUCTION

The Inparanoid Tool

In the first step, Inparanoid runs whole-genome comparisons between the genomes and within each genome. By demanding that matches involve at least 50% of either sequence, a local-alignment program such as Blast can be used in a "pseudoglobal" mode. This pairwise comparison step is followed by a rule-based clustering step that gathers inparalogs around a central ortholog seed-pair. The underlying model in this clustering method is that for any given protein, one would expect its ortholog to show the highest similarity in a whole genome comparison, while any inparalogs of this ortholog that may have arisen postspeciation would be the next-best hit. Furthermore, inparalogs are expected to show higher similarity to each other than to the protein they are co-orthologous to; using the

example shown in Figure 1, M2 and M3 are more similar to each other than to H2, their ortholog, and thus all three will form an ortholog cluster. In addition, a protein would be expected to show a higher degree of similarity to its ortholog than its outparalog. Since by definition, orthologs share a common ancestor upon speciation, whereas outparalogs do not, H1 should be more similar to M1 than to H2 and will thus be excluded from the H2 cluster to form a cluster of its own with M1 (Fig. 1b). An Inparanoid cluster is seeded by a reciprocally best-matching ortholog pair, around which inparalogs (should they exist) are gathered independently, while outparalogs are excluded. Here, seed-ortholog pair refers to the two seed members that are orthologous to each other, around which their inparalogs are clustered. Each is referred to the seed-inparalog when comparing against inparalogs in its own genome. Each member of the cluster receives an inparalog score, which reflects the relative distance to the seed-inparalog (1.0=identical to the seed-inparalog; 0.0=of equal distance to the seed-inparalog as the distance between the seed-ortholog pair). The confidence that the original seed-ortholog pair are true orthologs is estimated by sampling how often the pair is found as reciprocally best matches by a bootstrapping procedure. Bootstrap values were generated by counting how many times the seed-pair genes were each other's best match in a "sampling with replacement" procedure that was applied to the original Blast alignment [Remm et al., 2001]. In summary, an Inparanoid ortholog cluster contains a seed-ortholog pair with bootstrap confidence values, and a list of inparalogs with inparalog scores. The accuracy of Inparanoid has previously been tested by comparison to a curated dataset of worm and mammalian transmembrane proteins that were generated by manual tree-based orthology analyses. Inparanoid generated comparable clusters, but often split up larger clusters. A false-positive and false-negative level of 5% or less was seen for this dataset [Remm et al., 2001]. In a more recent study, Inparanoid was compared to OrthoMCL, and both performed equally well on the same dataset [Li et al., 2003].

The construction of the Inparanoid dataset was performed as follows: The complete protein sequence database from SWISSPROT and trEMBL was obtained from ftp://ftp.ebi.ac.uk/pub/databases/sp_tr_nrd [Boeckmann et al., 2003; Wheeler et al., 2004]. Relevant organisms were extracted according to the taxonomy ID number and converted to multiple-fasta format using the SWISS module of the BIOPERL package [Stajich et al., 2002]. In a step aimed at reducing the redundancy observed between trEMBL and Swissprot, due to ongoing submissions to EMBL, all proteins with 100% matches to other proteins over the full length were removed, with preference show to Swissprot entries. trEMBL proteins with 99% matches were removed only if the protein name matched a Swissprot entry or were annotated as containing partial sequences. The dataset size for each organism used in this study following this processing step was: *Arabidopsis thaliana* (34,170), *Escherichia coli* (8,901), *Caenorhabditis elegans* (20,627), *Homo sapiens* (36,379), *Saccharomyces cerevisiae* (6,706), *Mus musculus* (34,499), and *Drosophila melanogaster* (18,932). Whole genome NCBI Blast [Altschul et al., 1997] comparisons using these reduced datasets were then performed between human and each species. The Blast output (human → organism, organism → human, human → human, and organism → organism) was used as the input for the Inparanoid program [Remm et al., 2001].

Disease, phenotype, and gene OMIM numbers were obtained from the OMIM Morbid and Gene Maps (ftp://ftp.ncbi.nih.gov/repository/OMIM/) [Hamosh et al., 2002]. The OMIM Morbid Map is a flat file containing all phenotypes or diseases in OMIM with a known cytogenetic location. The OMIM Gene Map, on the other hand, is a list of all genes with known cytogenetic locations mentioned in OMIM, but, unlike the Morbid Map, it contains genes that may have no proven link to a human disease. OMIM numbers obtained from these flat files were used to obtain LocusLink gene locations in the mim2loc table, which were, in turn, used to get protein accession numbers from the LocusLink loc2acc table (ftp://ftp.ncbi.nih.gov/refseq/LocusLink/) [Maglott et al., 2000; Pruitt et al., 2000]. Thus, where information from these three tables permitted, a list of redundant GenBank entries was obtained for proteins with known alterations in the OMIM diseases. Each protein sequence obtained was compared to the reduced Swissprot/trEMBL dataset, using Blast to identify the longest high-scoring match in that dataset, the result of which was processed using MSPcrunch (with the -d option) [Sonnhammer and Durbin, 1994]. Only matches above 99% identity were considered. This nonredundant list of proteins was then used to parse Inparanoid clusters from results generated above. For each model organism, three types of text files were the final output of this process: one containing OMIM number, protein symbol, protein name, associated disease(s), and gene aliases; another containing parsed Inparanoid clusters; and a third flat file containing the entire OrthoDisease dataset for that organism. Access to the disease-associated clusters on the web server is handled through Common Gateway Interface (CGI) scripts written in Perl. The corresponding Gene Ontology (GO) entries were extracted from Swissprot entries where available [Harris et al., 2004].

RESULTS

OrthoDisease Online

The complete dataset of Inparanoid clusters and associated disease information was made available through an online searchable tool named OrthoDisease (http://orthodisease.cgb.ki.se/). Querying a disease or OMIM number using "Disease search" returns a list of Morbid Map–derived diseases that are linked to pages listing all human genes associated with that disease/phenotype in OMIM. Querying a protein/gene name using "Protein search" returns a similar listing. "Free Text Search" combines both of these searches and returns a protein list derived from the Morbid Map. Where applicable, these lists of proteins are clickable links to

the corresponding Inparanoid clusters from the various species present in OrthoDisease. Full information concerning the selected gene is displayed, along with disease-association, OMIM numbers, and Inparanoid clusters. Clusters are shown for each organism separately, and can be restricted to selected organisms. For each protein in the cluster, the inparalog score is displayed, and external links are provided. These include Pfam; a database of protein domain families that can be used to view the domain structures or to find other related proteins at the domain level (http://Pfam.cgb.ki.se/) [Bateman et al., 2004]. The full sequence of the protein can be obtained, as well as the full ExPASy descriptions of each protein (www.expasy.ch/) [Gasteiger et al., 2003].

Each ortholog group in OrthoDisease is made up of a seed-ortholog pair and, potentially, a set of inparalogs. For each seed-ortholog pair in OrthoDisease, an Inparanoid bootstrap value is provided that indicates the level of support for the seed-pair to be orthologs. For inparalogs, an inparalog score is provided that indicates the level of support for the gene to be an inparalog (see Fig. 1b). The seed-ortholog pair receives an inparalog score of 1.0 by definition, since this is the orthologous pair around which all inparalogs are clustered. To recap its meaning, inparalogs are genes which have resulted from a gene-duplication *after* the two species being compared have diverged [Remm et al., 2001; Sonnhammer and Koonin, 2002]. Thus, inparalog status is dependent on the species compared. For instance, a disease gene may have a large cluster of inparalogs in human when compared to *C. elegans*, but none when compared to mouse.

In addition to genes from the OMIM Morbid Map, OrthoDisease contains an alternative resource derived from the OMIM Gene Map that contains genes described in OMIM but without necessarily being mutated in any disorder. This table contains both disease gene candidates and genes that may have already been investigated for involvement in a particular disease. The Gene Map was processed in a similar way to the OMIM Morbid Map, and can also be searched independently using the "OMIM Gene Map search" in OrthoDisease. It consists of a larger dataset that does not overlap with the Morbid Map dataset. Through the same process as for the Morbid Map, 2,259 genes from the Gene Map were analyzed, and the number of ortholog clusters derived from the Gene Map per species is currently: *M. musculus*, 1,316; *D. melanogaster*, 1,114; *C. elegans*, 782; *A. thaliana*, 581; *S. cerevisiae*, 472; and *E. coli*, 161. Searching the Gene Map can be useful when a gene of interest is a strong disease gene candidate, but has not been definitively proven to be involved in the disease, and is therefore absent from the Morbid Map. For example, the Alpha helical coiled–coil rod protein HCR is not mutated in any Mendelian disease, but is included in the OMIM Gene Map due to its strong candidature in the PSORS1 locus on chromosome 6, where the allele HCR*WWCC is strongly associated with psoriasis susceptibility, a complex genetic disease [Asumalahti et al., 2002; O'Brien et al., 2001]. OrthoDisease also contains a section "Table Generator" where one can download the entire dataset for each organism. This dataset can be derived either from OMIM's Morbid or Gene Map. The output, which can be saved locally as a tab-delineated text file, contains model organism ortholog ID, human disease gene ID, inparalog scores, and associated disease(s). This can be restricted according to inparalog scores of either the model organism ortholog or of the human disease gene. The format can also be altered so that only one human disease gene per ortholog is shown (ortholog clusters may contain more than one human disease gene).

OrthoDisease Content

As of March 2004, 2,401 OMIM numbers were extracted from the OMIM Morbid Map. A total of 7,896 Genbank accession numbers were associated with 1,589 of these OMIM numbers using the Locuslink mim2loc and loc2acc tables. After mapping the Genbank entries to 2,757 Swissprot/trEMBL proteins, a total of 2,466 disease-genes were associated with an Inparanoid cluster in at least one organism. A flow chart of the processing from OMIM and Swissprot to OrthoDisease is shown in Figure 2. The total number of disease gene orthologs present in OrthoDisease is currently 6,785, and the breakdown of Inparanoid clusters and disease-gene orthologs per species is shown in Table 1.

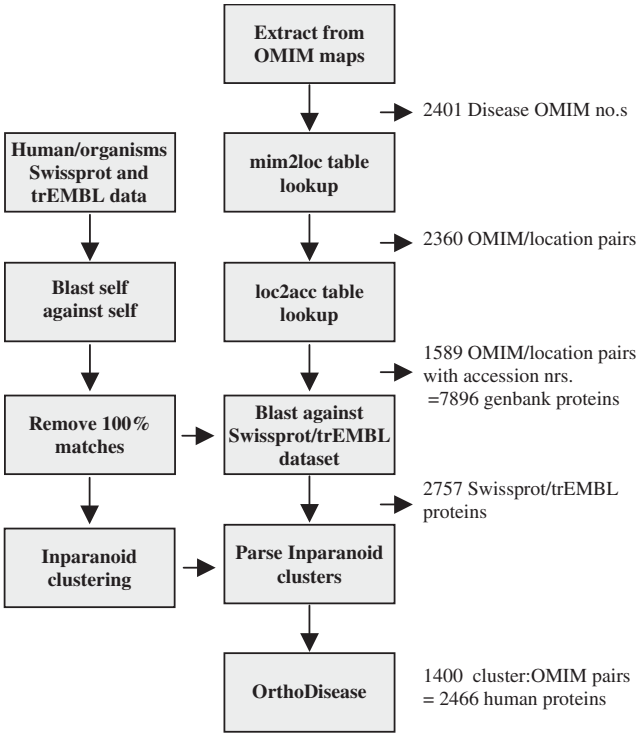

FIGURE 2. **The processing of records from OMIM to OrthoDisease. Actions are indicated in shaded boxes and the results from OMIM Morbid Map and Swissprot/trEMBL are shown to the right. OMIM numbers from the Morbid Map consist of only disease/phenotype entries. Whole genome Inparanoid analysis was performed in parallel and matching clusters were parsed and coupled with disease information to form OrthoDisease.**

An example output of OrthoDisease is shown (Fig. 3). Here, the human Insulin gene (INS) is included in OrthoDisease due to its mutation in rare forms of diabetes mellitus. One Inparanoid cluster was detected containing the human gene and its two orthologs in mouse. Mouse INS2 received an inparalog score of 1.0, as it formed the seed-ortholog pair with human INS and is, therefore, more similar to human insulin than INS1 is. Mouse INS1 scores 0.182, since it has diverged from its inparalog since speciation. This ortholog cluster is consistent with the results seen in knockout studies in mouse. Separate knockout of either INS1 or INS2 results in healthy mice; however, in the case of INS2-null mice, a dramatic overexpression of the more divergent insulin ortholog INS1 is seen, while the reverse is not the case [Leroux et al., 2003]. Furthermore, double knockout mice (INS1–/– and INS2–/–) develop fatal hyperglycemia in the first week of life [Leroux et al., 2001]. Thus, the presence of two insulin inparalogs in mouse clearly confounded knockout studies due to the subsequent redundancy in the mouse insulin system.

TABLE 1. Inparanoid Ortholog Cluster and Gene Content of OrthoDisease

Organisms	Ortholog clusters	Orthologs
M. musculus	1354	1569
D. melanogaster	724	1406
C. elegans	533	1033
A. thaliana	398	2171
S. cerevisiae	290	383
E. coli	153	223

The average Inparanoid cluster size in OrthoDisease varies widely across species, with the lowest seen in *M. musculus* (1.2 mouse orthologs per Inparanoid cluster) and the highest seen in *A. thaliana* (5.5), while *D. melanogaster* and *C. elegans* were similar in cluster size (1.9). The full distribution of Inparanoid clusters across all species can be seen in Table 2. Only 2% of mouse clusters contain more than two members; this is likely due to the relatively recent divergence of the human and mouse lineages. Investigation of the larger clusters seen in mouse reveals that the majority of very large clusters contain either histocompatibility antigens and interferons, which have duplicated into multiple inparalogs both in human and mouse [Hughes, 1995; Takada et al., 2003]. This is consistent with the need for diversity in the immune system, which is strongly linked to fitness and is under significant selection pressure, even over a short evolutionary time span. Other families of orthologs to human disease genes that have expanded in mouse are trypsins, anti-trypsins, and serpins, which all function together in such processes as fibrinolysis, complement activation, and blood coagulation [Reid et al., 1993]. The pattern of cluster-size distribution differed in the *D. melanogaster*, *S. cerevisiae*, and *C. elegans* OrthoDisease datasets, when compared with the complete genome analysis of these organisms, in which a higher percentage of large clusters (four members or greater) was seen in OrthoDisease (9.3, 6.2, and 6.3% vs. 2.1, 0.6, and 2.2%, respectively). Why expansion of inparalogs would be more frequent in disease genes and their orthologs is unclear. It cannot be due to observational bias (i.e., disease gene orthologs are more

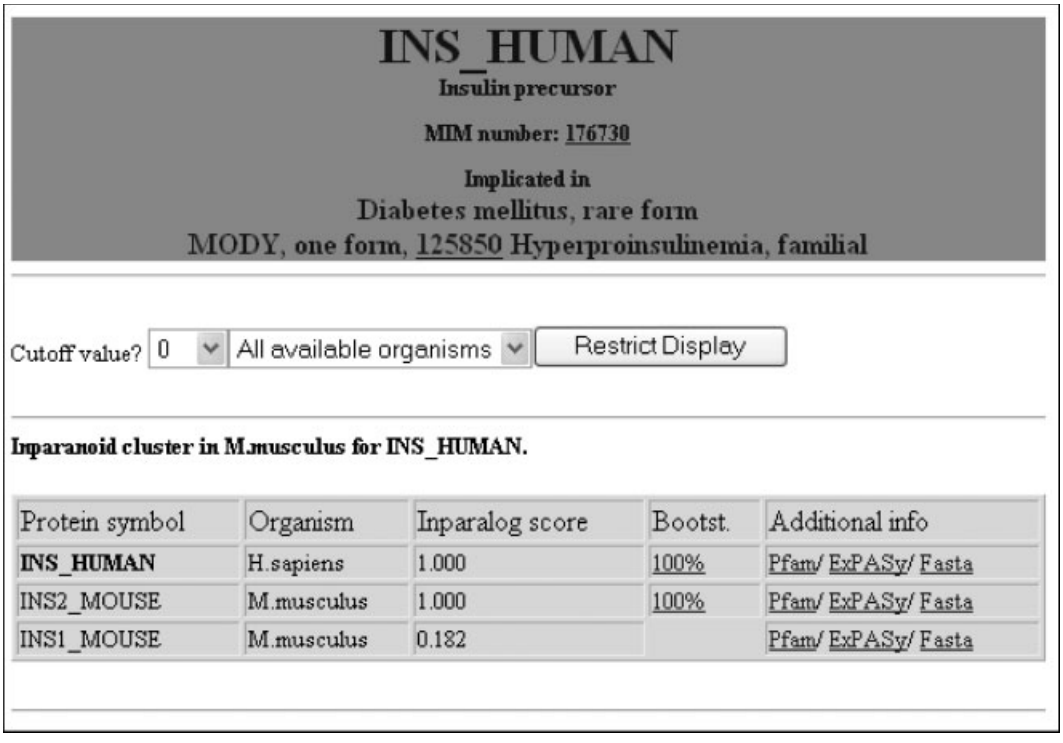

FIGURE 3. **A typical output of OrthoDisease for mouse is shown. Ortholog clusters are displayed separately for each organism where a cluster was found. Here, Insulin is used as the query protein for demonstration purposes.**

TABLE 2. Distribution of Inparanoid Cluster Size in OrthoDisease

Cluster size	MM	CE	DM	AT	SC	EC
1	1240	362	377	99	229	124
2	85	93	209	101	43	18
3	15	38	71	53	12	4
4	5	8	31	36	2	3
5	2	7	11	18	1	2
6	1	7	6	15	2	1
>6	6	18	19	76	1	1
Total	1354	533	724	398	290	153

TABLE 3. Distribution of Human Disease Gene Orthologs in All Organisms

Total Organisms	Total no.
0	291
1	945
2	549
3	432
4	284
5	177
6	79

TABLE 4. Functional Summary of 679 Human Genes With Orthologs in *Drosophila melanogaster*, *Caenorhabditis elegans*, and *mus musculus*

Gene ontology derived function	Total no.
Regulation of transcription	353
ATP synthesis coupled electron transport	165
DNA ligation	158
Mismatch repair	137
Glycolysis	119
Mucilage metabolism	96
Amino acid metabolism	96
Gluconeogenesis	69
Fatty acid biosynthesis	59
Xenobiotic metabolism	50
Proteoglycan metabolism	49
Ion transport	49
Pigment metabolism	48
Monosaccharide metabolism	47
Carbohydrate phosphorylation	47

frequently studied/cloned), since these analyses were based on complete genomes, nor was it found upon manual inspection to be due to the overrepresentation of sequence duplicates. Expansions in these clusters tended to be mirrored by a matched expansion in humans, thus implying enrichment in redundancy in human disease genes.

Of the 2,757 genes in OMIM Morbid Map, 291 lacked orthologs, while the 2,466 genes that form OrthoDisease often had orthologs in more than one organism (Table 3). A total of 679 human disease genes had orthologs in *M. musculus*, *D. melanogaster*, and *C. elegans*. GO numbers were extracted from each of the corresponding Swissprot entries, and a redundant list of function-ontologies was obtained. Because each entry was linked to multiple GO numbers, only the most relevant annotations regarding cellular processes were considered, e.g., "visual perception" was not used, while "ion transport" was. A list of the most consistently referred ontologies is shown (Table 4). As would be expected, fundamentally important cellular processes such as carbohydrate and amino acid metabolism are common to this group, as are maintenance of DNA and regulation of transcription. Such processes are often involved in disease [Chien et al., 2002; Rubin et al., 2000].

DISCUSSION

Inparanoid is an ortholog/inparalog finding algorithm that analyzes whole-genome Blast comparisons in order to cluster genes together according to both intra- and intergenome similarity [Remm et al., 2001]. Its power lies in its ability to identify orthologs, while simultaneously differentiating between outparalogs and inparalogs, in real-time. The performance of Inparanoid has been assessed and found to attain levels of ortholog assignment equal to or better than other whole genome analysis tools [Li et al., 2003; Remm et al., 2001]. Only OrthoMCL performed equally well, but since the Inparanoid tool is freely available, it was chosen as the tool of choice for orthology analysis in this study.

Here we report the release of OrthoDisease, an online database that presents the orthology analysis of the subset of the human genome known to be mutated in inherited disease. OrthoDisease is unique in this area, as it addresses the confounding effect of outparalogs in proper human disease gene ortholog detection. TIGR's Orthologs of Human Disease Genes database is a comprehensive resource that contains orthologs of OMIM genes identified using COGs, and is the only resource that is comparable with OrthoDisease in content. However, OrthoDisease restricts human disease genes to those that are known to have alterations in disease, while making available as a separate resource those genes that have been mentioned in OMIM (disease candidates and nondisease genes). TIGR's resource does not make this distinction, nor does it differentiate between inparalogs and outparalogs, and may report the latter as orthologs [Li et al., 2003].

It should be noted that in OrthoDisease, the human disease gene of interest is often not the seed-ortholog pair of a cluster, or the only human gene that has an inparalog score of 1.0. Inparalogs may exist that are identical to the seed-ortholog and therefore also get an inparalog score of 1.0. OrthoDisease will report Inparanoid clusters in organisms in which the inparalog score of the disease-gene is less than 1.0, i.e., is not part of the seed-ortholog pair. Irrespective of where a disease gene occurs in a cluster, both it and its inparalogs are co-orthologous to model organism genes that occur in the Inparanoid cluster.

Analysis of the distribution of orthologs demonstrated that genes mutated in human disease tend to be conserved across species. Genes that are shared between common model organisms tend to be those involved in basic cellular processes such as metabolism and cellular control. Analyses of ortholog duplications yielded the unexpected finding that the average cluster size seen in

human disease ortholog groups was considerably larger than the average cluster size from the respective whole genomes. This trend may lie in the process of human disease; i.e., the most common phenotype of a perturbed system in nature is not in fact disease, but death. Gene alterations that produce disease are generally based on subtle gene perturbations, in such a manner that an individual survives, albeit with reduced fitness. Redundancy and specialization of inparalogs protect an organism against fatal alterations, while maintaining efficiency of the system. An overrepresentation of large ortholog clusters for disease genes is consistent with a lack of fatal-mutation data.

OrthoDisease is automatically and regularly updated to reflect the additional information that is deposited into SWISS/trEMBL, OMIM, and LocusLink. The number of organisms in the database is set to grow as more genomes are completed. Since analysis is performed pairwise (human vs. model organism), the database will expand in a manageable linear fashion. This may also benefit the selection of which model organism(s) to select for disease mimicry and give a more complete evolutionary picture of disease genes. While the OMIM Morbid Map is a reliable source of disease genes, it encompasses only diseases that tend to be both Mendelian in character and have experimentally-confirmed and published mutations. Hence other sources of disease-genes will be explored. For example, CandiGene is a database currently under development; it contains the best candidate genes in major diseases, including complex genetic diseases. This database is curated manually and updated and expanded automatically, using natural language processing of Medline entries. Incorporation of this database into OrthoDisease will allow researchers to analyze orthology to new disease candidate genes as they are discovered.

REFERENCES

Aboobaker AA, Blaxter ML. 2000. Medical significance of Caenorhabditis elegans. Ann Med 32:23–30.

Ahringer J. 1997. Turn to the worm! Curr Opin Genet Dev 7: 410–415.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Asumalahti K, Veal C, Laitinen T, Suomela S, Allen M, Elomaa O, Moser M, de Cid R, Ripatti S, Vorechovsky I, Marcusson JA, Nakagawa H, Lazaro C, Estivill X, Capon F, Novelli G, Sarrialho-Kere U, Barker J, Trembath R, Kere J. 2002. Coding haplotype analysis supports HCR as the putative susceptibility gene for psoriasis at the MHC PSORS1 locus. Hum Mol Genet 11:589–597.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. 2004. The Pfam protein families database. Nucleic Acids Res 32:D138-D141.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365–370.

Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, Weng S, Cherry JM, Botstein D. 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science 282:2022–2028.

Chien S, Reiter LT, Bier E, Gribskov M. 2002. Homophila: human disease gene cognates in Drosophila. Nucleic Acids Res 30:149–151.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Zool 19:99–113.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31:3784–3788.

Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30:52–55.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261.

Hughes AL. 1995. The evolution of the type I interferon gene family in mammals. J Mol Evol 41:539–548.

Leroux L, Desbois P, Lamotte L, Duvillie B, Cordonnier N, Jackerott M, Jami J, Bucchini D, Joshi RL. 2001. Compensatory responses in mice carrying a null mutation for Ins1 or Ins2. Diabetes 50(suppl 1):S150–S153.

Leroux L, Durel B, Autier V, Deltour L, Bucchini D, Jami J, Joshi RL. 2003. Ins1 gene up-regulated in a beta-cell line derived from Ins2 knockout mice. Int J Exp Diabetes Res 4:7–12.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189.

Maglott DR, Katz KS, Sicotte H, Pruitt KD. 2000. NCBI's LocusLink and RefSeq. Nucleic Acids Res 28:126–128.

Mushegian AR, Garey JR, Martin J, Liu LX. 1998. Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. Genome Res 8:590–598.

Nembaware V, Crum K, Kelso J, Seoighe C. 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. Genome Res 12:1370–1376.

O'Brien KP, Holm SJ, Nilsson S, Carlen L, Rosenmuller T, Enerback C, Inerot A, Stahle-Backdahl M. 2001. The HCR gene on 6p21 is unlikely to be a psoriasis susceptibility gene. J Invest Dermatol 116:750–754.

Pruitt KD, Katz KS, Sicotte H, Maglott DR. 2000. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet 16:44–47.

Reid KB, Nolan KF, Lijnen HR, Collen D. 1993. Proteolytic enzymes in coagulation, fibrinolysis, and complement activation. Introduction. Methods Enzymol 223:1–9.

Reiter LT, Potocki L, Chien S, Gribskov M, Bier E. 2001. A systematic analysis of human disease-associated gene sequences in Drosophila melanogaster. Genome Res 11:1114–1125.

Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314:1041–1052.

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH, Lewis S. 2000. Comparative genomics of the eukaryotes. Science 287:2204–2215.

Sonnhammer EL, Durbin R. 1994. A workbench for large-scale sequence homology analysis. Comput Appl Biosci 10:301–307.

Sonnhammer EL, Koonin EV. 2002. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet 18:619–620.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12:1611–1618.

Storm CE, Sonnhammer EL. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics 18:92–99.

Takada T, Kumanovics A, Amadou C, Yoshino M, Jones EP, Athanasiou M, Evans GA, Fischer Lindahl K. 2003. Species-specific class I gene expansions formed the telomeric 1 mb of the mouse major histocompatibility complex. Genome Res 13: 589–600.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28: 33–36.

Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Suzek TO, Tatusova TA, Wagner L. 2004. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res 32:D35–D40.

Xie T, Ding D. 2000. Investigating 42 candidate orthologous protein groups by molecular evolutionary analysis on genome scale. Gene 261:305–310.