BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 5. 463–473
Advance Access published on 12 May 2011
doi:10.1093/bib/bbr024

OrthoDisease: tracking disease gene orthologs across 100 species

Kristoffer Forslund, Fabian Schreiber*, Nattaphon Thanintorn and Erik L.L. Sonnhammer*

Submitted: 4th March 2011; Received (in revised form): 30th March 2011

Abstract

Orthology is one of the most important tools available to modern biology, as it allows making inferences from easily studied model systems to much less tractable systems of interest, such as ourselves. This becomes important not least in the study of genetic diseases. We here review work on the orthology of disease-associated genes and also present an updated version of the InParanoid-based disease orthology database and web site OrthoDisease, with 14-fold increased species coverage since the previous version. Using this resource, we survey the taxonomic distribution of orthologs of human genes involved in different disease categories. The hypothesis that paralogs can mask the effect of deleterious mutations predicts that known heritable disease genes should have fewer close paralogs. We found large-scale support for this hypothesis as significantly fewer duplications were observed for disease genes in the OrthoDisease ortholog groups.

Keywords: *Orthology; disease; database; gene duplication*

INTRODUCTION

Orthology

The increasing availability of completely sequenced genomes of model organisms and their use in comparative genomics allows us to further understand genes and the processes they take part in. A subset of genes that is of major importance are disease genes. This can mean either genes that are targets for bacterial or viral exploitation, or genes where defective alleles have been identified as the cause of diseases. The exploration of the underlying evolutionary processes affecting these genes, as well as mapping them to functionally equivalent genes in model organisms to create disease models, enables many relevant applications such as drug design.

To find genes in one species that play a functionally equivalent role to particular genes in another species, biologists search for orthologous genes. Orthologs are homologous genes that diverged by a speciation event [1]. In contrast, paralogs are homologs that diverged by a gene duplication event. Moreover, in the context of a given species comparison, paralogs are further divided into genes that arose through gene duplication events predating the last common ancestor of the pair of species in question, called outparalogs, and genes that arose through gene duplication after the two lineages diverged, called inparalogs [2, 3]. A gene that has a recent same-species duplicate may still be orthologous to a gene in another species, and it will share this co-ortholog status with its inparalog(s). Since the ancestral function can be expected to be retained to a high extent, the functions represented by the set of group members in each lineage should be roughly

Corresponding author. Erik L.L. Sonnhammer, Swedish eScience Research Center and Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, Albanova, 10691 Stockholm, Sweden. Tel: +46-8-55378567; Fax: +46-8-55378214; E-mail: erik.sonnhammer@sbc.su.se

*These authors contributed equally to the work.

Kristoffer Forslund is a PhD student at the Stockholm Bioinformatics Centre. His research involves how protein functions evolve, as well as the role of protein domains in this process.

Fabian Schreiber is a Postdoc at the Stockholm Bioinformatics Centre. His research focusses on orthology detection, taxonomic profiling in metagenomics and phylogenomics of early animals.

Nattaphon Thanintorn is a project student at the Stockholm Bioinformatics Centre.

Erik Sonnhammer is Professor of Bioinformatics at Stockholm University and director of the Stockholm Bioinformatics Centre. He heads a bioinformatics research lab focusing on protein function prediction, protein domain evolution, orthology analysis and protein networks.

the same, whereas outparalogous genes are expected to functionally have diverged relatively more.

The process of orthology assignment involves the correct identification of ortholog groups between species. Ortholog groups between two species can be seen as each descending from a single gene in the last common ancestral species, with one or more genes found in the two extant species. Genes from the same species in an ortholog group are inparalogous to each other, while genes from different species are orthologous to each other. Homologous genes in different groups are outparalogous to each other, since they diverged previous to the speciation.

Orthology assignment approaches

In general, there are two different approaches to assigning orthology: tree- and graph-based methods. Tree-based methods try to reconcile gene trees with a species tree by labeling the internal nodes of the gene tree as either speciation or duplication events. This approach is generally computationally expensive [4].

Graph-based approaches perform all versus all searches of all input genes using a sequence similarity search tool, such as BLAST [5]. Pairs of genes, one from either species, that are each other's highest scoring match in this search (reciprocal best hits) are considered edges in a graph. Ortholog groups can be detected in the graph by finding cliques of interconnected nodes. However, it is not trivial to extend these groups with inparalogs, and several algorithms have been developed to this end. The graph-based approach can sometimes be misleading, such as when a gene loss has occurred. [2, 6–8]. Examples of graph-based implementations include COG [9], ORTHOMCL [10], OMA [11] and InParanoid [12].

We have continued to use InParanoid as the underlying orthology inference tool for the second release of OrthoDisease for several reasons. Its relatively low computational cost allows its application to the largest eukaryotic genomes, and to fairly large sets of species. Moreover, few other tools provide such a clear mapping of orthologous, inparalogous and outparalogous relationships between genes.

Databases of human disease orthologs

The study of sequence changes associated with human diseases has accumulated and produced a large body of literature. This information has been combined and organized into databases such as Online Inheritance in Man (OMIM) [13], LocusLink [14], the Human Gene Mutation Database [15] and Genecards [16]. OMIM is a widely used catalog of human diseases and their phenotypes. It lists heritable diseases along with genes whose mutations have been associated with disease or other phenotypes.

Human disease genes are difficult to elucidate functionally by direct study. For many types of molecular functions, it is more fruitful to study their orthologs in model organisms [17]. To assist such studies, a number of databases have been developed to map disease genes to orthologs. These are listed in Table 1 and include Homophila [18], EGO (previously TOGA) [19] and OrthoDisease [20].

In general, the construction of these databases usually involves two steps. First, a set of disease genes of interest is extracted from a database of human diseases such as OMIM and is mapped to the corresponding protein sequences. Once a set of protein sequences of human disease genes is assembled, different approaches are used to find human orthologs in target species. The Homophila database was

Table 1: Overview of existing databases of disease orthologs

Database	URL	No. of species	Analyzed OMIM entries	Disease information source	Ortholog identification approach	Distinguishes inparalogs from outparalogs?
Homophila	http://homophila.sdsc.edu (defunct)	2	1858	OMIM	Best Blast hit against *Drosophila melanogaster*	No
EGO (formerly TOGA)	http://compbio.dfci.harvard.edu/tgi/ego/	90	Not given	OMIM	Reciprocally best Blast hit	No
OrthoDisease 2.0	http://orthodisease.sbc.su.se/	100	2935	OMIM	InParanoid (algorithm version 4)	Yes

The table lists information about the database name, the number of species, the source of disease information, the number of analyzed OMIM entries, the ortholog identification method and whether the database distinguishes between inparalogs and outparalogs.

constructed based on the best (one-way) BLAST hits in pairwise species comparisons, whereas the EGO database defines orthology on the basis of the best reciprocal (two-way) BLAST hit. OrthoDisease was built using the InParanoid algorithm.

InParanoid orthology inference starts by considering reciprocal best matches, but applies additional criteria to rule out sources of errors or inconsistencies such as fragmentary matches, which are avoided by requiring that matches must be longer than a certain fraction of the length of the proteins. In a subsequent step, the resulting ortholog groups are expanded by a set of empirically optimized rules to identify inparalogs of the seed orthologs. A third step may merge or split groups in some cases to produce the final set of group assignments. See [12] for the full algorithm in its most recent form.

Properties of disease-associated genes

The use of model organisms allows study, through experimentation and observation, of the underlying mechanisms of biology, including disease [21]. This is relevant for a wide range of species, as for example even the distantly related yeast includes orthologs of 22% of human disease genes in the updated OrthoDisease database presented here.

Beside locating suitable model systems to study disease mechanisms, we can explore whether disease-associated genes have any distinguishing characteristics. Previous comparisons of genes with and without disease annotations have revealed a number of trends, which are reviewed by Dalkilic *et al.* [22]. Among other things, disease-associated genes tend to be longer on average. They also tend to be more likely to have homologs in distant species, but less likely to have close paralogs than non-disease-associated genes [22].

'Paralog compensation' refers to the case when some or all roles of a loss-of-function mutant gene may nevertheless be filled by duplicates of that gene. A survey by Lopez-Bigas and Ouzounis [23] of the properties of disease-associated genes observed that human genes with many paralogs were less likely to be associated with diseases, and attributed this finding to a mechanism of paralog compensation. Under this model, which we call Scenario 1 (Figure 1), even if a gene is mutated so that its function is lost or reduced, a sufficiently recent duplicate gene may still retain enough vestiges of the ancestral function for the pathway to function, reducing the fitness impact of the mutation. If this effect is strong enough, a disease phenotype might never be observed and no disease association will be recorded. It should be noted, however, that Lopez-Bigas and Ouzounis [23] did not distinguish between inparalogs and outparalogs.

Under an alternative model, which we call Scenario 2 (Figure 1), paralog compensation would have the opposite effect on gene–disease association. By reducing the fitness impact of the mutation, the defective allele would become less likely to be purged from the gene pool, and thus more likely to be observed by disease geneticists. Under this model, we would expect a greater proportion of disease genes in larger ortholog groups, at least for closely related species. In O'Brien *et al.* [20], data from three species indicated that disease genes are found more often in larger ortholog groups, i.e. with more same-species inparalog 'siblings', which would be consistent with paralog compensation acting under Scenario 2.

Gene essentiality may be thought of as in one sense opposite to disease gene involvement. A gene is essential if its disruption causes complete lethality or sterility to the organism. Systematically determining gene essentiality is not possible in human, but has been done in several other species, including yeast [24], worm [25] and mouse [26]. As mutants deficient in essential genes do not procreate, nonfunctional variants of such genes should be rare or nonexistent in the gene pool. Deleterious mutations of essential genes should only rarely be associated with known genetic diseases, as mutants would be non-viable. While not every mutation to an essential gene might disrupt gene function enough to be lethal, this still makes the proportion of mutations resulting in a viable but noticeable disease phenotype lower, and should deplete essential genes from disease associations. We would thus expect traits associated with nonessentiality to be enriched among disease genes.

It should be noted that the above reasoning applies only to loss of function mutations. Gain of function mutations should normally not be subject to paralog compensation. Any trend we observe is thus likely to derive from disease genes with loss of function mutations.

Contents of present study

Here we present an updated version of the OrthoDisease disease orthology database containing groups of disease gene orthologs between human and 99 other species, ranging from chimpanzee to

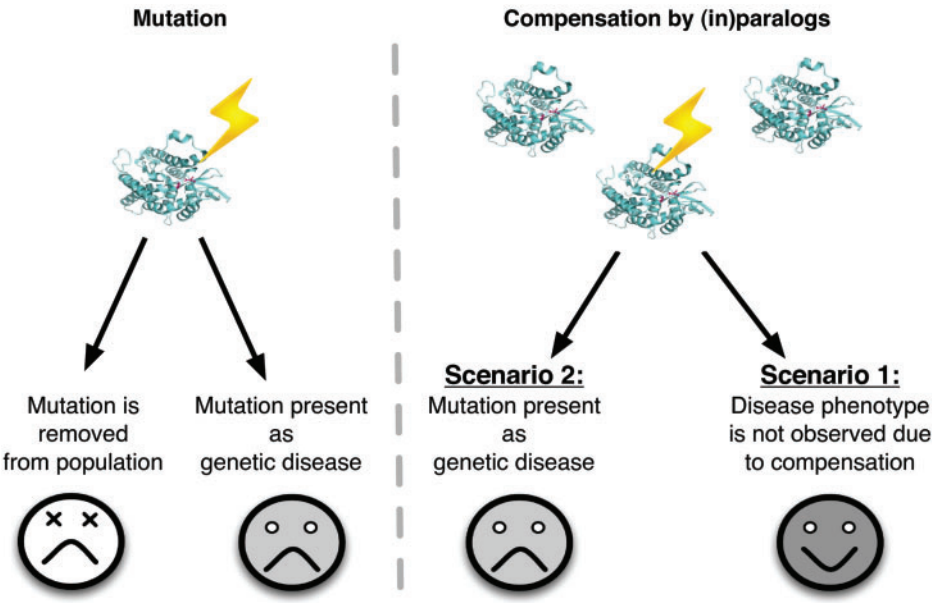


Figure 1: Possible effects of paralog compensation on the likelihood that a gene is disease associated. Suppose a gene with no paralogs is mutated. If it is completely disabled, the mutant allele is likely removed through strong negative selection. If function is merely impaired, chances of a disease allele remaining is higher. Close paralogs may however be able to fulfill the function of a defective gene, which would decrease the chance of purging the mutant from the population. Under Scenario 1, this paralog compensation is strong enough to obscure the mutation's effect entirely, so that no fitness decrease occurs. The mutant allele will thus remain in the population, but will not be known as a disease gene. Under Scenario 2, the paralog compensation is less complete, with enough fitness decrease to make the mutant a target for disease genetics, and thus potentially become known as a disease gene, but not enough fitness decrease for it to have been completely removed by negative selection yet.

Escherichia coli. We updated OrthoDisease using the latest version of InParanoid, which increased the number of species by a factor of 14 (7 versus 100 species) and improved the accuracy of the underlying orthology inferences through increased robustness to error sources such as low-complexity sequence regions or fragmentary matches.

Furthermore, we used OrthoDisease as a basis for a survey of the taxonomic distribution of distinct categories of disease genes. We also investigated whether genes with close paralogs are more or less likely to be associated with disease.

METHODS

Updating the OrthoDisease database

The InParanoid database

Information on gene orthology was taken from version 7 of the InParanoid database [12]. This more recent version has both improved coverage (to 99 eukaryotes and 1 prokaryote) and accuracy, through more stringent filters for fragmentary matches during the sequence comparison step as well as adaptions to reduce false positive matches due to sequence regions that have low complexity or a very biased composition [27].

Disease gene information

We used the OMIM database as a resource for disease-annotated genes that have been implicated or shown to be mutated in diseases. OrthoDisease includes all OMIM entries of type *phenotype* or *gene+phenotype*, excluding pure gene entries as well as entries corresponding to nondisease phenotypes such as hair or skin pigmentation (entries in square brackets in the MorbidMap). This filtering was performed using information from the OMIM MorbidMap (ftp://ftp.ncbi.nih.gov/repository/OMIM/morbidmap), downloaded on 18 February 2011. To link OMIM entries to the Ensembl Gene IDs used by InParanoid to index human genes, mappings were taken from the UniProt flatfile (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz), downloaded on 18 February 2011. All information is stored in SQL

tables for easy access and data analysis, and is made available to the user community through a web server. An outline of this workflow can be seen in Figure 2.

The OMIM diseases were grouped into 22 categories according to Goh *et al*. [28]. We however excluded the categories 'multiple' and 'unclassified' in the analysis.

RESULTS AND DISCUSSION

OrthoDisease online

All InParanoid groups with disease annotations may be accessed through the web server OrthoDisease (http://orthodisease.sbc.su.se), with a number of different interfaces available to the user. It is possible to search for orthologs of genes involved in particular diseases, in one or all model organisms, as well as to list all disease gene ortholog groups that exist between human and a particular species. Furthermore, the web server provides gene identifier, OMIM number and free text search options. Additionally, the download view allows the user to download all disease gene ortholog groups between human and another species as a file.

Each disease gene ortholog group in OrthoDisease consists of the seed ortholog pair and, possibly, a set of genes inparalogous to the seeds. For each gene in the group, an inparalog score is provided that corresponds to the certainty of its inclusion.

OrthoDisease content

As of February 2011, OrthoDisease contains 2935 distinct human disease phenotypes from OMIM, mapped to 2313 out of 21 673 human genes and their orthologs in 99 other species. Supplementary Table 1 shows the distribution across species. The raw numbers of disease gene orthologs mainly reflect proteome size as well as evolutionary proximity, with the plants and, curiously, the lancelet *Branchiostoma floridae* at the top. For the plants, this generally reflects multiple whole-genome duplications multiplying the number of orthologs of some early genes that later came to be disease associated in human. The lancelet is a chordate that diverged from the lineage that led to the jawed vertebrates before a series of whole-genome duplications in that lineage [29]. This tiny chordate split from the vertebrate lineage over 500 million years ago, yet the available proteome is the largest among the animals. Potentially, this may be an artifact of poor genome sequencing, assembly and annotation.

An example of the view provided by the web interface is shown in Figure 3. The human

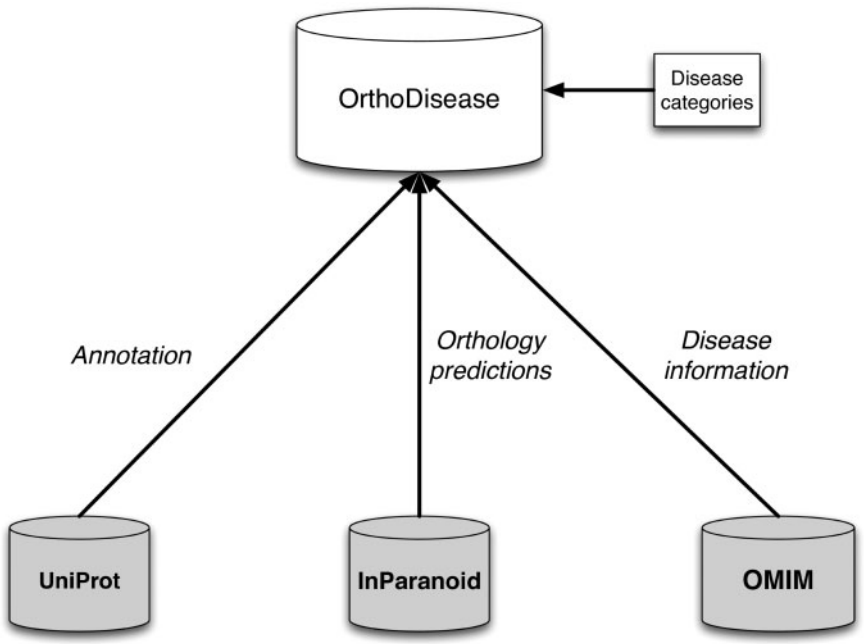
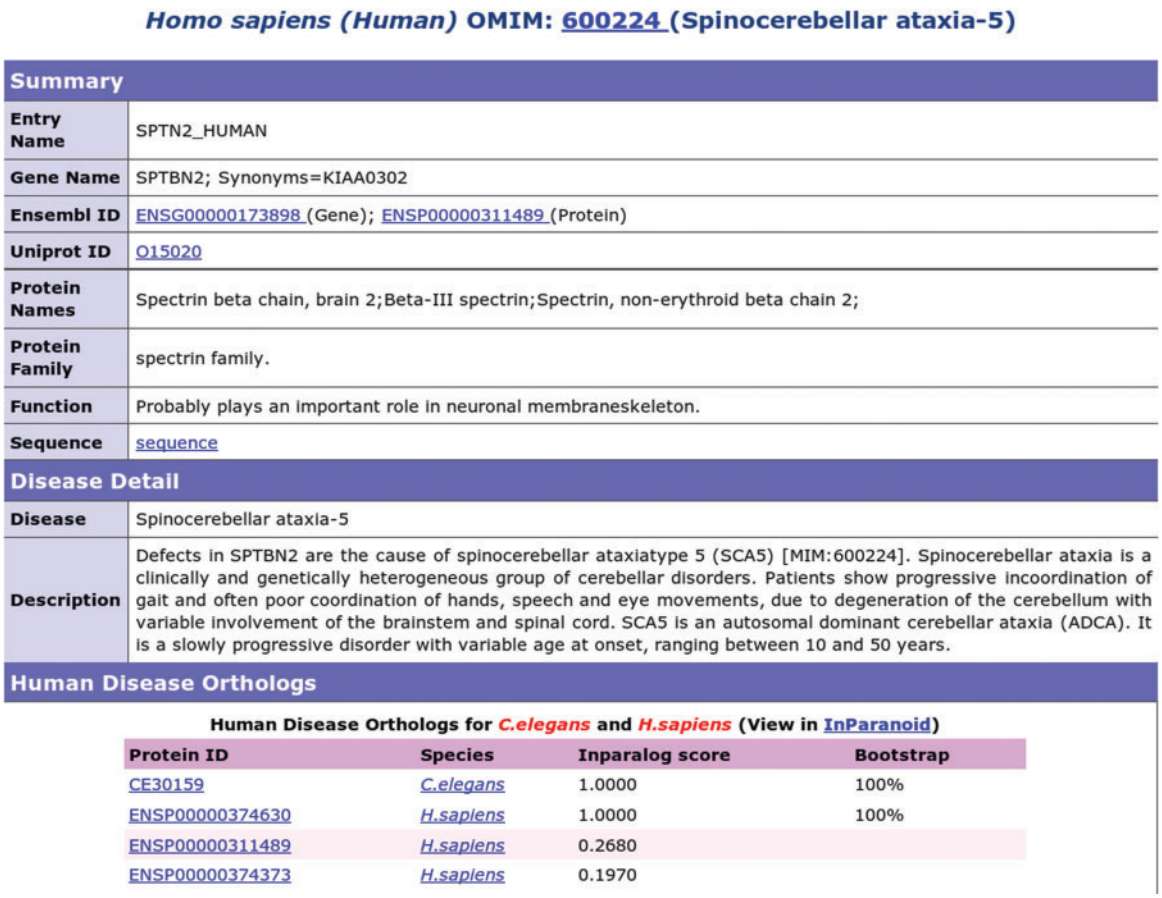


Figure 2: Outline of the main steps in the construction of the OrthoDisease database. Disease information is taken from the OMIM database and is mapped to Ensembl IDs with the help of UniProt. It can then be assigned to the genes in the ortholog groups inferred by InParanoid.



Homo sapiens (Human) OMIM: 600224 (Spinocerebellar ataxia-5)

Summary	
Entry Name	SPTN2_HUMAN
Gene Name	SPTBN2; Synonyms=KIAA0302
Ensembl ID	ENSG00000173898 (Gene); ENSP00000311489 (Protein)
Uniprot ID	O15020
Protein Names	Spectrin beta chain, brain 2;Beta-III spectrin;Spectrin, non-erythroid beta chain 2;
Protein Family	spectrin family.
Function	Probably plays an important role in neuronal membraneskeleton.
Sequence	sequence
Disease Detail	
Disease	Spinocerebellar ataxia-5
Description	Defects in SPTBN2 are the cause of spinocerebellar ataxiatype 5 (SCA5) [MIM:600224]. Spinocerebellar ataxia is a clinically and genetically heterogeneous group of cerebellar disorders. Patients show progressive incoordination of gait and often poor coordination of hands, speech and eye movements, due to degeneration of the cerebellum with variable involvement of the brainstem and spinal cord. SCA5 is an autosomal dominant cerebellar ataxia (ADCA). It is a slowly progressive disorder with variable age at onset, ranging between 10 and 50 years.

Human Disease Orthologs

Human Disease Orthologs for *C.elegans* and *H.sapiens* (View in InParanoid)

Protein ID	Species	Inparalog score	Bootstrap
CE30159	*C.elegans*	1.0000	100%
ENSP00000374630	*H.sapiens*	1.0000	100%
ENSP00000311489	*H.sapiens*	0.2680	
ENSP00000374373	*H.sapiens*	0.1970	

Figure 3: This example entry from the OrthoDisease web site lists the human *SPTBN2* gene, which is associated with spinocerebellar ataxia. The upper part of the entry gives detailed information about the gene along with disease-related information. The lower part shows the InParanoid group the gene belongs to. This group has three members in human and one member in *C. elegans*. For each gene present in the orthology group, the Inparalog score reflects the certainty that the gene is a group member, while the bootstrap score reflects the certainty that the seed orthologs are correctly chosen.

SPTBN2 gene is present in OrthoDisease because mutations in it lead to spinocerebellar ataxia. The entry contains detailed information about the protein's function, the disease and external identifiers. The corresponding InParanoid ortholog group contains three human genes and one gene in *Caenorhabditis elegans*. Phenotypes resulting from a knockout of this gene in worm would thus be relevant for all its human orthologs. For genes associated with more than one disease phenotype, a separate view is available for each association.

Taxonomic and disease category distribution of disease orthologs

We selected a set of representative model organisms and ordered them according to their evolutionary separation from human according to the NCBI Taxonomy common tree. Figure 4 shows the fraction of human disease-associated genes and nondisease-associated genes with orthologs in a selection of model organisms. This indicates whether the system components are ancestral (and well conserved) to the two species or not. Not surprisingly, the more distantly related the species have fewer orthologs, both with and without disease associations. Strikingly, human genes with disease associations are more likely than genes without disease associations to have orthologs, in 95 of 99 species. This is in agreement with the findings of Lopez-Bigas and Ouzounis [23]. Potentially, this reflects the fact that disease genes with orthologs are more amenable to study, and thus more likely to have been recorded in the databases. Also, since disease genes are more studied they may contain fewer errors that could obfuscate orthology relationships.

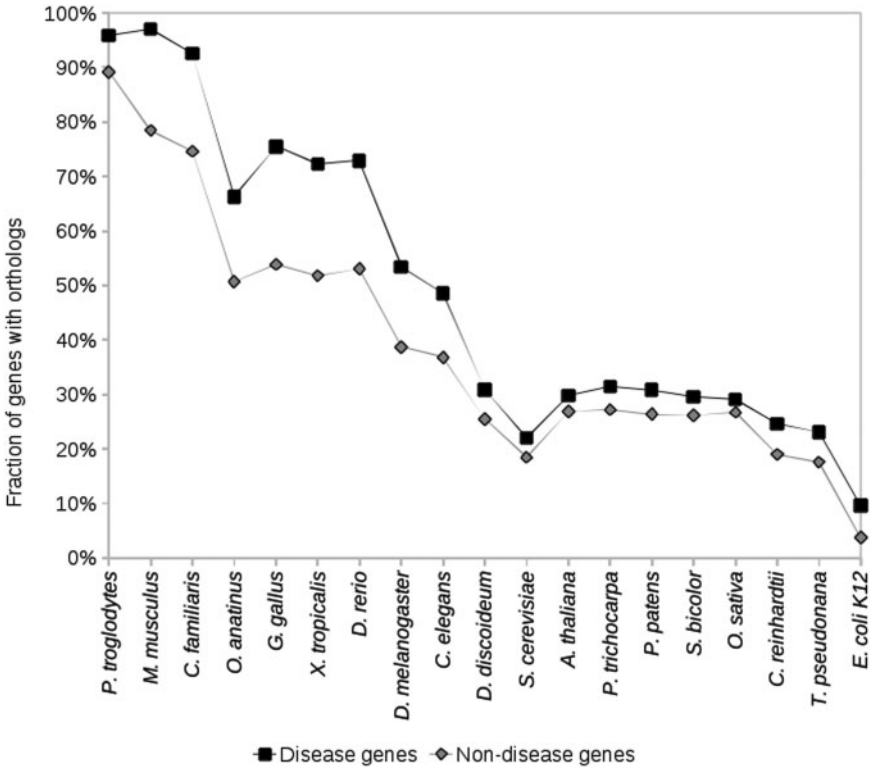


Figure 4: The fraction of human genes with and without disease association that have InParanoid orthologs in selected species.

For a given disease gene, how likely is it to have an ortholog in a given non-human species, and how does this chance vary between disease categories and over evolutionary separations? Figure 5 shows, in the form of a heatmap, the presence of disease gene orthologs in selected species for each of the 20 disease categories. Disease genes in the metabolic category retain a considerably larger fraction of orthologs in the more distant species than other categories; evidently, these are very ancient functions. In most other categories the distant species have much fewer orthologs, indicating more recent functions. For instance, a group of six categories including immunological, bone and psychiatric drops sharply when leaving the vertebrates. This correlates well with vertebrate physiology.

Number of inparalogs of disease genes

Again, by disease genes, we mean human genes with an OMIM disease annotation. Do disease genes differ from non-disease genes in how many human inparalogs they typically have?

We considered a random model where every human gene has an equal chance of being a disease gene, taken as the ratio of disease genes to all human genes included in the experiment. For each species comparison, ortholog groups were retained, but human disease gene annotation status was randomized, by randomly assigning genes as disease associated while maintaining the same fraction of disease genes as in the original data set. The average number of inparalogs of all disease and nondisease genes was computed from 1000 such randomizations of the data set and compared with the original observed values. This provided a test for whether the observed numbers of inparalogs for the different classes of genes was significantly different from those produced by a random model of this type.

As seen in Supplementary Table 2, the observed average number of inparalogs of nondisease genes is much higher than for disease genes. Randomization testing generally fails to produce this difference in the numbers of inparalogs. Hence, disease genes have significantly fewer human inparalogs than nondisease genes when compared with the null model of uniform disease gene distribution, implying that large ortholog groups with many human inparalogs are depleted, or smaller ortholog groups enriched, for

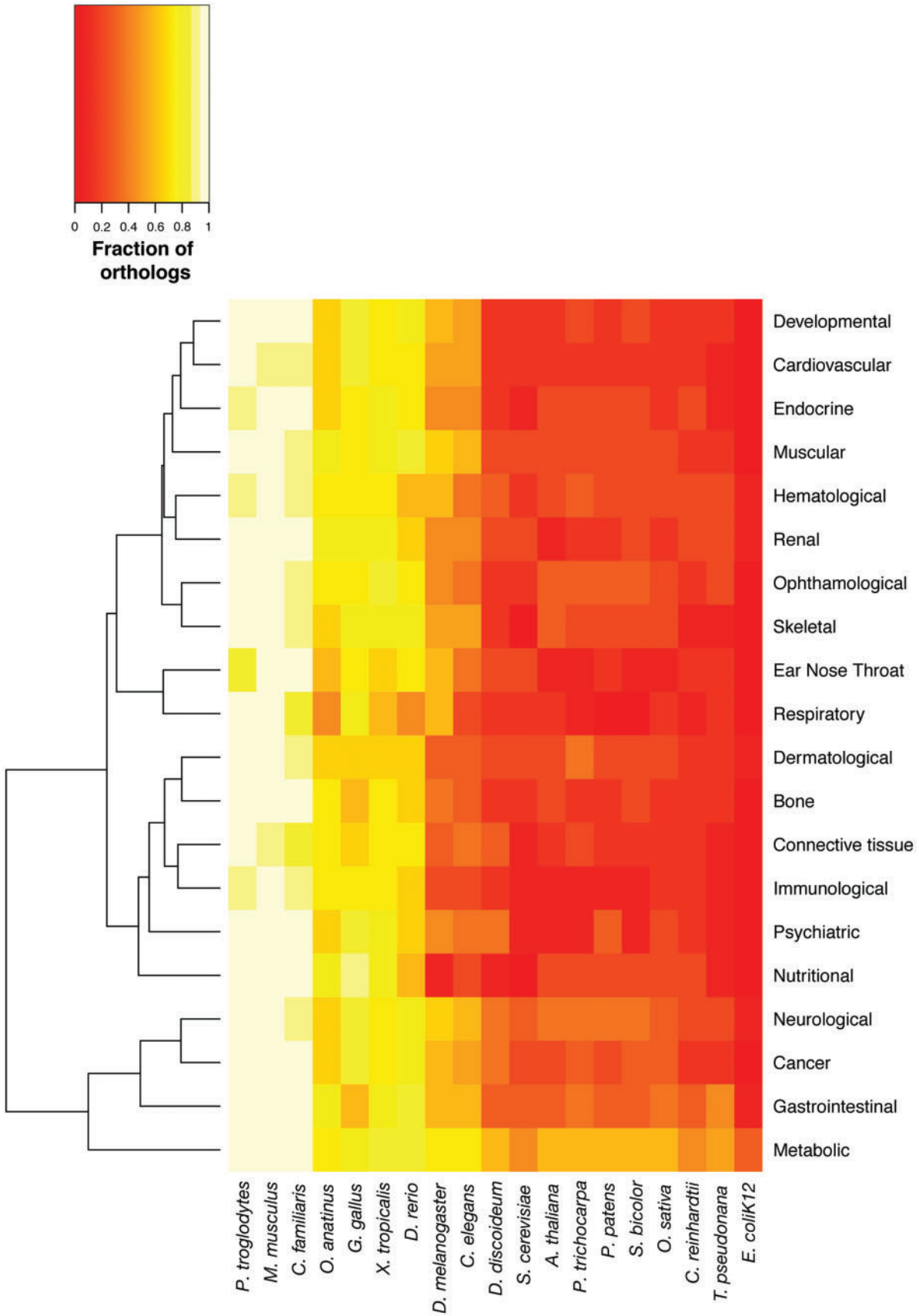


Figure 5: Heat map displaying the fraction of genes in each disease category that have orthologs in each model species.

disease genes. Figure 6 shows for selected species the ratio of the average number of disease gene inparalogs to the average number of inparalogs of all genes, and the equivalent ratio for nondisease genes. For 94 of the 99 species, the average number of inparalogs was significantly lower for disease genes than for nondisease genes. Thus, disease-associated genes tend to have significantly fewer inparalogs.

It was previously reported [20] that ortholog groups containing disease genes were enriched for groups with more than foure members, which seems to conflict with what we observe here. However, from a statistical perspective, ortholog groups containing one or more disease genes are expected to be enriched for larger groups even when randomly assigning disease annotation status to genes. The reason is that larger groups contain more genes that could conceivably be annotated with a disease. Thus, even if genes in larger groups are individually less likely to be disease annotated than those in smaller groups, the likelihood of finding one such gene in a larger group would still be increased. This is merely a statistical effect, which was previously misinterpreted because no control was made. A control randomization experiment similar to the one we report above (data not shown) shows that OrthoDisease groups containing at least one disease-annotated gene are significantly smaller on average than they would be under a model of disease annotation randomly distributed across genes regardless of group size. We believe that this explains the misleading results of the previous study.

CONCLUSION

We have updated the OrthoDisease database both in terms of species and orthology prediction accuracy. We increased the species coverage to the 100 included in the most recent version of the InParanoid database to allow more detailed studies of the distribution of human disease orthologs over a wide taxonomic range.

If no paralog compensation had occurred at all, we would have expected from our null model to see a

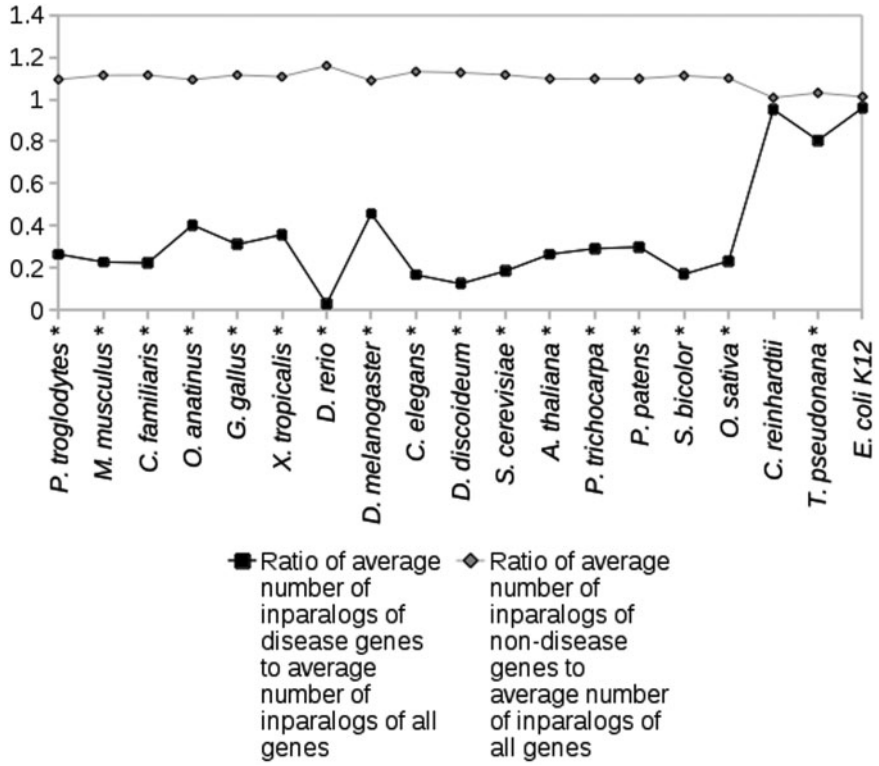


Figure 6: Average number of human inparalogs of genes with and without disease annotations, respectively. On the *Y*-axis is shown the ratio of the average number of human inparalogs of genes of either category, normalized by the average number of human inparalogs across all genes in each species comparison. An asterisk following the species name indicates that the difference in average number of inparalogs between genes of the two categories is significant at $P < 0.05$ under a randomization test.

uniform distribution of disease genes across ortholog group sizes. Alternately, under our two scenarios of paralog compensation, Scenario 1 predicted that large ortholog groups would be depleted for known disease genes, whereas they would be enriched under Scenario 2. This enrichment would come from reduced selective pressure on large ortholog groups to lose disease genes compared with that of singletons or small ortholog groups. We have shown that, for the genes included in OrthoDisease, there is no statistical support for larger ortholog groups, where lineage-specific gene duplication has occurred to a greater extent, to be enriched for disease genes. Instead, we found statistically supported evidence for the reverse trend where less duplicated genes more often are associated with an OMIM genetic disease entry. We are thus able to validate the observation of López-Bigas and Ouzounis [23] on a much larger data set, even when distinguishing between in- and outparalogs. From this, we conclude that paralog compensation is a relevant factor in disease gene evolution, and that its mechanisms and effects predominantly are those we described as Scenario 1—paralogs may reduce disease phenotypes and thus make observation and study of a distinct disease associated with the mutation less likely.

Another possible reason behind the depletion of disease genes in large ortholog groups could be research bias—it may be easier for geneticists to discern disease—gene associations for non-duplicated genes. From a population genetics perspective, each gene duplication represents a recombination event in a single individual, which may spread in the population either by chance or by increasing fitness. The likelihood of this should be smaller if the original individual was carrying a defective allele of the duplicated gene. As the particular duplication was indeed fixed in the population, it makes sense that duplicated genes are depleted for disease alleles. Further population genetic analysis might shed light on this issue, particularly as it applies to diploid genomes.

The bias toward non-duplicated disease genes may also reflect the poor understanding of complex diseases, that are caused by a combination of many weakly defective gene alleles. This would be the case if such diseases preferentially tend to involve systems involving many redundant genetic pathways resulting from gene duplications. If progress is made toward mapping causative genes in complex diseases, this bias may be diminished.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key points

- When exploring genes associated with genetic diseases, knowledge of their orthology relationships is important, not least for the purpose of selecting suitable model systems.
- We present a revision of the OrthoDisease (http://orthodisease.sbc.su.se) database to provide this information, based on disease information from OMIM and orthology relationships between human and 99 other species from InParanoid.
- Disease-associated genes have significantly fewer inparalogs than other human genes. This may be because close homologs retain vestiges of their shared ancestral function, so that mutations disabling only one of them still does not result in a clear disease phenotype.

FUNDING

This work was supported by the Department of Biochemistry and Biophysics at Stockholm University; and the Wenner-Gren Foundations [to F.S.].

References

1. Fitch W. Distinguishing homologous from analogous proteins. *System Zool* 1970;**19**:99–113.
2. Remm M, Storm C, Sonnhammer E. Automatic clustering of orthologs and inparalogs from pairwise species comparisons. *J Mol Biol* 2001;**314**:1041–52.
3. Sonnhammer E, Koonin E. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* 2002;**18**:619–20.
4. Fang G, Bhardwaj N, Robilotto R, *et al*. Getting started in Gene Orthology and functional analysis. *Plos Comput Biol* 2010;**6**:e1000703.
5. Altschul S, Madden T, Schaffer A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
6. Chervitz S, Aravind L, Sherlock G, *et al*. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 1998;**282**:2022–8.
7. Xie T, Ding D. Investigating 42 candidate orthologous protein groups by molecular evolutionary analysis on genome scale. *Gene* 2000;**261**:305–10.
8. Nembaware V. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Res* 2002;**12**:1370–6.
9. Tatusov RL, Fedorova ND, Jackson JD, *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003;**4**:41.
10. Li L, Stoeckert C Jr, Roos D. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;**13**:2178–89.

11. Altenhoff A, Schneider A, Gonnet G, *et al*. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res* 2011;**39**(Suppl. 1):D289–94.
12. Östlund G, Schmitt T, Forslund K, *et al*. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* 2010;**38**(Suppl. 1):D196–203.
13. Hamosh A, Scott A, Amberger J, *et al*. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005; **33**(Suppl. 1):D514–17.
14. Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 2001;**29**: 137–40.
15. Stenson PD, Ball EV, Howells K, *et al*. The Human Gene Mutation Database:providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum Genomics* 2009;**4**:69–72.
16. Safran M, Dalah I, Alexander J, *et al*. GeneCards Version 3: the human gene integrator. *Database*. Advanced access published 5 August 2010; doi: 10.1093/database/baq020.
17. Aboobaker AA, Blaxter ML. Medical significance of *Caenorhabditis elegans*. *Ann Med* 2000;**32**:23–30.
18. Chien S, Reiter L, Bier E, *et al*. Homophila: human disease gene cognates in Drosophila. *Nucleic Acids Res* 2002;**30**: 149–51.
19. Quackenbush J, Cho J, Lee D, *et al*. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res* 2001;**29**: 159–64.
20. O'Brien K, Westerlund I, Sonnhammer E. OrthoDisease: a database of human disease orthologs. *Hum Mutat* 2004;**24**: 112–19.
21. Bedell MA, Jenkins NA, Copeland NG. Mouse models of human disease. Part I: techniques and resources for genetic analysis in mice. *Genes Dev* 1997;**11**:1–10.
22. Dalkilic M, Costello J, Clark W, *et al*. From protein-disease associations to disease informatics. *Front Biosci* 2008;**13**: 3391–407.
23. Lopez-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* 2004;**32**:3108–14.
24. Gu Z, Steinmetz L, Gu X, *et al*. Role of duplicate genes in genetic robustness against null mutations. *Nature* 2003;**421**: 63.
25. Conant G, Wagner A. Asymmetric sequence divergence of duplicate genes. *Genome Res* 2003;**13**:2052–8.
26. Liao B, Zhang J. Mouse duplicate genes are as essential as singletons. *Trends Genet* 2007;**23**:378–381.
27. Forslund K, Sonnhammer EL. Benchmarking homology detection procedures with low complexity filters. *Bioinformatics* 2009;**25**:2500–2505.
28. Goh K-I, Cusick M, Valle D, *et al*. The human disease network. *Proc Natl Acad Sci USA* 2007;**104**:8685–90.
29. Putnam NH, Butts T, Ferrier DEK, *et al*. The amphioxus genome and the evolution of the chordate karyotype. *Nature* 2008;**453**:1064–1071.