Informatics

Incorporating Mouse Genome
Mammalian Genome
Genes and Phenotypes

HCOP: The HGNC comparison of orthology predictions search tool

Mathew W. Wright, Tina A. Eyre, Michael J. Lush, Sue Povey, Elspeth A. Bruford

HUGO Gene Nomenclature Committee, The Galton Laboratory, Department of Biology, University College London, Wolfson House, 4, Stephenson Way, London, NW1 2HE, UK

Received: 28 July 2005 / Accepted: 4 October 2005

Abstract

The HGNC Comparison of Orthology Predictions search tool, HCOP (http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/hcop.pl), enables users to compare predicted human and mouse orthologs for a specified gene, or set of genes, from either species according to the ortholog assertions from the Ensembl, HGNC, Homologene, Inparanoid, MGI and PhIGs databases. Users can assess the reliability of the prediction from the number of these different sources that identify a particular orthologous pair. HCOP provides a useful one-stop resource to summarise, compare and access various sources of human and mouse orthology data.

Orthologs are genes in different species that derive from a common ancestor without duplication, and generally share the same function. The HUGO Gene Nomenclature Committee (HGNC) collaborates with the Mouse Genomic Nomenclature Committee in order to reduce interspecies nomenclature confusion by assigning, where possible, the equivalent gene symbol to orthologous human and mouse genes (e.g., *PON2* in human, Pon2 in mouse).

Establishing orthology relationships is a major task in the post-genomic era. Various groups report orthology information but a single tool for comparison of these data to identify a consensus of the orthology predictions has not previously been available. The new HGNC Comparison of Orthology Predictions search tool, HCOP, found at http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/hcop.pl, fulfills this role for human and mouse genes. It returns orthology assertions made by Ensembl (Hubbard et al. 2005), HGNC (Wain et al. 2004), Homologene (Wheeler et al. 2005), Inparanoid (Remm et al. 2001; O''Brien et al. 2005), MGI (Eppig et al. 2005) and PhIGs (Dehal & Boore 2005). Documentation and help is available at http://www.gene.ucl.ac.uk/nomenclature/data/humot_documentation.html

A comparison of orthology assertions for a human or mouse gene can be obtained by searching HCOP with either an approved symbol (e.g., *PON2*), a term from an approved gene name (e.g., ''oxonase''), Entrez Gene ID, HGNC ID or MGI ID, or RefSeq accession (e.g., NM_000305; Pruitt et al. 2005). A file containing a list of identifiers can also be uploaded to facilitate multiple searches. The results return the official nomenclatures, sequence accessions, database identifiers, aliases, and chromosomal locations for each putative ortholog pair. As conservation of synteny occurs throughout mammalian genomes and is a reliable indicator of orthology, HCOP also indicates whether the human and mouse ortholog are from syntenic chromosomes. A list of databases that support each assertion and links to sources of further information about these databases are also provided.

Paralogs are genes within a specific genome related by duplication from a common ancestor. Ensembl, Inparanoid and Homologene sometimes return more than one predicted ortholog, which usually indicates the presence of closely related paralogs. Inparanoid also differentiates between inparalogs and outparalogs; inparalogs arise through a gene duplication event after speciation, whereas outparalogs arise following a gene duplication before speciation (Remm et al. 2001; O''Brien et al. 2005). Hence, where an HCOP search shows more than one predicted ortholog for any gene in one species, this may be indicative of paralogy as reported by the above databases.

Correspondence to: Mathew W. Wright; E-mail: nome@galton.ucl.ac.uk

Table 1. Number of human/mouse orthology assertions predicted and shared by each database.

TOTAL 70711	*ENS*	*HGNC*	*HGENE*	*INPAR*	*MGI*	*PHIGS*
ENS	15185	2791	12799	11273	12763	9557
HGNC	2791	3296	2929	2348	3098	2004
HGENE	12799	2929	15649	10693	13874	9026
INPAR	11273	2348	10693	11506	10658	8150
MGI	12763	3098	13874	10658	15441	9039
PHIGS	9557	2004	9026	8150	9039	9634

ENS: Ensembl; HGNC: HUGO Gene Nomenclature Committee; HGENE: Homologene; INPAR: Inparanoid; MGI: Mouse Genome Informatics; PHIGS: PhIGs.

A particular search results page showing a set of genes of interest can be easily returned to by bookmarking the URL, or linked to from external webpages. HCOP has also been used to generate comparative files listing predicted human/mouse ortholog pairs. These data are available for download at http://www.gene.ucl.ac.uk/nomenclature/hcop_ hum_mus.pl.

Orthology dataset comparisons

Analysis of HCOP data has also provided a simple means to assess the level of agreement between the orthology assertions made by different databases (Table 1). This analysis has shown that there is general agreement between the different databases, although most sources provide orthology assertions for only a small number of genes. HGNC provides the smallest number of assertions (3,296), and Homologene the greatest number (15,649). While the coverage of HGNC is small, its data is the only set to have been entirely manually curated, providing a good quality check of the automated or semi-automated methods used by other groups. There are 1,613 examples where an orthology prediction has been agreed by all 6 databases and 8,337 examples where the orthology is agreed by 5 of the databases. Once complete coverage of the human and mouse genomes is available, it should be possible to identify human/mouse ortholog pairs with even greater confidence.

Conclusions

HCOP provides a useful tool to obtain and compare human/mouse orthology data and will greatly increase the speed and ease with which such data can be obtained. Potentially this resource can also be expanded to show orthology assertions between other mammalian genomes, facilitating our understanding of evolutionary relationships throughout mammals.

Acknowledgment

HCOP has been developed as part of the Human & Mouse Orthologous Gene Nomenclature (HUMOT) project (http://www.gene.ucl.ac.uk/nomenclature/humot) which is funded by the Wellcome Trust. The work of the HGNC is also supported by NHGRI grant P41 HG003345 and the UK Medical Research Council.

References

1. Dehal P, Boore JL (2005) Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. PLoS Biol 3, e314
2. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, Anagnostopoulos A, et al. (2005) The Mouse Genome Database (MGD): from genes to mice–a community resource for mouse biology. Nucleic Acids Res 33(Database issue), D471–D475
3. Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, et al. (2005) Ensembl 2005. Nucleic Acids Res 33(Database issue), D447–D453
4. O''Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33(Database issue), D476–D480
5. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33(Database issue), D501–D504
6. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314, 1041–1052
7. Wain HM, Lush MJ, Ducluzeau F, Khodiyar VK, Povey S (2004) Genew: the Human Gene Nomenclature Database, 2004 updates. Nucleic Acids Res 32 (Database issue), D255–D257
8. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, et al. (2005) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 33(Database issue), D39–D45