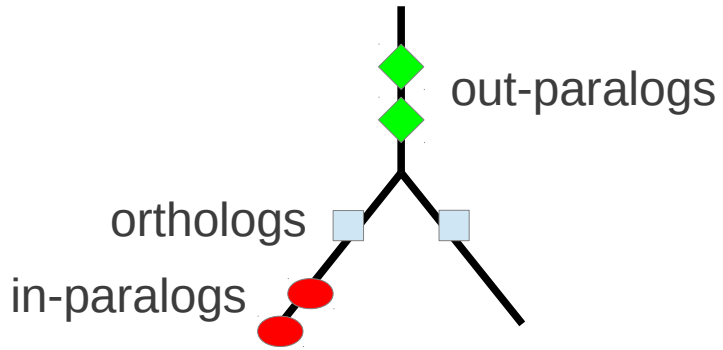
The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data

Chen X, Zhang J (2012) The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data. PLoS Comput Biol 8(11): e1002784. doi:10.1371/journal.pcbi.1002784

Patrick Tan

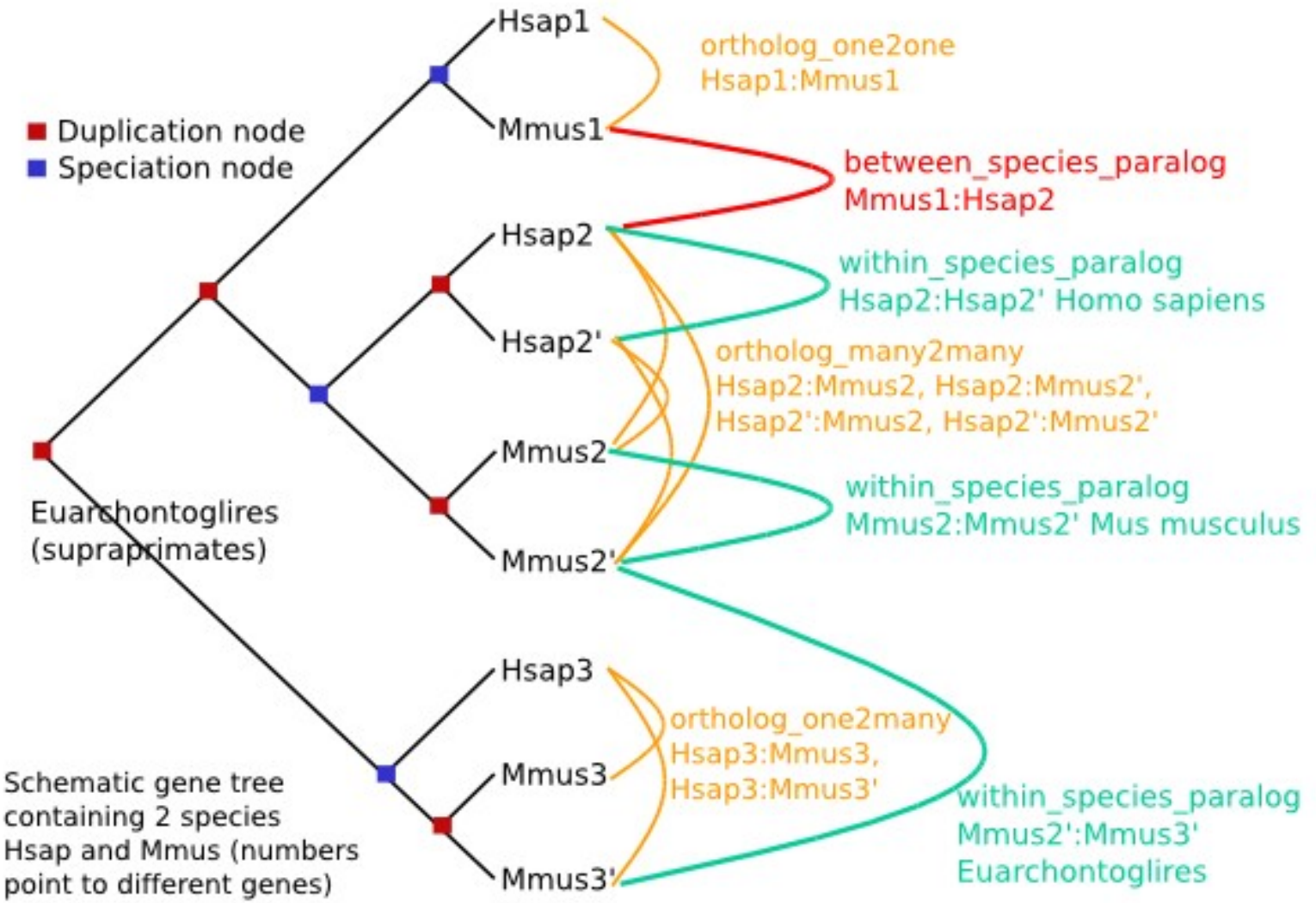
Homology 101

- homology – shared ancestry
- ortholog – genes differentiated by a speciation event; usually have the same function (let's assume one-to-one mapping) e.g. mouse and human alpha hemoglobin
- paralog – genes differentiated by a duplication event; can be within or between species; may have the same or new function; e.g. human alpha and beta hemoglobin
  - out-paralogs – gene duplication that happened **before** the speciation event
  - in-paralogs – gene duplication that happened **after** the speciation event
- Useful for predicting gene function



http://en.wikipedia.org/wiki/Homology_(biology)#Sequence_homology

EnsEMBL Homology types

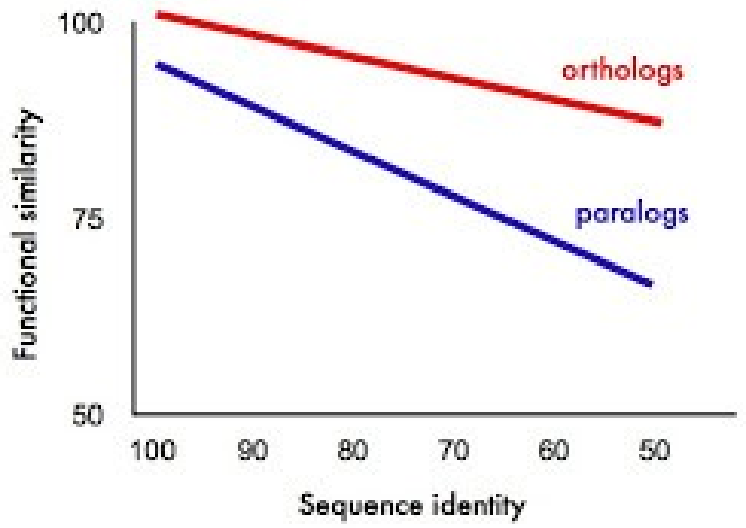


http://uswest.ensembl.org/info/docs/compara/homology_method.html

EnsemblCompara GeneTrees: Analysis of complete, duplication aware phylogenetic trees in vertebrates. Vilella AJ, Severin J, Ureta-Vidal A, Durbin R, Heng L, Birney E. Genome Research 2008 Nov 4.

Ortholog conjecture

- orthologs are assumed to be more functionally similar than paralogs
- stems from evolutionary biologists, there's still no hard evidence yet



*http://phylogenomics.blogspot.ca/2011/09/special-guest-post-discussion.html

Overview

- A recent paper (Nehrt et al. 2011) challenges the ortholog conjecture
  - used Gene Ontology (GO) and microarray expression and found that paralogs are more similar in function than orthologs (for the same level of protein sequence divergence)*
  - the cellular context (ie. the genome that it is found in) drives evolution function
- Chen and Zhang
  - argued that GO has annotation errors and experimental biases (reviewers of Nehrt said something similar but convinced PLoS with the microarray data)
  - used RNA-seq instead of microarray to test the ortholog conjecture

Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. PLoS Comput Biol 7(6): e1002073. doi:10.1371/journal.pcbi.1002073

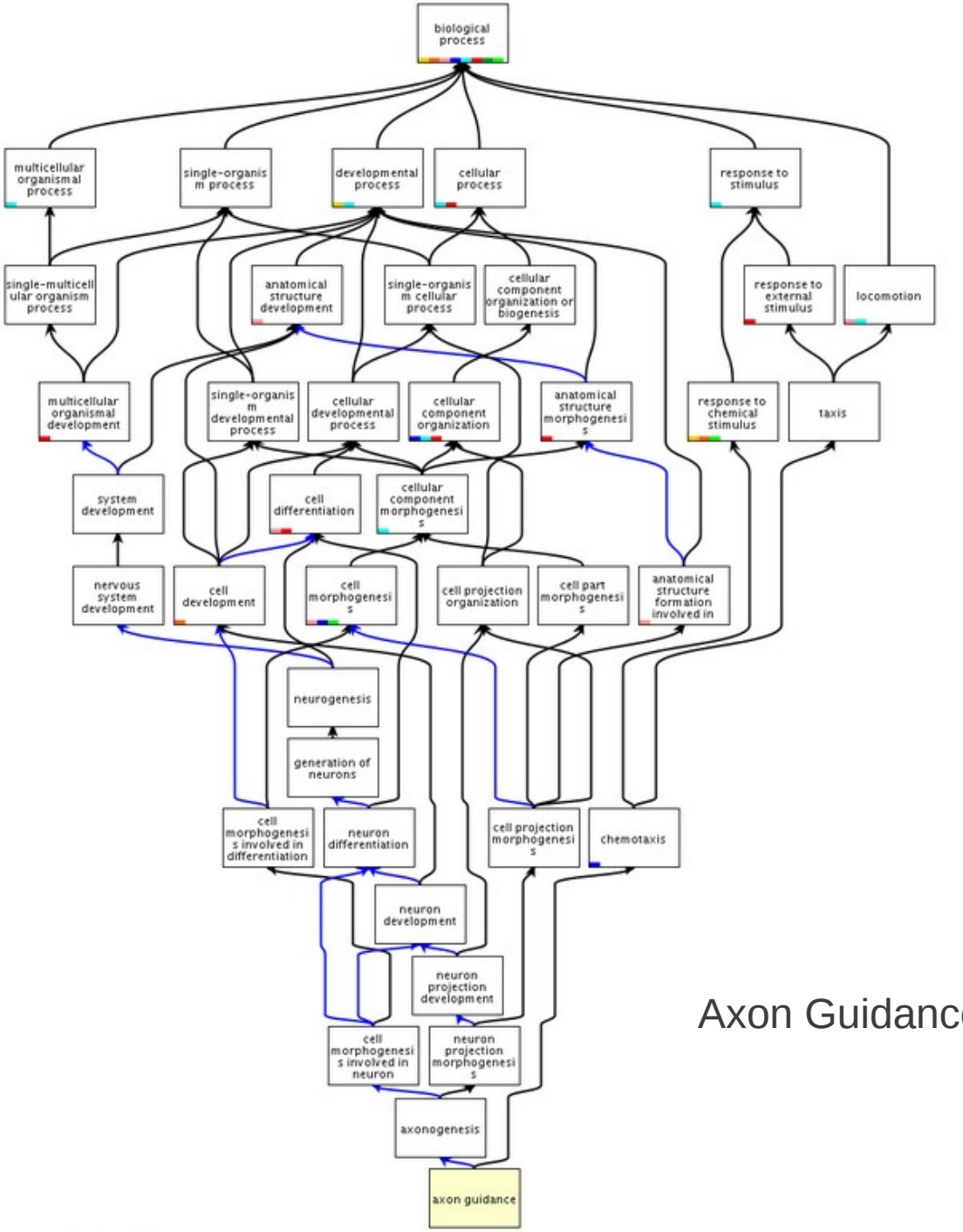
*http://phylogenomics.blogspot.ca/2011/09/special-guest-post-discussion.html

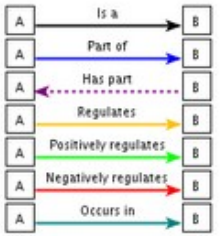
Gene Ontology (GO)

- structured and controlled vocabulary of gene annotation
- 3 parts:
  - molecular function
  - biological process
  - cellular component
- manually curated (and some inferred)
- changes over time
  - some terms are added / removed

The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat. Genet.. May 2000;25(1):25-9.

www.geneontology.org/





Axon Guidance GO Biological Process

The problems with GO

Method:

- paralog and ortholog information, protein sequence identity, most recent common ancestor were all downloaded from EnsEMBL
- ~16k ortholog pairs (1:1?), ~56k inparalog pairs, ~233K outparalog pairs between mouse and human in EnsEMBL
- ~1k ortholog pairs, ~100 inparalog pairs, ~1k outparalog pairs in GO time series
- ~1k ortholog pairs, ~200 inparalog pairs, ~5k outparalog pairs were co-studied

Figure 1. GO-based functional similarities of orthologs and paralogs vary in the last five years.

functional similarity = propagated GO term overlap

orthologs have become more similar

fraction = overlap / total unique

Due to previous underreporting of orthologs

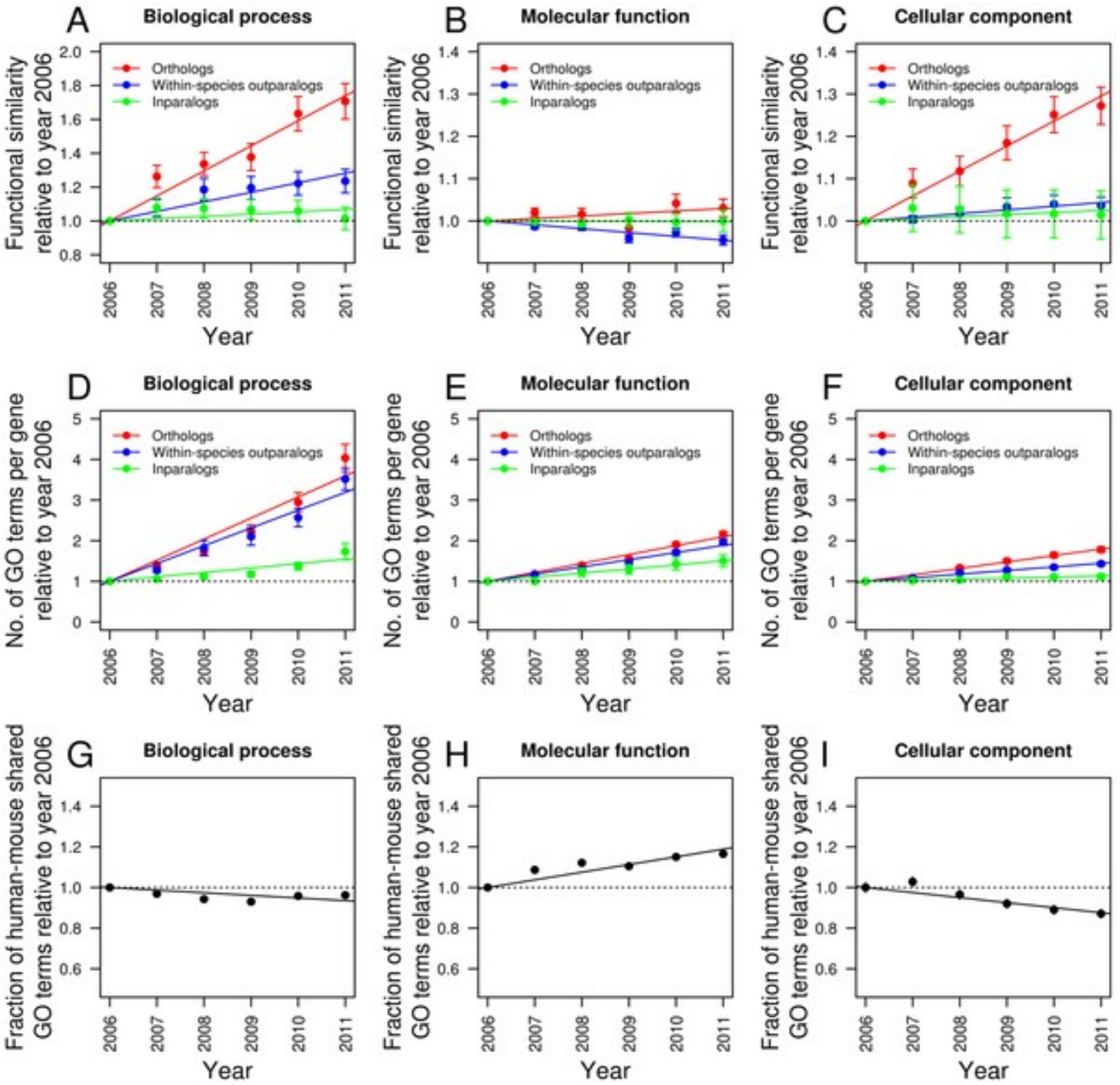


Figure 2. Biases in co-study papers that result in underestimation of functional similarity of ortholog, compared with paralogs.

co-study = homologous pairs share at least one PubMed ID

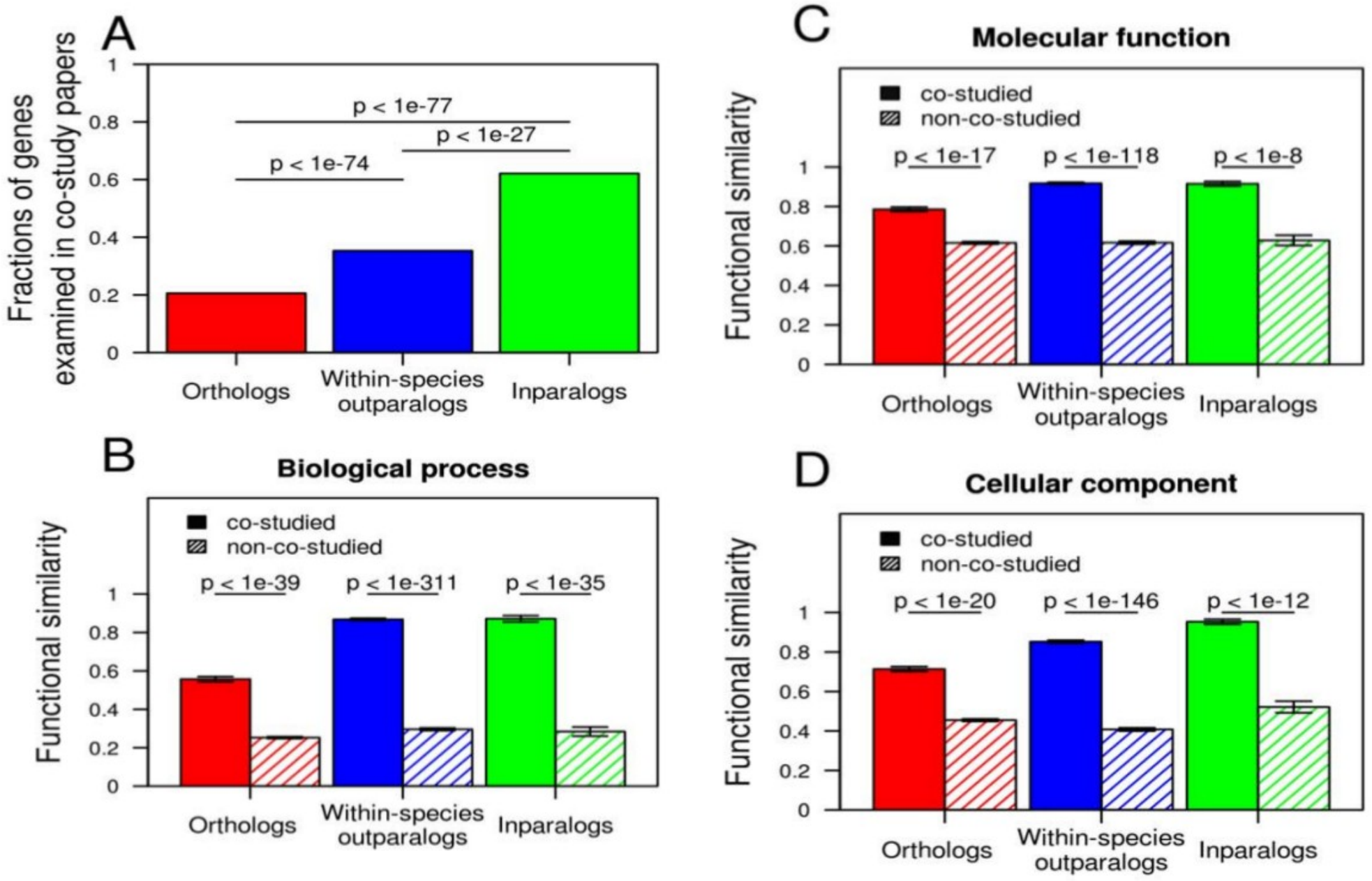


Table 1. Eight pairs of human-mouse orthologs with identical protein sequences but no overlapping GO annotations based on co-study papers.

One of the eight pairs

PubMed ID	Ensembl gene ID	GO term	GO category	GO term description	Experimental systems[1]	Bias/error[2]
PMID:11595183	ENSG00000155849	GO:0005886	Cellular component	plasma membrane	Hamster	Annotation error
		GO:0005737	Cellular component	cytoplasm	Hamster	Experimental bias
		GO:0016601	Biological process	Rac protein signal transduction	Hamster	Inferred function
		GO:0006928	Biological process	cellular component movement	Human, hamster	
		GO:0006911	Biological process	phagocytosis, engulfment	Human, hamster	
		GO:0030036	Biological process	actin cytoskeleton organization	Hamster	
	ENSMUSG00000041112	GO:0030029	Biological process	actin filament-based process	Hamster	
		GO:0006909	Biological process	phagocytosis	Hamster	
....						

[1] Experimental systems may mean organisms or cell lines.

[2] Bias/error indicates (1) experimental bias (i.e., different functional aspects were examined for orthologs), (2) inferred function (i.e., function in a species is inferred from that in a related species), or (3) annotation error (i.e., mistake in annotation).

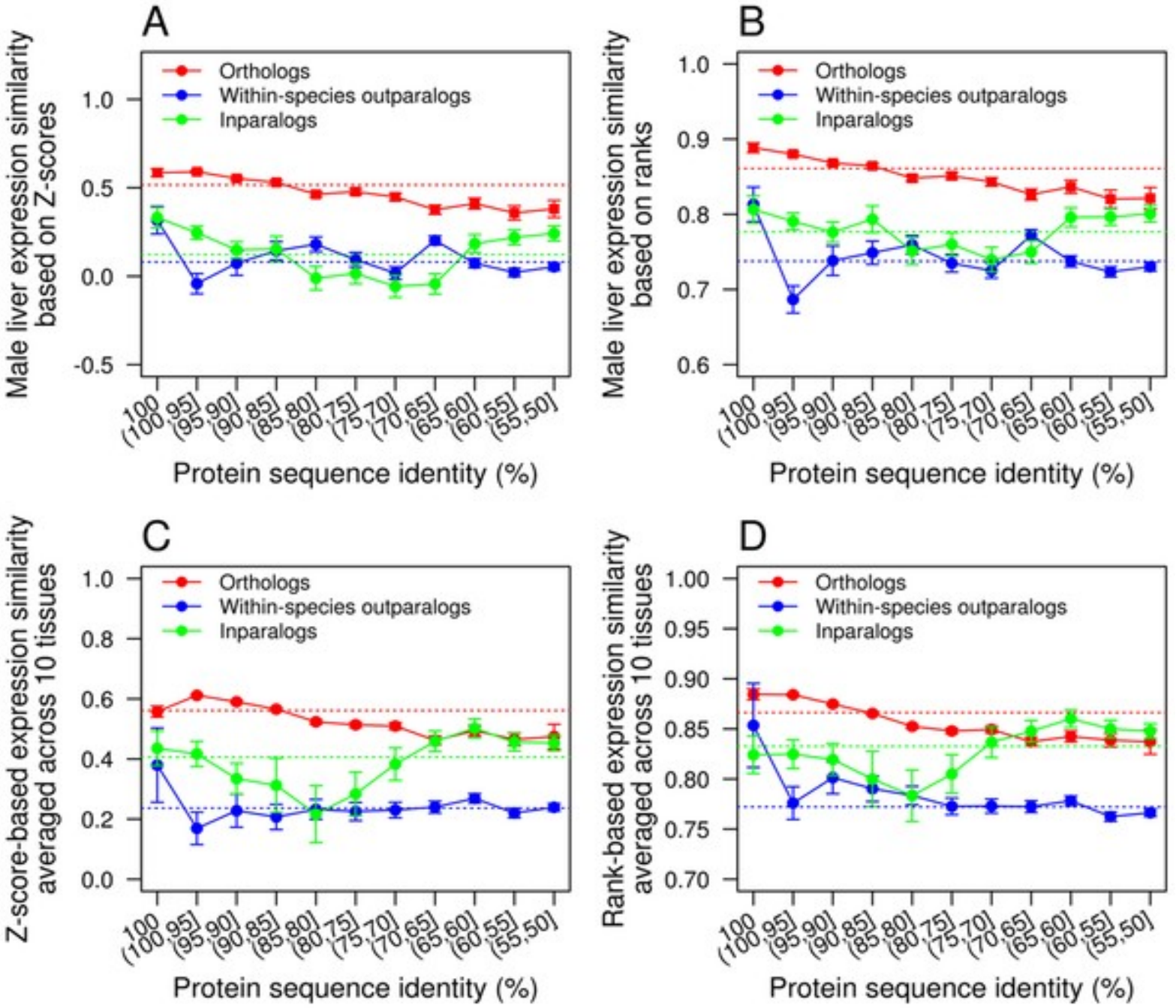
doi:10.1371/journal.pcbi.1002784.t001

RNA-seq expression comparisons

Method:

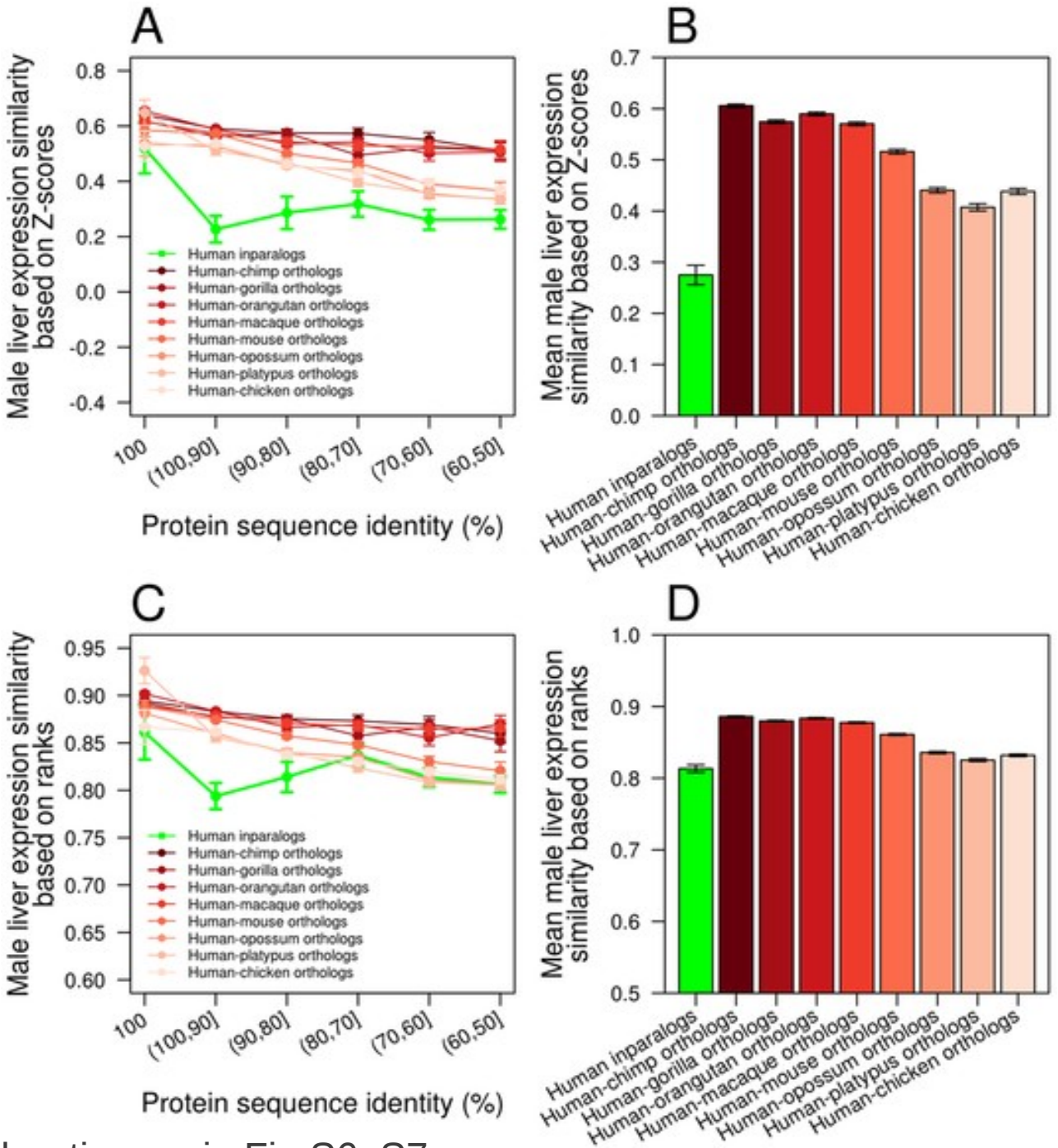
- 10 tissues from human, mouse and other species (Brawand et al. 2012)
- log2(RPKM) values were converted to Z-score (reads per kilobase per million mapped reads) and also to ranks

Figure 3. Expression similarity of homologous genes.



Similar trends in Fig S3, S4 for other tissues

Figure 4. Male liver expression similarity of homologous genes from multiple species.



pretty high values compared to z-scores?

Same trend for other tissues in Fig S6, S7

Figure 5. Z-score-based male liver expressions are more similar between human-mouse orthologs than between within-species paralogs.

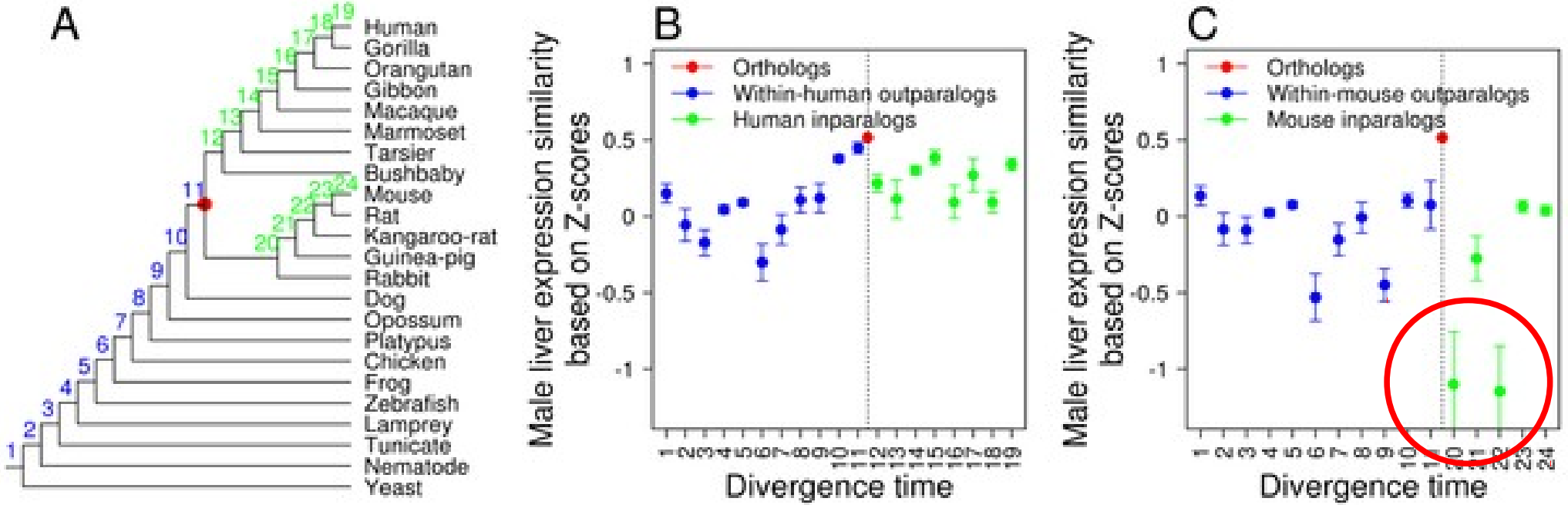
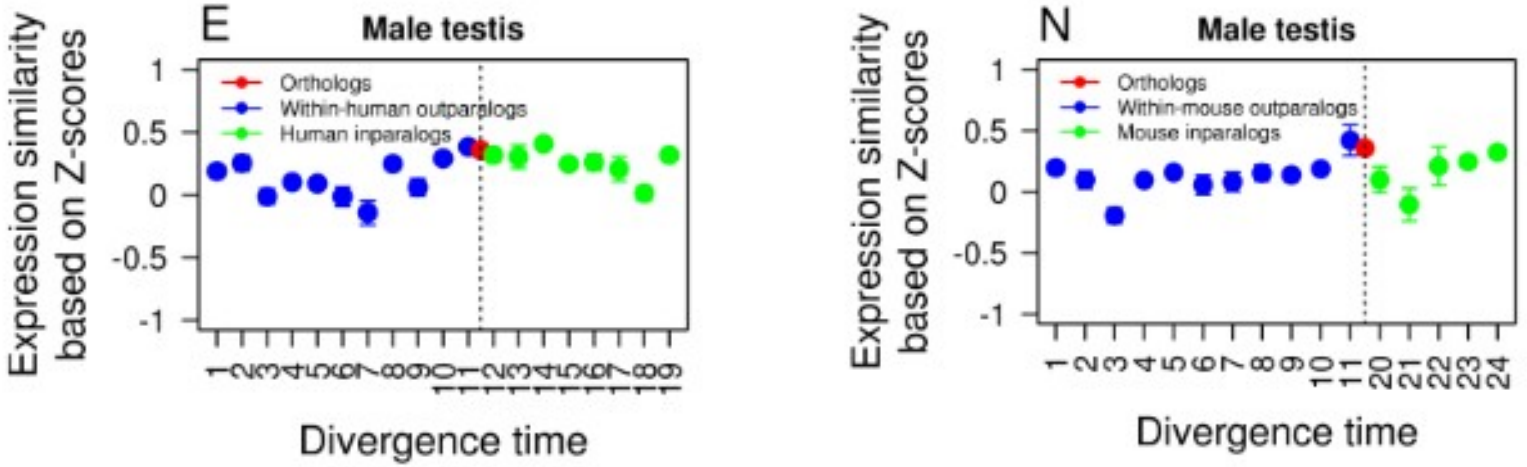


Figure S8 – Other tissues



Summary

- Reasons why GO can't be used to test the ortholog conjecture
  - Orthologous genes are under-represented in GO relative to paralogs (because people accept them and have no reason to report?)
  - Circular conjecture – inferred GO from other species
  - Annotation errors ie. annotating with the wrong species
- Higher similarity in paralogs in Nehrt et al. (context specific hypothesis) may be due to species specific platform differences in microarrays

Discussion points

- Analyse at the level of protein domains? Each protein family is unique? Some protein families are just more conserved than others
- Gene expression <> protein (examples of structurally similar proteins with low sequence similarity and vice-versa) (Krissinel 2007)
  - core residues are more conserved
  - mutations in amino acids important to structure may have a dramatic effect
- Genes with similar function but not in sequence

Bioinformatics (2007) 23 (6): 717-723.doi:10.1093/bioinformatics/btm006

Supplement

Figure S1.

GO-based functional similarities of orthologs and paralogs of different years, relative to those in 2006, in (A) biological process, (B) molecular function, and (C) cellular component. This figure is identical to Fig. 1A–C, except that we randomly sample equal numbers of orthologs and outparalogs as that of inparalogs. The averages of 1000 replications of the random sampling are presented.

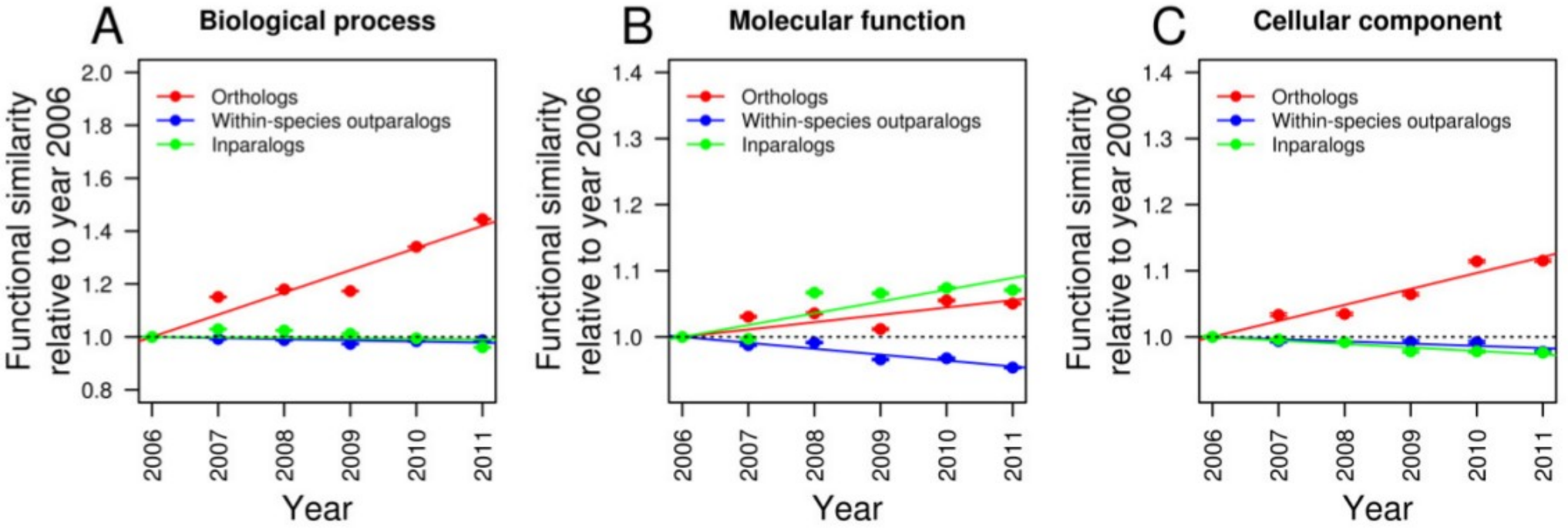


Figure S1

Gene Orthology/Paralogy prediction method

The gene orthology and paralogy prediction pipeline has 6 basic steps:

1) Load a representative translation of each gene from all species used in Ensembl. We currently choose the longest translation annotated by the CCDS project, if any, or the longest protein-coding translation otherwise.
2) run WUBlastp+SmithWaterman of every gene against every other (both self and non-self species) in a genome-wise manner
3) Build a sparse graph of gene relations based on Blast scores and generate clusters using hcluster_sg1
4) For each cluster, build a multiple alignment based on the protein sequences using a combination of multiple aligners, consensified by M-Coffee2
5) For each aligned cluster, build a phylogenetic tree using TreeBeST3 using the CDS back-translation of the protein multiple alignment from the original DNA sequences**. A rooted tree with internal duplication tags is obtained at this stage, reconciling it with the species tree in ensembl-compara/scripts/pipeline/species_tree_njtree.taxon_id.nh (refer to section "Create the species tree file" in ensembl-compara/scripts/pipeline/README-genetree for a more detailed explanation).
6) From each gene tree, infer gene pairwise relations of orthology* and paralogy types.

http://uswest.ensembl.org/info/docs/compara/homology_method.html

EnsemblCompara GeneTrees: Analysis of complete, duplication aware phylogenetic trees in vertebrates. Vilella AJ, Severin J, Ureta-Vidal A, Durbin R, Heng L, Birney E. Genome Research 2008 Nov 4.