Phylogenetic Identification and Functional Characterization of Orthologs and Paralogs across Human, Mouse, Fly, and Worm – Supplementary Material

Yi-Chieh Wu, Mukul S. Bansal, Matthew D. Rasmussen, Javier Herrero, Manolis Kellis

Supplemental Figures

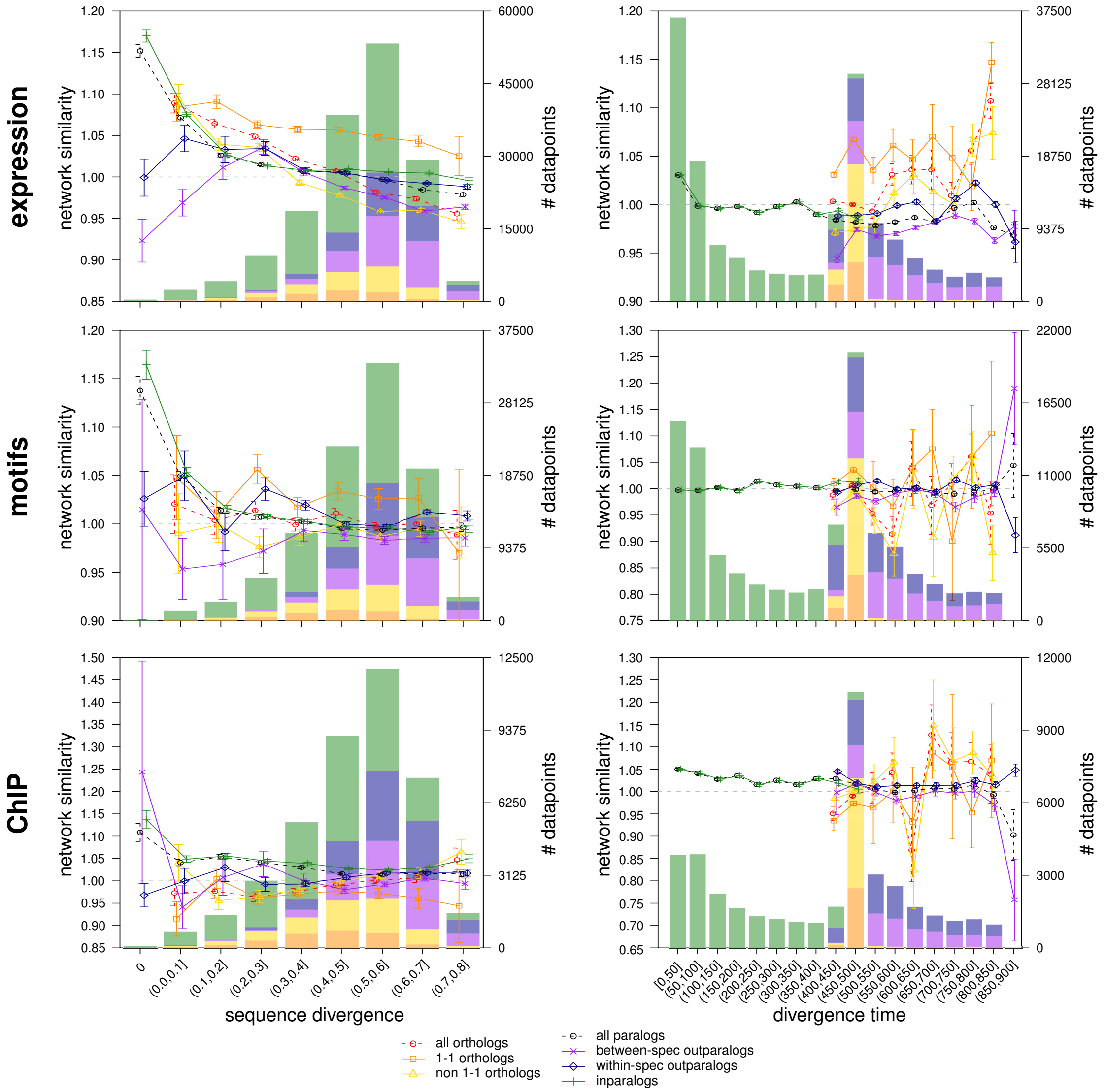

Figure S1. Network similarity of different types of homologs. See Figure 3 for details.

Supplemental Tables

Table S1. Species tree parameters.

species	parameter[a]	value	reference
human–mouse	t	75	Mouse Genome Sequencing Consortium et al. (2002)
human	N_e	$10,400$	Yu et al. (2004)
human	g	20	Rasmussen and Kellis (2012)
mouse	N_e	$460,000$	Piganeau and Eyre-Walker (2009)
mouse	g	0.5	Piganeau and Eyre-Walker (2009)
fly[b]	topology	-	Tamura et al. (2004)
fly root	t	62	Tamura et al. (2004)
D. melanogaster	N_e	10^7	Shapiro et al. (2007)
D. melanogaster	g	0.1	Sawyer and Hartl (1992)
worm[c]	topology	-	Kiontke et al. (2004)
C. elegans–C. briggsae	t	18.6	Cutter (2008)
C. elegans	N_e	$50,000$	Rockman and Kruglyak (2009)
C. remanei	N_e	10^6	Hillier et al. (2007)
C. elegans	g	1/6	Cutter (2008)
outgroup			
S. cerevisiae[d]	N_e	10^7	Tsai et al. (2008)
S. cerevisiae[d]	g	0.9	Rasmussen and Kellis (2012)
all species	λ, μ	0.002, 0.002	See footnote[e]

[a] t: divergence time (myr); N_e: effective population size; g: generation time (yr); λ, μ: duplication, loss rate (events/site/myr)

[b] Parameters were assumed to be the same across all fly species.

[c] Population size for *C. elegans* was propagated to other hermaphroditic species (*C. briggsae*), and population size for *C. remanei* was propagated to other dioecious species (*C. brenneri*, *C. japonica*). Generation time was assumed to be the same across all worm species.

[d] Parameters were estimated from *S. paradoxus*.

[e] Parameters were estimated using the procedure of Rasmussen and Kellis (2011).

[f] For branches outside of the major clades, population size and generation time were calculated by taking the average of the respective parameter for its children.

Table S2. Statistics for the different homolog subtypes.

homolog class	# of homologs[a]	mean[b]	median[b]	p-value[c]	ρ[c]
all orthologs	23,518	1.010	0.990	3.579×10^{-32}	-0.077
one-to-one orthologs	7,769	1.039	1.021	3.023×10^{-1}	-0.012
non-one-to-one orthologs	15,749	0.995	0.975	1.495×10^{-10}	-0.051
all paralogs	140,191	1.006	1.003	2.515×10^{-233}	-0.087
inparalogs	88,186	1.013	1.013	2.033×10^{-66}	-0.058
within-species outparalogs	23,422	1.008	1.008	2.296×10^{-5}	-0.028
between-species outparalogs	28,583	0.981	0.965	9.864×10^{-9}	-0.034

[a] Only homologs with sequence divergence ≤ 0.8 are retained.

[b] Mean and median network similarity.

[c] Correlation test, with p-value and Spearman's correlation coefficient shown.

Table S3. Significance test for difference in network similarity between orthologs and paralogs, binned by sequence divergence.

sequence divergence	% difference[a]	p-value$_{\text{orthologs}>\text{paralogs}}$[b]	p-value$_{\text{orthologs}<\text{paralogs}}$[b]
$[0.0, 0.1]$	4.672 [6.706]	1	1.342×10^{-6}
$[0.1, 0.2]$	0.255 [1.041]	0.861	1.394×10^{-1}
$[0.2, 0.3]$	1.224 [0.606]	0.155	8.452×10^{-1}
$[0.3, 0.4]$	0.163 [2.195]	1	3.073×10^{-6}
$[0.4, 0.5]$	0.590 [1.047]	0.966	3.382×10^{-2}
$[0.5, 0.6]$	0.172 [1.221]	1	3.339×10^{-4}
$[0.6, 0.7]$	0.024 [1.227]	0.996	4.296×10^{-3}
$[0.7, 0.8]$	0.103 [1.712]	0.888	1.119×10^{-1}

[a] Percent difference between orthologs and paralogs in mean [median] network similarity.

[b] P-values based on one-tailed Mann-Whitney tests.

Table S4. Significance test for difference in network similarity between orthologs and paralogs, binned by divergence time.

divergence time	% difference[a]	p-value$_{\text{orthologs}>\text{paralogs}}$[b]	p-value$_{\text{orthologs}<\text{paralogs}}$[b]
$[400, 450]$	1.520 [0.406]	1.570×10^{-6}	1
$[450, 500]$	1.492 [0.116]	3.757×10^{-9}	1
$[500, 550]$	1.526 [0.630]	2.978×10^{-2}	0.970
$[550, 600]$	4.751 [2.935]	3.561×10^{-6}	1
$[600, 650]$	5.207 [2.528]	3.592×10^{-4}	1
$[650, 700]$	5.745 [4.480]	1.057×10^{-3}	0.999
$[700, 750]$	3.728 [2.587]	2.074×10^{-4}	1
$[750, 800]$	1.620 [2.416]	9.773×10^{-1}	0.023
$[800, 850]$	5.471 [5.631]	3.461×10^{-4}	1

[a] Percent difference between orthologs and paralogs in mean [median] network similarity.

[b] P-values based on one-tailed Mann-Whitney tests.

Table S5. Significance test for difference in network similarity between homolog subtypes.[a]

	1-1 orthologs	non 1-1 orthologs	inparalogs	ws outparalogs	bs outparalogs
1-1 orthologs	–	4.335 [4.647] 5.7979×10^{-151} (>)	2.521 [0.834] 2.516×10^{-32} (>)	3.083 [1.320] 8.391×10^{-47} (>)	5.698 [5.726] 9.827×10^{-302} (>)
non 1-1 orthologs		–	1.815 [3.813] 1.255×10^{-154} (<)	1.252 [3.328] 2.065×10^{-78} (<)	1.364 [1.080] 4.935×10^{-26} (>)
inparalogs			–	0.563 [0.486] 2.14×10^{-11} (>)	3.179 [4.893] $< 2.225 \times 10^{-308}$ (>)
ws outparalogs				–	2.616 [4.407] 1.276×10^{-272} (>)
bs outparalogs					–

[a] In the first row of each cell are percent differences in mean [median] network similarity. In the second row of each cell are p-values based on one-tailed Mann-Whitney tests, with the alternative hypothesis shown in parenthesis; that is, > (<) tests the alternative hypothesis that the row header is more (less) similar than the column header.

References

Cutter A. D. 2008. Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. Mol Biol Evol **25**:778–786.

Hillier D. W, Miller R. D, Baird S. E, Chinwalla A, Fulton L. A, Koboldt D. C and Waterston R. H. 2007. Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. PLoS Biol **5**:e167.

Kiontke K, Gavin N. P, Raynes Y, Roehrig C, Piano F and Fitch D. H. A. 2004. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. PNAS **101**:9003–9008.

Mouse Genome Sequencing Consortium, Chinwalla A. T, Cook L. L, Delehaunty K. D, et al. (319 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520–562.

Piganeau G and Eyre-Walker A. 2009. Evidence for variation in the effective population size of animal mitochondrial dna. PLoS One **4**:e4396.

Rasmussen M. D and Kellis M. 2011. A Bayesian approach for fast and accurate gene tree reconstruction. Mol Biol Evol **28**:273–290.

Rasmussen M. D and Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res **22**:755–765.

Rockman M. V and Kruglyak L. 2009. Recombinational landscape and population genomics of *Caenorhabditis Elegans*. PLoS Genet **5**:e1000419.

Sawyer S. A and Hartl D. L. 1992. Population genetics of polymorphism and divergence. Genetics **132**:1161–1176.

Shapiro J. A, Huang W, Zhang C, Hubisz M. J, et al. (13 co-authors). 2007. Adaptive genic evolution in the *Drosophila* genomes. PNAS **104**:2271–2276.

Tamura K, Subramanian S and Kumar S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol Biol Evol **21**:36–44.

Tsai I. J, Bensasson D, Burt A and Koufopanou V. 2008. Population genomics of the wild yeast *saccharomyces paradoxus*: Quantifying the life cycle. PNAS **105**:4957–4962.

Yu N, Jensen-Seaman M. I, Chemnick L, Ryder O and Li W.-H. 2004. Nucleotide diversity in gorillas. Genetics **166**:1375–1383.