RESEARCH

Non-synonymous to synonymous substitutions suggest that orthologs tend to keep their functions, while paralogs are a source of functional novelty

Juan Miguel Escorcia-Rodríguez[2], Mario Esposito[1], Julio Augusto Freyre-González[2] and Gabriel Moreno-Hagelsieb[1*]

*Correspondence: gmorenohagelsieb@wlu.ca
[3]Wilfrid Laurier University, Department of Biology, 75 University Ave W, N2L 3C5 Waterloo, ON, Canada
Full list of author information is available at the end of the article

Abstract

Background: Orthologs diverge after speciation events and paralogs after gene duplication. It is thus expected that orthologs would tend to keep their functions, while paralogs could be a source of new functions. Because protein functional divergence follows from non-synonymous substitutions, we performed an analysis based on the ratio of non-synonymous to synonymous substitutions (dN/dS) as proxy for functional divergence. We used four working definitions of orthology, including reciprocal best hits (RBH), among other definitions based on network analyses and clustering.

Results: The results showed that orthologs, by all definitions tested, had values of dN/dS noticeably lower than those of paralogs, not only suggesting that orthologs keep their functions better, but also that paralogs are a readily source of functional novelty. The differences in dN/dS ratios remained favouring the functional stability of orthologs after eliminating gene comparisons with potential problems, such as genes having a high codon usage bias, low coverage of either of the aligned sequences, or sequences with very high similarities. The dN/dS ratios kept suggesting better functional stability of orthologs regardless of overall sequence divergence. Separating orthologs and paralogs into groups with similar overall substitution rates kept showing dN/dS differences favouring the functional stability of orthologs over that of paralogs.

Availability: A couple of programs for obtaining orthologs and dN/dS values as tested in this manuscript are available at github: https://github.com/Computational-conSequences/SequenceTools.

Keywords: orthologs; paralogs; functional divergence; non-synonymous to synonymous substitution; dN/dS; function

Background

In this report, we present an analysis of non-synonymous and synonymous substitutions (dN/dS) what suggests that orthologs keep their functions better than paralogs.

Since the beginning of comparative genomics, the assumption has been made that orthologs should be expected to conserve their functions more often than paralogs [1, 2, 3, 4]. The expectation is based on the definitions of each homolog type: orthologs are characters diverging after a speciation event, while paralogs are characters diverging after a duplication event [5]. Given those definitions, orthologs could be considered the "same" genes in different species, while paralogy has been

proposed as a mechanism for the evolution of new functions under the assumption that, since one of the copies can perform the original function, the other copy would have some freedom to functionally diverge [6]. This does not mean that functional divergence cannot happen between orthologs. However, it is very hard to think of a scenario whereby orthologs would diverge in functions at a higher rate than paralogs. Therefore, it has been customary to use some working definition for orthology to infer the genes whose products most likely perform the same functions across different species [1, 2, 3, 4, 7].

Despite such an obvious expectation, a report was published making the surprising claim that orthologs diverged in function more than paralogs [8]. The article was mainly based on the examination of Gene Ontology annotations of orthologs and paralogs from two species: humans and mice. Again, it is very hard to imagine a scenario where orthologs, as a group, would diverge in functions more than paralogs. At worst, we could expect equal functional divergences. If the report were correct, it would mean mice myoglobin could be performing the function that human alpha-haemoglobin performs. How could such a thing happen? How could paralogs exchange functions with so much freedom?

Later work showed that the article by Nehrt et all [8] was in error. For example, some work reported that Gene Ontologies suffered from "ascertainment bias," which made annotations more consistent within an organism than without [9, 10]. These publications also proposed solutions to such and other problems [9, 10]. Another work showed that gene expression data supported the idea that orthologs keep their functions better than paralogs [11].

Most of the publications we have found focused on gene annotations and gene expression in Eukaryotes. We thus wondered whether we could perform some analyses that did not suffer from annotator bias, and that could cover most of the homologs found between any pair of genomes. Not only that, we also wanted to analyze prokaryotes (Bacteria and Archaea). We thus thought about performing analyses of non-synonymous to synonymous substitutions (dN/dS), which compare the relative strengths of positive and negative (purifying) selection [12, 13]. Since changes in function require changes in amino-acids, it can be expected that the most functionally stable homologs will have lower dN/dS ratios compared to less functionally stable homologs. Thus, comparing the dN/dS distributions of orthologs and paralogs could indicate whether either group has a stronger tendency to conserve their functions.

Materials and methods

Genome Data

We downloaded the analyzed genomes from NCBI's RefSeq Genome database [14]. We performed our analyses by selecting genomes from three taxonomic classes, using one genome within each order as a query genome (Table 1): *Escherichia coli* K12 MG1655 (class Gammaproteobacteria, domain Bacteria), *Bacillus subtilis* 168 (Bacilli, Bacteria), and *Pyrococcus furiosus* COM1 (Thermococci, Archaea).

Orthologs as reciprocal best hits (RBH)

We compared the proteomes of each of these genomes against those of other members of their taxonomic order using BLASTP [15], with a maximum e-value of

Table 1 **Genomes used in this study.**

Genome ID	Taxonomic Order	Strain
Domain Bacteria, Class Gammaproteobacteria		
GCF_000005845	Enterobacterales	*Escherichia coli* K-12 MG1655
GCF_001989635	Enterobacterales	*Salmonella enterica* Typhimurium 81741
GCF_000695935	Enterobacterales	*Klebsiella pneumoniae* KPNIH27
GCF_000834215	Enterobacterales	*Yersinia frederiksenii* Y225
GCF_002591155	Enterobacterales	*Proteus vulgaris* FDAARGOS366
GCF_000017245	Pasteurellales	*Actinobacillus succinogenes* 130Z
GCF_001558255	Vibrionales	*Grimontia hollisae* ATCC33564
GCF_002157895	Aeromonadales	*Oceanisphaera profunda* SM1222
GCF_000013765	Alteromonadales	*Shewanella denitrificans* OS217
GCF_001654435	Pseudomonadales	*Pseudomonas citronellolis* SJTE-3
GCF_000021985	Chromatiales	*Thioalkalivibrio sulfidiphilus* HL-EbGr7
Domain Bacteria, Class Bacilli		
GCF_000009045	Bacillales	*Bacillus subtilis* 168
GCF_000284395	Bacillales	*Bacillus velezensis* YAU B9601-Y2
GCF_002243625	Bacillales	*Aeribacillus pallidus* KCTC3564
GCF_001274575	Bacillales	*Geobacillus stearothermophilus* 10
GCF_000299435	Bacillales	*Exiguobacterium antarcticum* B7 Eab7
GCF_002442895	Bacillales	*Staphylococcus nepalensis* JS1
GCF_000283615	Lactobacillales	*Tetragenococcus halophilus* NBRC12172
GCF_001543285	Lactobacillales	*Aerococcus viridans* CCUG4311
GCF_000246835	Lactobacillales	*Streptococcus infantarius* CJ18
GCF_001543145	Lactobacillales	*Aerococcus sanguinicola* CCUG43001
GCF_002192215	Lactobacillales	*Lactobacillus casei* LC5
Domain Archaea, Class Thermococci		
GCF_000275605	Thermococcales	*Pyrococcus furiosus* COM1
GCF_001577775	Thermococcales	*Pyrococcus kukulkanii* NCB100
GCF_000246985	Thermococcales	*Thermococcus litoralis* DSM5473
GCF_000009965	Thermococcales	*Thermococcus kodakarensis* KOD1
GCF_001433455	Thermococcales	*Thermococcus barophilus* CH5
GCF_001647085	Thermococcales	*Thermococcus piezophilus* CDGS
GCF_002214505	Thermococcales	*Thermococcus siculi* RG-20
GCF_000585495	Thermococcales	*Thermococcus nautili* 30-1
GCF_002214585	Thermococcales	*Thermococcus profundus* DT5432
GCF_002214545	Thermococcales	*Thermococcus thioreducens* OGL-20P
GCF_000769655	Thermococcales	*Thermococcus eurythermalis* A501

1×10^{-6} (*-evalue 1e-6*), soft-masking (*-seg yes -soft_masking true*), a Smith-Waterman final alignment (*-use_sw_tback*), and minimal alignment coverage of 60% of the shortest sequence. Orthologs were defined as reciprocal best hits (RBHs) as described previously [16, 17]. Except where noted, paralogs were all BLASTP matches left after finding RBHs.

OrthoFinder

OrthoFinder defines an orthogroup as the set of genes derived from a single gene in the last common ancestor of the all species under consideration [18]. First, OrthoFinder performs all-versus-all blastp comparisons and uses an e-value of 1×10^{-3} as a threshold. Then, it normalizes the gene length and phylogenetic distance of the BLAST bit scores. It uses the lowest normalized value of the RBHs for either gene in a gene pair as the threshold for their inclusion in an orthogroup. Finally, it weights the orthogroup graph with the normalized bit scores and clusters it using MCL. OrthoFinder outputs the orthogroups and orthology relationships, which can be many to many (co-orthology).

We ran OrthoFinder with all the proteomes for each of the taxonomic groups listed (Table 1). From the OrthoFinder outputs with the orthology relationships between every two species, we used those considering the query organism. From an orthogroup containing one or more orthology relationships, we identified the

out-paralogs as those genes belonging to the same orthogroup but not to the same orthology relationship. We identified the inparalogs for the query organisms as its genes belonging to the same orthogroup since they derived from a single ancestor gene.

ProteinOrtho

ProteinOrtho is a graph-based tool that implements an extended version of the RBH heuristic and is intended for the identification of ortholog groups between many organisms [19]. First, from all-versus-all blast results, ProteinOrtho creates subnetworks using the RBH at the seed. Then, if the second best hit for each protein is almost as good as the RBH, it is added to the graph. The algorithm claims to recover false negatives and to avoid the inclusion of false positives [19].

We ran ProteinOrtho pairwise since we need to identify orthologs and paralogs between the query organisms and the other members of their taxonomic data, not those orthologs shared between all the organisms. ProteinOrtho run blast with the following parameters: an E-value cutoff of 1x10-6, minimal alignment coverage of 50% of the shortest sequence, and 25% of identity. We used the orthologs as the combinatorial of genes from a different genome that belongs to the same orthogroup, and in-paralogs as the combinatorial of genes from the same genome that belong to the same orthogroup. We rerun ProteinOrtho with a similarity value of 75 instead of 90, to identify out-paralogs as those interactions not identified in the first run.

InParanoid

InParanoid is a graph-based tool to identify orthologs and in-paralogs from pairwise sequence comparisons [20]. InParanoid first runs all-versus-all blast and identify the RBHs. Then, use the RBHs as seeds to identify co-orthologs for each gene (which the authors define as in-paralogs), proteins from the same organism that obtain better bits score than the RBH. Finally, through a series of rules, InParanoid cluster the co-orthologs to return non-overlapping groups. Furthermore, the authors define out-paralogs as those blast-hits outside of the co-ortholog clusters [20].

We ran InParanoid for each query genome against those of other members of their taxonomic order. InParanoid was run with the following parameters: double blast and 40 bits as score cutoff. The first pass run with compositional adjustment on and soft masking. This removes low complexity matches but truncates alignments [20]. The second pass run with compositional adjustment off to get full-length alignments. We used as in-paralogs the combinatorial of the genes of the same organism from the same cluster, and as out-paralogs those blast-hits outside of the co-ortholog clusters.

Non-synonymous to synonymous substitutions

To perform dN/dS estimates, we used the CODEML program from the PAML software suite [21]. The DNA alignments were derived from the protein sequence alignments using an *ad hoc* program written in PERL. The same program ran pairwise comparisons using CODEML to produce Bayesian estimates of dN/dS [22, 23]. The results were separated between ortholog and paralog pairs, and the density distributions were plotted using R [24]. Statistical analyses were also performed with R.

Codon Adaptation Index

To calculate the Codon Adaptation Index (CAI) [25], we used ribosomal proteins as representatives of highly expressed genes. To find ribosomal proteins we matched the COG ribosomal protein families described by Yutin *et al.* [26] to the proteins in the genomes under analysis using RPSBLAST (part of NCBI's BLAST+ suite) [15]. RPSBLAST was run with soft-masking (*-seg yes -soft_masking true*), a Smith-Waterman final alignment (*-use_sw_tback*), and a maximum e-value threshold of 1×10^{-3} (*-evalue 1e-3*). A minimum coverage of 60% of the COG domain model was required. To produce the codon usage tables of the ribosomal protein-coding genes, we used the program *cusp* from the EMBOSS software suite [27]. These codon usage tables were then used to calculate the CAI for each protein-coding gene within the appropriate genome using the *cai* program also from the EMBOSS software suite [27].

Results and Discussion

Reciprocal best hits have lower dN/dS ratios than paralogs

During the first stages of this research, we ran a few tests using other methods for estimating dN/dS, which showed promising results (with the very same tendencies as those presented in this report). However, we chose to present results using Bayesian estimates, because they are considered the most robust and accurate [22, 23]. To compare the distribution of dN/dS values between orthologs and paralogs, we plotted dN/dS density distributions using violin plots (Fig. 1). These plots demonstrated evident differences, with orthologs showing lower dN/dS ratios than paralogs, thus indicating that orthologs keep their functions better. Wilcoxon rank tests showed the differences to be statistically significant, with probabilities much lower than 1×10^{-9}. Since most comparative genomics work is done using reciprocal best hits (RBHs) as a working definition for orthology [28, 29], this result alone suggests that most research in comparative genomics has used the proteins/genes that most likely share their functions.

Differences in dN/dS resist other working definitions of orthology

A concern with our analyses might arise from our use of reciprocal best hits (RBHs) as a working definition of orthology. RBHs is arguably the most usual working definition of orthology [28, 29]. It is thus important to start these analyses with RBHs, at the very least to test whether it works for the purpose of inferring genes most likely to keep their functions. Also, analyses of the quality of RBHs for inferring orthology, based on synteny, show that RBH error rates tend to be lower than 95% [16, 28, 17]. Other analyses show that databases of orthology, based on RBHs, tend to contain a higher rate of false positives (paralogs mistaken for orthologs), than databases based on phylogenetic and network analyses [30]. This means that RBHs are mostly contaminated by paralogs, rather than paralogs with orthologs. Therefore, we can assume that orthologs dominate the RBHs dN/dS distributions.

Despite the above, we considered three other definitions of orthology (see methods). As shown using the comparison of *E. coli* K12 with *Yersinia frederiksenii* Y225 as an example (Fig. 2), orthologs obtained with different working definitions show very similar dN/dS rations, and all of those were lower than the values for paralogs (Fig. 2A). That the dN/dS values would be almost identical is explained by the high number of ortholog pairs shared by all ortholog sets (Fig. 2B).

Differences in dN/dS persist after testing for potential biases

While the tests above suggest that RBHs separate homologs with higher tendencies to preserve their functions from other homologs, we decided to test for some potential biases. A potential problem could arise from comparing proteins of very different lengths. We thus filtered the dN/dS results to keep those where the pairwise alignments covered at least 80% of the length of both proteins. The results showed shorted tails in both ortholog and paralog density distributions, but the tendency for orthologs to have lower dN/dS values remained (Fig. 3, Figure S1).

Another parameter that could bias the dN/dS results is high sequence similarity. In this case, the programs tend to produce high dN/dS rations. While we should expect this issue to have a larger effect on orthologs, we still filtered both groups, orthologs and paralogs, to contain proteins less than 70% identical. This filter had very little effect (Fig. 3, Figure S2).

To try and avoid the effect of sequences with unusual compositions, due, for example, to horizontal gene transfer, we filtered out sequences with unusual codon usages as measured using the Codon Adaptation Index (CAI) [25]. For this test, we eliminated sequences showing CAI values below the 15 percentile and above the 85 percentile of the respective genome's CAI distribution. After filtering, orthologs still exhibited dN/dS values below those of paralogs (Fig. 3, Figure S3).

Different models for background codon frequencies can also alter the dN/dS results [31]. Thus, we performed the same tests using the Muse and Gaut model for estimating background codon frequencies [32], as advised in [31]. Again, the results show orthologs to have lower dN/dS ratios than paralogs (Fig. 3, Figure S4).

Differences in dN/dS ratios remain regardless of overall sequence divergence

Orthologs will normally contain more similar proteins than paralogs. Thus, a similarity test alone would naturally make orthologs appear less divergent and, apparently, less likely to have evolved new functions. However, synonymous substitutions attest for the strength of purifying / negative selection in dN/dS analyses, making these ratios independent of the similarity between proteins. However, we still wondered whether more divergent sequences would show different tendencies to those found above.

To test whether dN/dS increases against sequence divergence, we separated orthologs and paralogs into ranges of divergence as measured by the t parameter calculated by CODEML (Fig. 4). The differences in dN/dS between orthologs and paralogs still suggested that orthologs keep their functions better.

Mouse/Human orthologs show lower dN/dS ratios than paralogs

Finally, we also performed dN/dS comparisons for human and mouse genes. The results were very similar to those obtained above (Fig. 5).

Conclusion

The results shown above use an objective measure of divergence that relates to the tendencies of sequences to diverge in amino-acid composition, against their tendencies to remain unchanged. Namely, the non-synonymous to synonymous substitution rates (dN/dS). Since changes in function require changes in amino-acids,

this measure might suggest which sequences have a higher tendency to keep their functions, which would show as a tendency towards lower dN/dS values. As should be expected from the conceptual definition of orthology, orthologs showed significantly lower values of dN/dS than paralogs. Thus, we can confidently conclude that orthologs keep their functions better than paralogs. It would also be proper to stop referring to the now confirmed expectation as a conjecture, since the expectation arises naturally from the definition. We did not expect the differences to be so obvious. Thus, our results also show that paralogs tend to acquire novel functions.

Ethical Approval and Consent to participate
Not applicable.

Consent for publication
The authors acknowledge that they consent to the publication of this manuscript.

Availability of supporting data
A couple of programs for obtaining orthologs and dN/dS values as tested in this manuscript are available at github: https://github.com/Computational-conSequences/SequenceTools.

Competing interests
The authors declare that they do not have any competing interest.

Funding
Funding: Work supported with a Discovery Grant to GM-H from the Natural Sciences and Engineering Research Council of Canada (NSERC). This work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT-UNAM) [IN205918 to JAF-G]. JME-R is supported by PhD fellowship 959406 from CONACyT-México.

Authors' contributions
JME-R contributed the description and prediction of orthologs by methods other than RBH. ME performed initial comparisons of dN/dS ratios between orthologs and paralogs. JAF-G directed the work by JME-R, checked manuscript. GMH designed the study, wrote the manuscript, wrote the programs available at github, and performed analyses.

Acknowledgements
We are grateful to Joe Bielawski for helpful advice. We thank The Shared Hierarchical Academic Research Computing Network (SHARCNET) for computing facilities.

Author details
[1]Wilfrid Laurier University, Department of Biology, 75 University Ave W, N2L 3C5 Waterloo, ON, Canada. [2]Center for Genomic Sciences, Universidad Nacional Autónoma de México (UNAM), Av. Universidad s/n, Col. Chamilpa, 62210 Cuernavaca, Morelos, Mexico. [3]Wilfrid Laurier University, Department of Biology, 75 University Ave W, N2L 3C5 Waterloo, ON, Canada.

References
1. Mushegian, A.R., Koonin, E.V.: Gene order is not conserved in bacterial evolution. Trends Genet **12**(8), 289–290 (1996)
2. Huynen, M.A., Bork, P.: Measuring genome evolution. Proc Natl Acad Sci USA **95**(11), 5849–5856 (1998)
3. Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., Yuan, Y.: Predicting function: from genes to genomes and back. J Mol Biol **283**(4), 707–725 (1998)
4. Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E.V.: The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res **28**(1), 33–36 (2000)
5. Fitch, W.M.: Homology a personal view on some of the problems. Trends Genet **16**(5), 227–231 (2000)
6. Ohno, S.: Evolution by Gene Duplication. Springer, ??? (1970)
7. Gabaldón, T., Koonin, E.V.: Functional and evolutionary implications of gene orthology. Nat Rev Genet **14**(5), 360–366 (2013)
8. Nehrt, N.L., Clark, W.T., Radivojac, P., Hahn, M.W.: Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol **7**(6), 1002073 (2011)
9. Thomas, P.D., Wood, V., Mungall, C.J., Lewis, S.E., Blake, J.A., Gene Ontology Consortium: On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. PLoS Comput Biol **8**(2), 1002386 (2012)
10. Altenhoff, A.M., Studer, R.A., Robinson-Rechavi, M., Dessimoz, C.: Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol **8**(5), 1002514 (2012)
11. Kryuchkova-Mostacci, N., Robinson-Rechavi, M.: Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. PLoS Comput Biol **12**(12), 1005274 (2016)
12. Ohta, T.: Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J Mol Evol **40**(1), 56–63 (1995)

13. Yang, Z., Nielsen, R.: Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17**(1), 32–43 (2000)
14. Haft, D.H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., Chitsaz, F., Derbyshire, M.K., Gonzales, N.R., Gwadz, M., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Yamashita, R.A., Zheng, C., Thibaud-Nissen, F., Geer, L.Y., Marchler-Bauer, A., Pruitt, K.D.: RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res **46**(D1), 851–860 (2018)
15. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L.: BLAST+: architecture and applications. BMC Bioinformatics **10**, 421 (2009)
16. Moreno-Hagelsieb, G., Latimer, K.: Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics **24**(3), 319–324 (2008)
17. Ward, N., Moreno-Hagelsieb, G.: Quickly Finding Orthologs as Reciprocal Best Hits with BLAT, LAST, and UBLAST: How Much Do We Miss? PLoS ONE **9**(7), 101850 (2014)
18. Emms, D.M., Kelly, S.: OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol **16**(1), 157 (2015)
19. Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P.F., Prohaska, S.J.: Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics **12**, 124 (2011)
20. Sonnhammer, E.L.L., Östlund, G.: InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res **43**(Database issue), 234–9 (2015)
21. Yang, Z.: PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24**(8), 1586–1591 (2007)
22. Anisimova, M., Bielawski, J.P., Yang, Z.: Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol **19**(6), 950–958 (2002)
23. Angelis, K., dos Reis, M., Yang, Z.: Bayesian estimation of nonsynonymous/synonymous rate ratios for pairwise sequence comparisons. Mol Biol Evol **31**(7), 1902–1913 (2014)
24. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2020). R Foundation for Statistical Computing. https://www.R-project.org/
25. Sharp, P.M., Li, W.H.: The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res **15**(3), 1281–1295 (1987)
26. Yutin, N., Puigbò, P., Koonin, E.V., Wolf, Y.I.: Phylogenomics of prokaryotic ribosomal proteins. PLoS ONE **7**(5), 36972 (2012)
27. Rice, P., Longden, I., Bleasby, A.: EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet **16**(6), 276–277 (2000)
28. Wolf, Y.I., Koonin, E.V.: A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol Evol **4**(12), 1286–1294 (2012)
29. Galperin, M.Y., Kristensen, D.M., Makarova, K.S., Wolf, Y.I., Koonin, E.V.: Microbial genome analysis: the COG approach. Brief Bioinformatics (2017)
30. Altenhoff, A.M., Dessimoz, C.: Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol **5**(1), 1000262 (2009)
31. Bielawski, J.P.: Detecting the signatures of adaptive evolution in protein-coding genes. Curr Protoc Mol Biol **Chapter 19**, 19–1 (2013)
32. Muse, S.V., Gaut, B.S.: A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol **11**(5), 715–724 (1994)
33. Goldman, N., Yang, Z.: A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11**(5), 725–736 (1994)

Figures

Additional Files

Figure S1

Full results on dN/dS ratios obtained with alignments covering at least 80% of both proteins. The results show orthologs to have lower dN/dS values than paralogs.

Figure S2

Full results on dN/dS ratios obtained with proteins no more that 70% identical. The results show orthologs to have lower dN/dS values than paralogs.

Figure S3

Full results on dN/dS rations obtained with proteins within usual codon usage. The Codon Adaptation Indexes [25] of the genes in the datasets had to be within the 15 and 85 percentile of the overall genomic CAI values. The results show orthologs to have lower dN/dS values than paralogs.

Figure S4

Full results on dN/dS rations obtained Muse and Gaut's estimate of background codon frequencies [32]. The results show orthologs to have lower dN/dS values than paralogs.

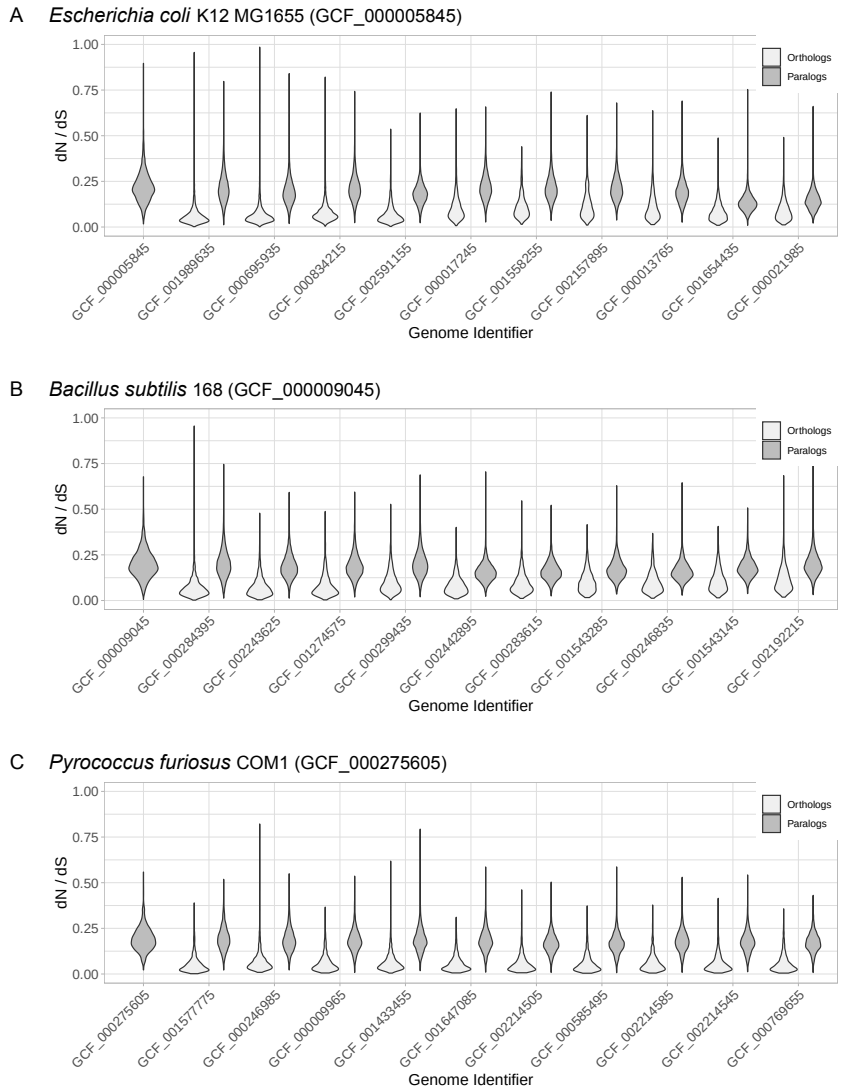

Figure 1 Non-synonymous to synonymous substitutions (dN/dS). The figure shows the dN/dS values for genes compared between query organisms, against genomes from organisms in the same taxonomic class. Namely: *E. coli* against Gammaproteobacteria, *B. subtilis* against Bacilli, and *P. furiosus* against Thermococci. The genome identifiers are ordered from most similar to least similar to the query genome. The dN/dS values tend to be higher for paralogs, suggesting that orthologs tend to keep their functions better.

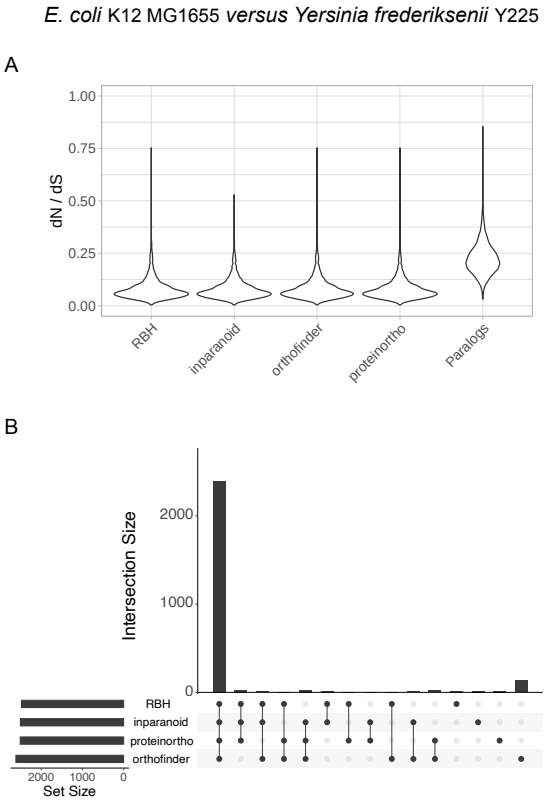

Figure 2 Working definitions of orthology. (A) The figure shows the dN/dS values of four working definitions of orthology compared to that of paralogs. (B) The ortholog sets are very similar, though not identical.

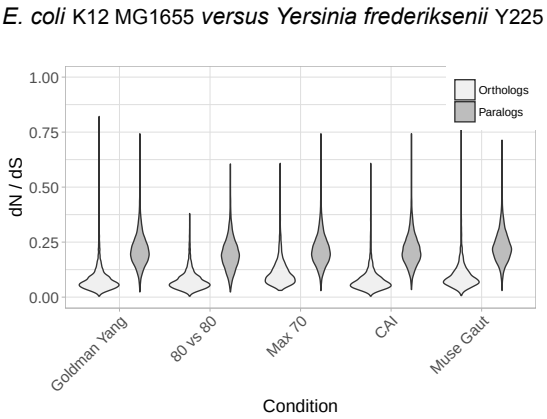

Figure 3 Quality controls. The figure shows examples of dN/dS values obtained testing for potential biases. The Goldman and Yang model for codon frequencies [33] was used for the results shown before, and it is included here for reference. The 80 vs 80 test used data for orthologs and paralogs filtered to contain only alignments covering at least 80% of both proteins. The maximum identity test filtered out sequences more than 70% identical. The CAI test filtered out sequences having Codon Adaptation Indexes (CAI) below the 15 percentile and above the 85 percentile of the genome's CAI distribution. We also tested the effect of the Muse and Gaut model for estimating background codon frequencies [32].

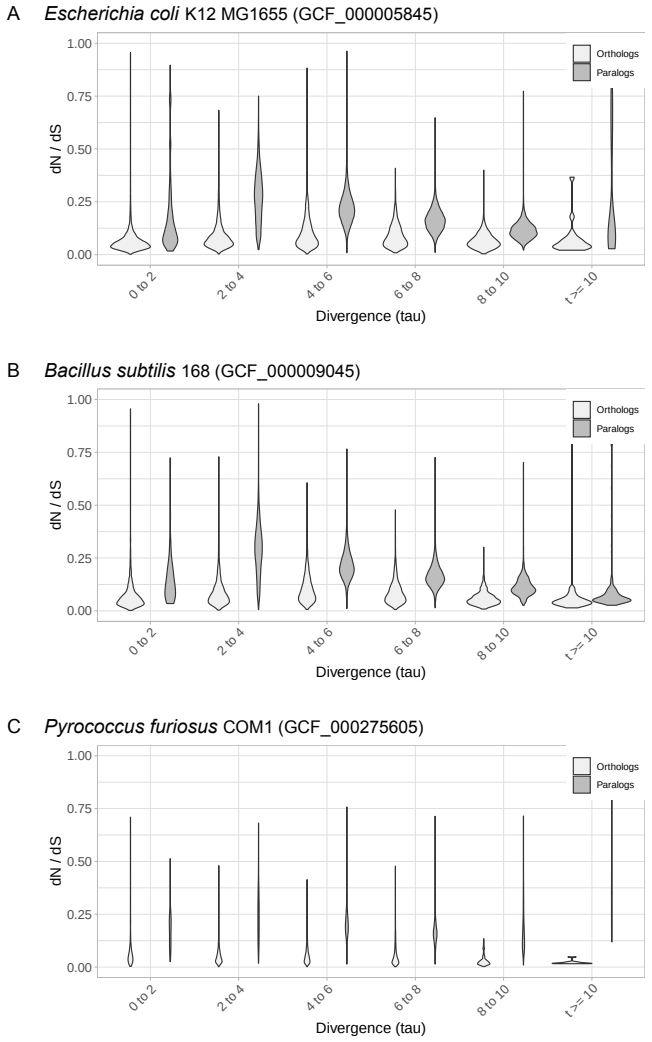

Figure 4 Non-synonymous to synonymous substitutions (dN/dS) and divergence. The figure shows the dN/dS values for genes compared between query organisms, against genomes from organisms in the same taxonomic class (Table 1). Orthologs and paralogs were separated into divergent bins. The dN/dS values tend to be higher for paralogs, suggesting that orthologs tend to keep their functions better.

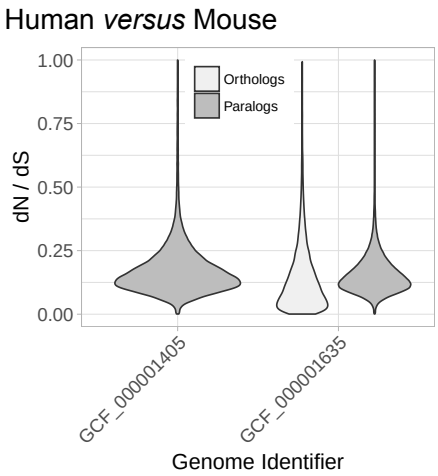

Figure 5 Mammals. The figure shows the dN/dS values obtained when comparing human and mouse genes.