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

Mario Esposito[1], Gabriel Moreno-Hagelsieb[1,*]

1 Dept of Biology, Wilfrid Laurier University, Waterloo, Ontario, Canada N2L 3C5

* gmoreno@wlu.ca

Abstract

Because orthologs diverge after speciation events, and paralogs after gene duplication, it is expected that orthologs should tend to keep their functions, while paralogs have been proposed as a source of new functions. This does not mean that paralogs should diverge much more than orthologs, but it certainly means that, if there is a difference, then orthologs should be more functionally stable. Since protein functional divergence follows from non-synonymous substitutions, here we present an analysis based on the ratio of non-synonymous to synonymous substitutions (dN/dS). The results showed orthologs to have noticeable and statistically significant lower values of dN/dS than paralogs, not only confirming that orthologs keep their functions better, but also suggesting that paralogs are a readily source of functional novelty.

Author summary

Homologs are characters diverging from a common ancestor, with orthologs being homologs diverging after a speciation event, and paralogs diverging after a duplication event. Given those definitions, orthologs are expected to preserve their ancestral function, while paralogs have been proposed as potential sources of functional novelty. Since changes in protein function require changes in amino-acids, we analyzed the rates of non-synonymous, mutations that change the encoded amino-acid, to synonymous, mutations that keep the encoded amino-acid, subtitutions (dN/dS). Orthologs showed the lowest dN/dS ratios, thus suggesting that they keep their functions better than paralogs. Because the difference is very evident, our results also suggest that paralogs are a source of functional novelty.

Introduction

In this report, we present evidence suggesting that orthologs keep their functions better than paralogs, by the analysis of non-synonymous and synonymous substitutions (dN/dS).

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

duplications, paralogy has been proposed as a mechanism for the evolution of new functions. Since one of the copies can perform the original function, the other copy would have some freedom to functionally diverge [2]. It is thus expected that the products of orthologous genes would tend to perform the same functions, while paralogs may functionally diverge. This is not to say 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, since the beginning of comparative genomics, to use some working definition for orthology to infer the genes whose products most likely perform the same functions across different species (see for example [3–5].

Despite such an obvious expectation, a report was published making the surprising claim that orthologs diverged in function more than paralogs [6]. 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 would. At worst, we could expect equal functional divergences. If the report were correct, it would mean that we should expect such things as mice myoglobin to perform 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 [6] 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 [7, 8]. These publications also proposed solutions to such and other problems [7, 8]. Another work showed that gene expression data supported the idea that orthologs keep their functions better than paralogs [9].

Most of the publications we have found have 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 [10, 11]. 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 than the 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 [12]. As of November 2017, our genome collection contained around 8500 complete genomes. 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). We compared the proteomes of each of these genomes against those of other members of their taxonomic order using BLASTP [13], with a maximum e-value of 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 [14, 15].

Except where noted, paralogs were all BLASTP matches left after finding RBHs.

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

Non-synonymous to synonymous substitutions

To perform dN/dS estimates, we used the CODEML program from the PAML software suite [16]. 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 [17, 18]. The results were separated between ortholog and paralog pairs, and the density distributions were plotted using R [19]. Statistical analyses were also performed with R.

Codon Adaptation Index

To calculate the Codon Adaptation Index (CAI) [20], 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.* [21] to the proteins in the genomes under analysis using RPSBLAST (part of NCBI's BLAST+ suite) [13]. 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. The codon usage tables of the the ribosomal protein-coding genes chosen above, we used the program *cusp* from the EMBOSS software suite [22]. 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.

Results and Discussion

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, since they are considered the most robust and accurate [17,18]. 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 rations 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}.

While the tests above suggest that reciprocal best hits 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 decided to check the results for alignments covering 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. 2, S1 Fig).

Another parameter that can bias the dN/dS results is when sequences are very similar. In this case, the programs tend to produce very high dN/dS rations. While we should expect this issue to have a larger effect on orthologs, we still filtered both homology groups to contain proteins less than 70% identical. This filter had very little effect (Fig. 2, S2 Fig).

To try and avoid the effect of horizontal gene transfers and/or of sequences with unusual compositions, we filtered out sequences with unusual codon usages as measured using the Codon Adaptation Index (CAI) [20]. 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. 2, S3 Fig).

Since different models for codon frequency can also alter the dN/dS results [25]. Thus, we performed the same tests using the Muse and Gaut model for estimating background codon frequencies [24], as advised in [25]. Again, the results show orthologs to have lower dN/dS ratios than paralogs (Fig. 2, S4 Fig).

To test whether the differences might be related to identity, we separated all of the orthologs and paralogs into bins of percent identity ranges. As expected, the dN/dS values tend to increase with lower percent identity values (Fig. 3). The differences in dN/dS between orthologs and paralogs seem less obvious than in the prior results. However, the values of dN/dS still tend to be lower for orthologs than for paralogs.

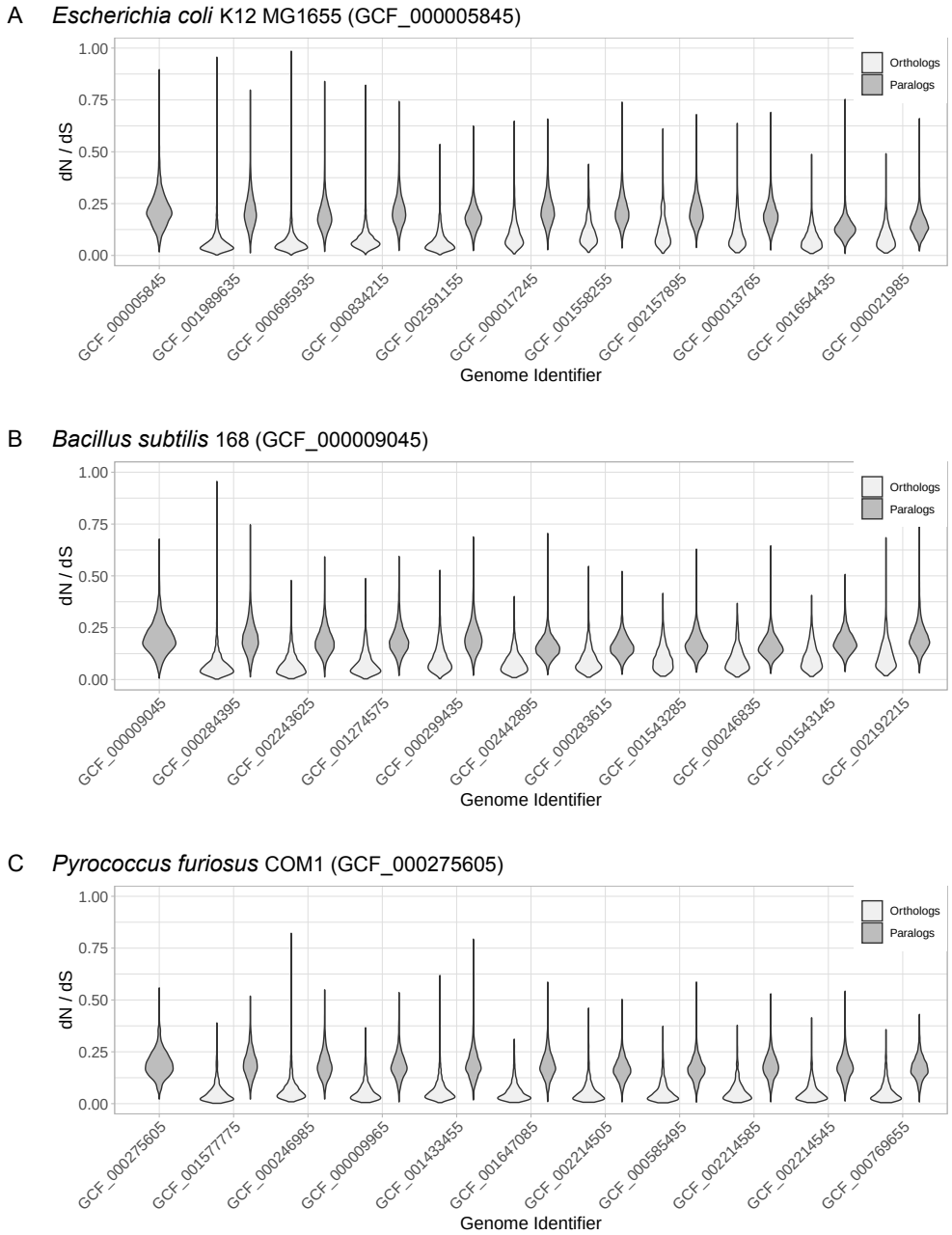


Fig 1. Non-synonymous to synonymous substitutions (dN/dS). The figure shows the dN/dS values for genes compared between a query organism, against some of the available genomes from organisms in the same taxonomic class. Namely: *E. coli* against Gammaproteobacteria, *B. subtilis* against Bacilli, and *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 than paralogs.

Finally, we also performed dN/dS comparisons for human and mouse genes. The results were very similar to those obtained above (Fig. 4).

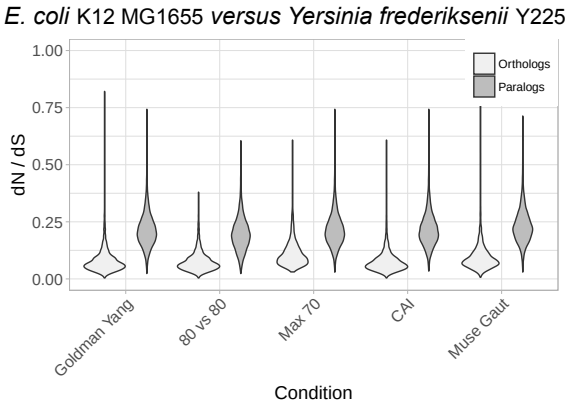


Fig 2. Quality controls. The figure shows examples of dN/dS values obtained testing for potential problems. The Goldman and Yang model for codon frequencies [23] was used for the results shown before, and it is included here for reference. The 80 vs 80 test used data for orthologs and paralogs filtered 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 [24].

A potential concern with these analyses might arise from our use of reciprocal best hits (RBHs) as a working definition of orthology. RBHs is arguably the main working definition of orthology [26]. It is thus important to start these analyses with RBHs, at the very least to show that it works for the purpose of inferring genes most likely to keep their functions when comparing any pair of genomes. Analyses of the quality of RBHs for inferring orthology, based on synteny, show that RBH error rates tend to be lower than 95% [14, 15, 26], while 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 analyses [27]. This means that our orthologs are mostly contaminated by paralogs. We can therefore infer that orthologs dominate the RBHs dN/dS distributions, since the mistakes in ortholog identification would tend to make the RBHs dN/dS values more similar to those found in paralogs.

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 that orthologs would tend to keep their functions arises naturally from the definition , while the expectation could vary for paralogs. We did not expect the differences to be so obvious. Thus, our

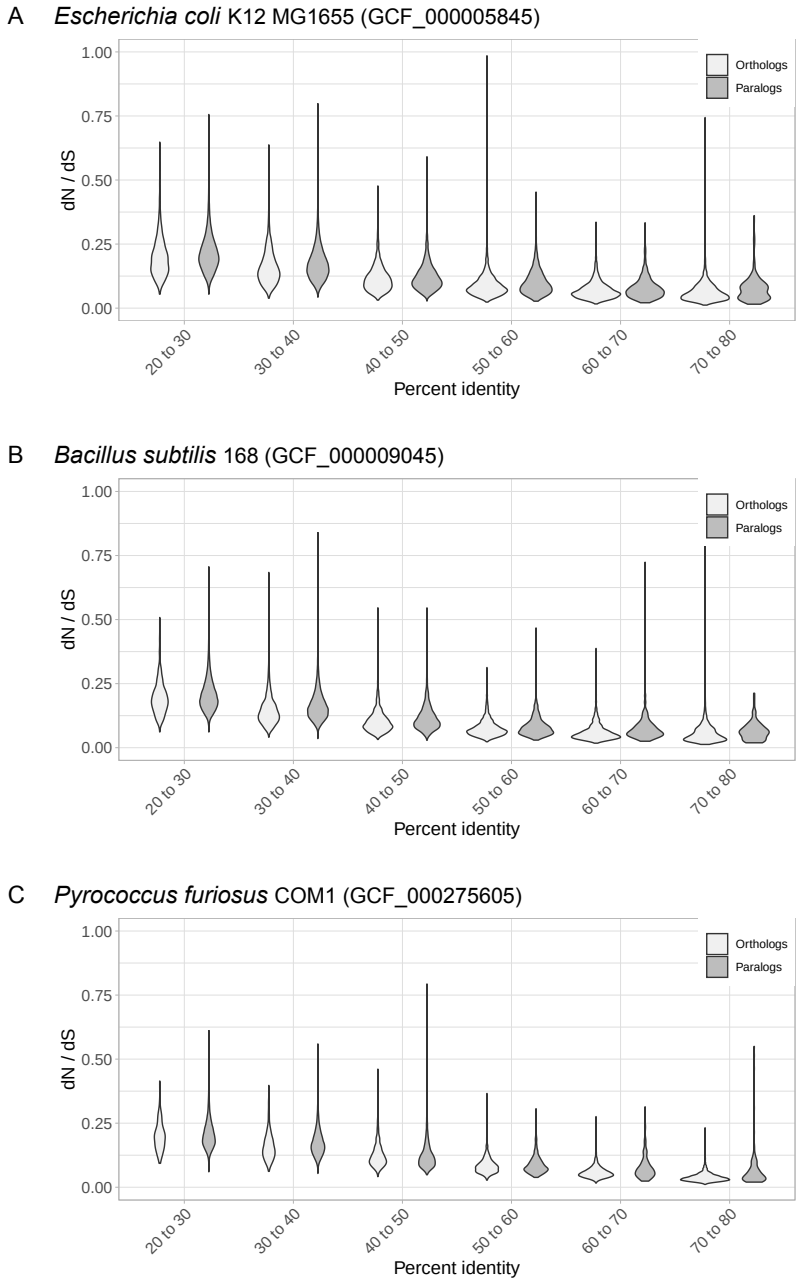


Fig 3. Non-synonymous to synonymous substitutions (dN/dS) and percent identity. The figure shows the dN/dS values for genes compared between a query organism, against some of the available genomes from organisms in the same taxonomic class (Table 1). Orthologs and paralogs were separated into pecent ideintity bins. While not as obvious as in prior results, the dN/dS values tend to be higher for paralogs, suggesting that orthologs tend to keep their functions better than paralogs.

results also show that paralogs tend to acquire novel functions.

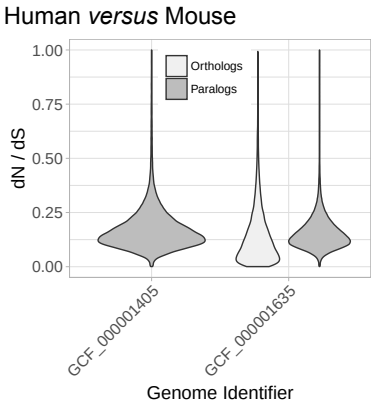


Fig 4. Mammals. The figure shows the dN/dS values obtained when comparing human and mouse genes.

Supporting information

S1 Fig. Full results on dN/dS ratios obtained with alignments covering at least 80% of both proteins. The results still show orthologs to have lower dN/dS values than paralogs.

S2 Fig. Full results on dN/dS ratios obtained with proteins no more that 70% identical. The results still show orthologs to have lower dN/dS values than paralogs.

S3 Fig. Full results on dN/dS rations obtained with proteins within usual codon usage. The Codon Adaptation Indexes [20] of the genes in the datasets had to be within the 15 and 85 percentile of the overall genomic CAI values. The results still show orthologs to have lower dN/dS values than paralogs.

S4 Fig. Full results on dN/dS rations obtained Muse and Gaut's estimate of background codon frequencies [24]. The results still show orthologs to have lower dN/dS values than paralogs.

Acknowledgments

We are grateful to Dr. Joe Bielawski for helpful advice. We thank The Shared Hierarchical Academic Research Computing Network (SHARCNET) for computing facilities. Work supported with a Discovery Grant to GM-H from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Fitch WM. Homology a personal view on some of the problems. Trends Genet. 2000;16(5):227–231.
2. Ohno S. Evolution by gene duplication. Berlin: Springer-Verlag; 1970.

3. Mushegian AR, Koonin EV. Gene order is not conserved in bacterial evolution. Trends Genet. 1996;12(8):289–290.

4. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. Predicting function: from genes to genomes and back. J Mol Biol. 1998;283(4):707–725.

5. Gabaldón T, Koonin EV. Functional and evolutionary implications of gene orthology. Nat Rev Genet. 2013;14(5):360–366.

6. Nehrt NL, Clark WT, Radivojac P, Hahn MW. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol. 2011;7(6):e1002073.

7. Thomas PD, Wood V, Mungall CJ, Lewis SE, Blake JA, Gene Ontology Consortium. On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. PLoS Comput Biol. 2012;8(2):e1002386.

8. Altenhoff AM, Studer RA, 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. 2012;8(5):e1002514.

9. Kryuchkova-Mostacci N, Robinson-Rechavi M. Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. PLoS Comput Biol. 2016;12(12):e1005274.

10. Ohta T. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J Mol Evol. 1995;40(1):56–63.

11. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 2000;17(1):32–43.

12. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, et al. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res. 2018;46(D1):D851–D860.

13. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

14. Moreno-Hagelsieb G, Latimer K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics. 2008;24(3):319–324.

15. Ward N, Moreno-Hagelsieb G. Quickly Finding Orthologs as Reciprocal Best Hits with BLAT, LAST, and UBLAST: How Much Do We Miss? PLoS ONE. 2014;9(7):e101850.

16. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–1591.

17. Anisimova M, Bielawski JP, Yang Z. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol. 2002;19(6):950–958.

18. Angelis K, dos Reis M, Yang Z. Bayesian estimation of nonsynonymous/synonymous rate ratios for pairwise sequence comparisons. Mol Biol Evol. 2014;31(7):1902–1913.

19. R Core Team. R: A Language and Environment for Statistical Computing; 2018. Available from: `https://www.R-project.org/`.

20. Sharp PM, Li WH. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15(3):1281–1295.

21. Yutin N, Puigbò P, Koonin EV, Wolf YI. Phylogenomics of prokaryotic ribosomal proteins. PLoS ONE. 2012;7(5):e36972.

22. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16(6):276–277.

23. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994;11(5):725–736.

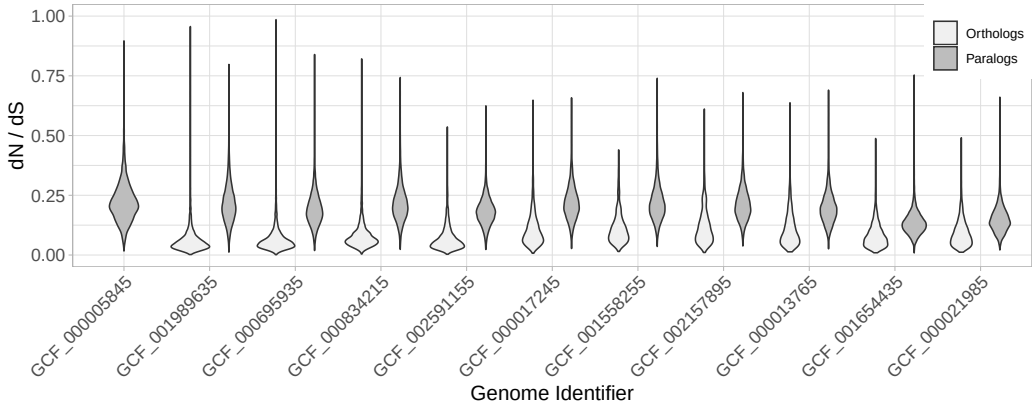
24. Muse SV, Gaut BS. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 1994;11(5):715–724.

25. Bielawski JP. Detecting the signatures of adaptive evolution in protein-coding genes. Curr Protoc Mol Biol. 2013;Chapter 19:Unit 19.1.

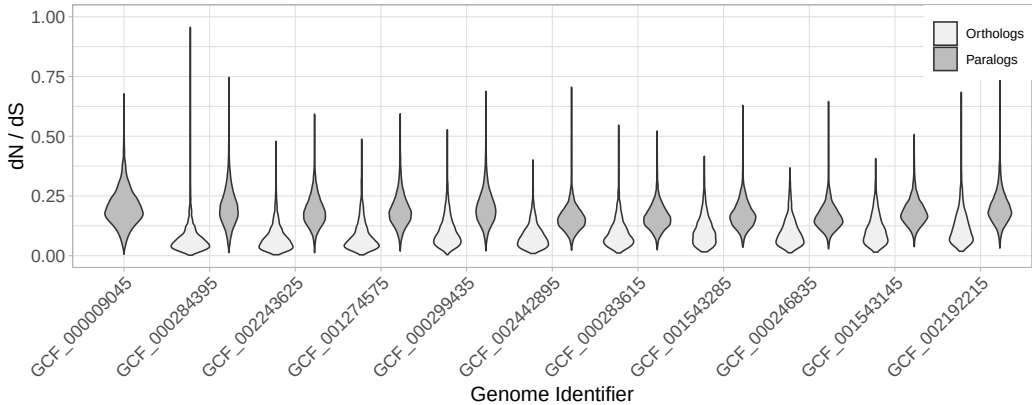
26. Wolf YI, Koonin EV. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol Evol. 2012;4(12):1286–1294.

27. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol. 2009;5(1):e1000262.

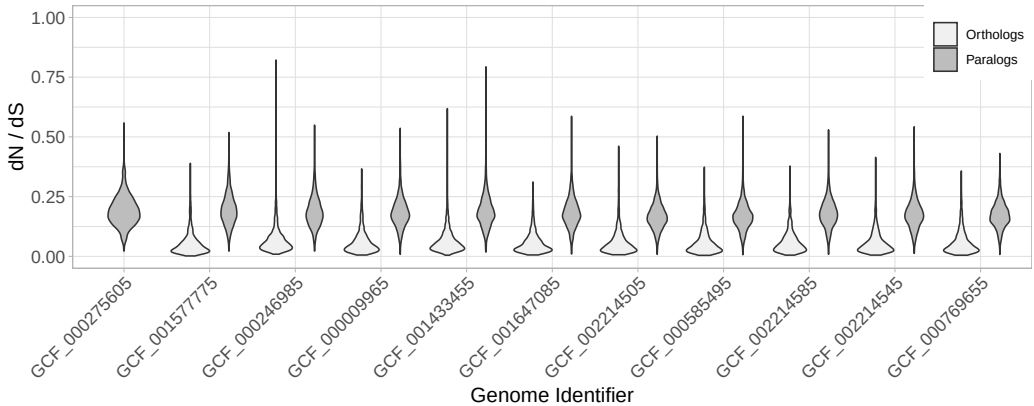
A *Escherichia coli* K12 MG1655 (GCF_000005845)

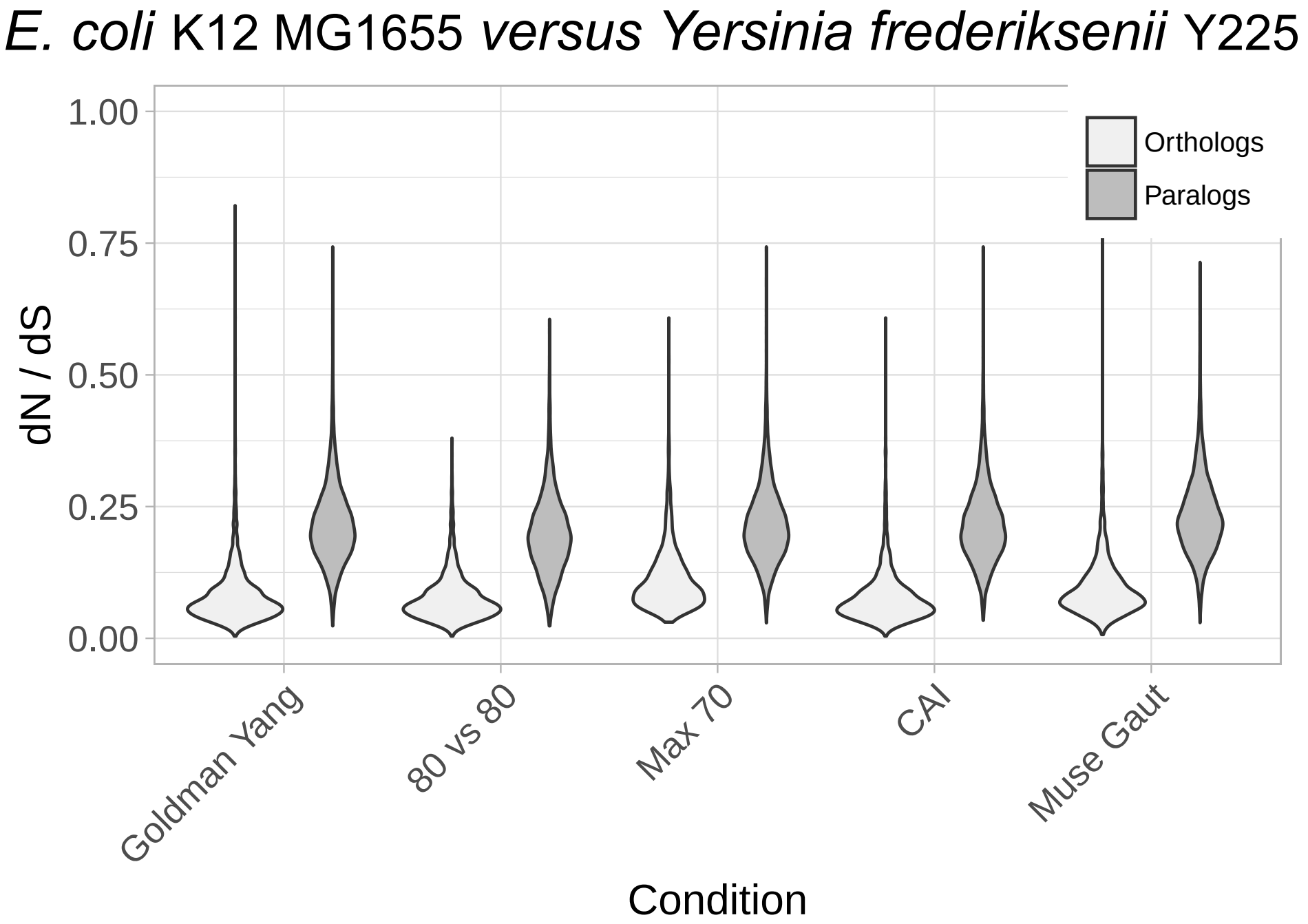


B *Bacillus subtilis* 168 (GCF_000009045)

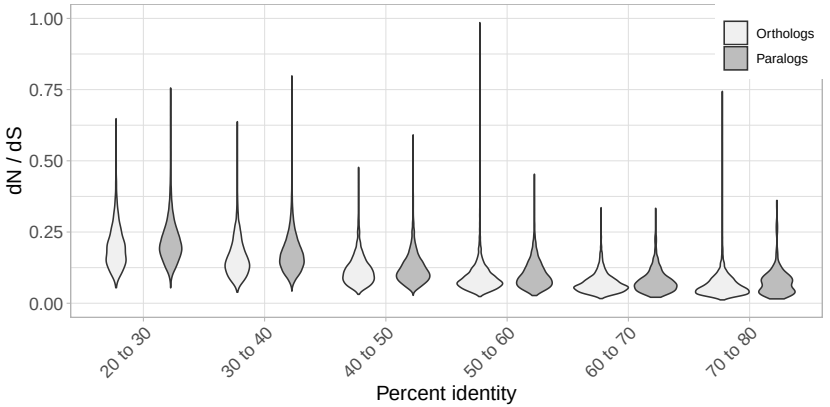


C *Pyrococcus furiosus* COM1 (GCF_000275605)

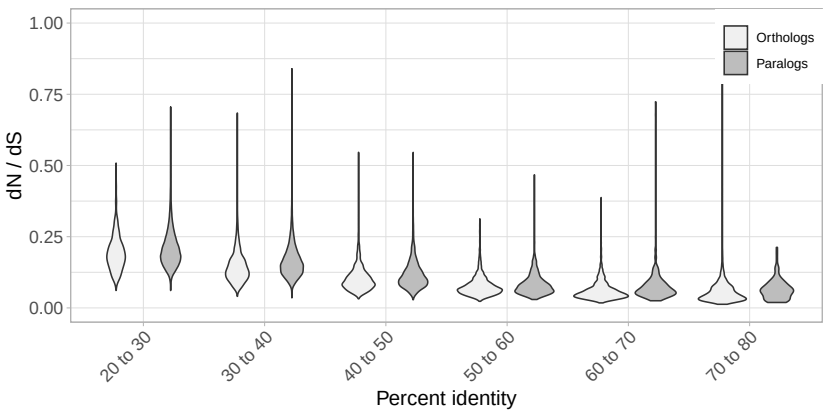



E. coli K12 MG1655 *versus Yersinia frederiksenii* Y225
1.00
0.75
0.50
0.25
0.00
dN / dS
Orthologs
Paralogs
Goldman Yang
80 vs 80
Max 70
CAI
Muse Gaut
Condition

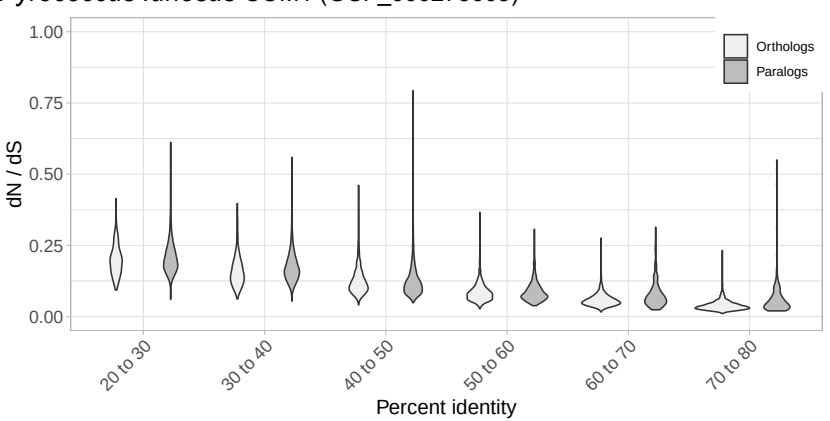
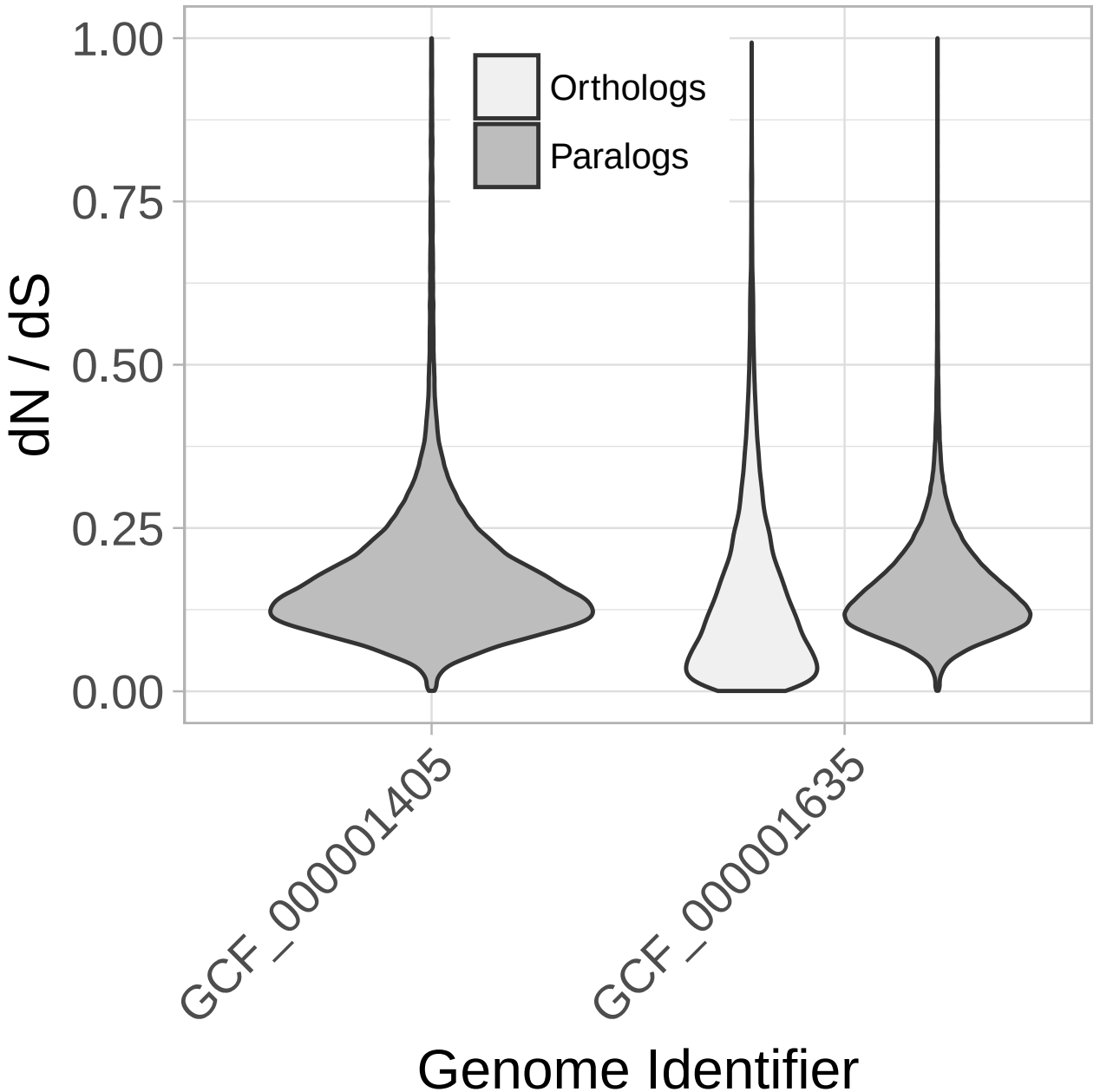
A *Escherichia coli* K12 MG1655 (GCF_000005845)



B *Bacillus subtilis* 168 (GCF_000009045)



C *Pyrococcus furiosus* COM1 (GCF_000275605)



Human *versus* Mouse