A Tight Link between Orthologs and Bidirectional Best Hits in Bacterial and Archaeal Genomes

Yuri I. Wolf and Eugene V. Koonin*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

*Corresponding author: E-mail: koonin@ncbi.nlm.nih.gov.

Accepted: November 9, 2012

Abstract

Orthologous relationships between genes are routinely inferred from bidirectional best hits (BBH) in pairwise genome comparisons. However, to our knowledge, it has never been quantitatively demonstrated that orthologs form BBH. To test this "BBH-orthology conjecture," we take advantage of the operon organization of bacterial and archaeal genomes and assume that, when two genes in compared genomes are flanked by two BBH show statistically significant sequence similarity to one another, these genes are bona fide orthologs. Under this assumption, we tested whether middle genes in "syntenic orthologous gene triplets" form BBH. We found that this was the case in more than 95% of the syntenic gene triplets in all genome comparisons. A detailed examination of the exceptions to this pattern, including maximum likelihood phylogenetic tree analysis, showed that some of these deviations involved artifacts of genome annotation, whereas very small fractions represented random assignment of the best hit to one of closely related in-paralogs, paralogous displacement in situ, or even less frequent genuine violations of the BBH–orthology conjecture caused by acceleration of evolution in one of the orthologs. We conclude that, at least in prokaryotes, genes for which independent evidence of orthology is available typically form BBH and, conversely, BBH can serve as a strong indication of gene orthology.

Key words: orthology, bidirectional best hit, genome comparison, synteny.

Gene orthology is the central concept of comparative and evolutionary genomics. Orthologs are defined as genes that derive from a single ancestral gene in the last common ancestor of the compared genomes (Fitch 1970, 2000; Sonnhammer and Koonin 2002; Koonin 2005). Robust identification of orthologs is essential for accurate reconstruction of genome evolution (Koonin 2005; Lemoine et al. 2007). Probably the best recognized and the most important implication of orthology is the "ortholog conjecture": orthologs perform "the same" function in different organisms (to the extent biological functions in different organisms can be considered equivalent) (Koonin 2005; Nehrt et al. 2011). This conjecture is the cornerstone of all functional annotation of sequenced genomes, that is, the justification of the transfer of functional assignments between genomes. Recently, the orthology conjecture has been severely challenged by observations that, at the same level of sequence similarity, paralogs within the same genome appeared to be more similar functionally (judged by the similarity in the Gene Ontology classification [Lomax 2005; Skunca et al. 2012]) and in terms of the expression profile than orthologs from different genomes (Nehrt et al. 2011). Testing the orthology conjecture involves many potential artifacts, especially when it comes to direct comparison of functional assignments. Indeed, two independent research groups have already published refutations of the conclusions of Nehrt et al. (2011) that paralogs are more functionally similar than orthologs, thus apparently upholding the orthology conjecture (Altenhoff et al. 2012; Thomas et al. 2012).

As a simple step in a thorough investigation of the orthology problem, we sought to test what could be denoted "bidirectional bets hits (BBH)-orthology equivalence conjecture" (BBHO conjecture for short): the sequences of orthologous genes (proteins) are more similar to each other than to any other sequences in the respective genomes. Indeed, in the great majority of comparative genomics studies, orthology is inferred from the BBH rather than by direct analysis of phylogenetic trees or on the basis of other, independent evidence (Kristensen et al. 2011), under the assumption of the BBH–orthology equivalence (fig. 1). Certain aspects of the BBHO conjecture have been explored previously. In particular, it has been shown that in the trade-off between sensitivity and selectivity of functional conservation prediction, BBH represent the conservative end of the spectrum, that is, low coverage but high functional correlation (Hulsen et al. 2006). Another

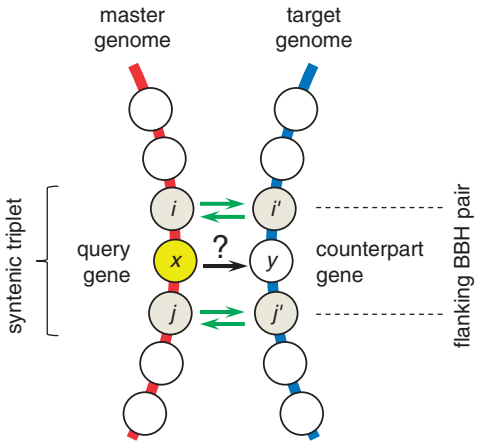

Fig. 1.—Schematic of the genome comparison for testing the BBH–orthology conjecture.

study has concluded that, at least for an extremely simple procedure, the BBH analysis performed remarkably well in terms of the agreement with phylogenetic trees (Altenhoff and Dessimoz 2009). However, there also have been reports that genes coming out as being most similar in sequence database searches do not necessarily form clades in phylogenetic trees (Koski and Golding 2001; Lemoine et al. 2007).

Evolution of orthologous gene families is almost always accompanied by gene duplications and losses. Genes that arose from duplications form paralogous families (in-paralogs) within the respective lineages while remaining, in a broad sense, (co)orthologous to each other and other, nonduplicated genes in other lineages. Gene loss complicates matters further, making the observed relationships to span the complete range of the many-to-many-to-none (Koonin 2005; Kristensen et al. 2011). However, if the similarity between sequences is comonotonic with evolutionary distance, each set of in-paralogs within a particular genome would form at least one BBH with the orthologous sets of in-paralogs (or solo orthologs) in other genomes. Thus, in a broad sense, the BBHO conjecture still holds for gene families with duplications and losses, applying to co-orthologous clusters instead of individual genes. This is how we treat BBHO in the subsequent text.

We tested the BBHO conjecture by taking advantage of the operon organization of bacterial and archaeal genomes. Operons are distinct units of evolution that typically consist of three or four genes. The results of testing the consistency of the BBH relationships in gene strings imply that BBH is an excellent proxy for orthology.

We sought to identify the simplest conceivable approach to test the BBHO conjecture. To this end, pairwise comparisons of bacterial and archaeal genomes were performed and all pairs of "syntenic triplets" of genes (*ixj*)–(*i'yj'*) were extracted such that *i*–*i'* and *j*–*j'* are BBH. The protein products of the genes *x* and *y* could either show significant sequence similarity to one another or not. Assuming that the genes *i*–*i'* and *j*–*j'* are pairs of orthologs (two nonorthologous, closely spaced BBH being extremely unlikely), we conjecture that, if the protein sequences of the genes *x* and *y* are significantly similar, these genes are orthologous. The alternative is the highly unlikely (but not impossible) in situ displacement of the ortholog by a paralog. Thus, to the extent the middle *x*–*y* pairs in syntenic gene triplets are BBH, the BBHO conjecture holds.

To test whether the middle genes in the syntenic gene triplets were BBH, we chose two well-characterized "master" genomes, the bacterium *Escherichia coli* and the archaeon *Haloarcula marismortui*, and compared all protein sequences encoded in these genomes to the proteins from each of the 573 representative bacterial and archaeal genomes using BLASTP (Altschul et al. 1997). The representative microbial genomes (the largest genome in a genus with addition of *E. coli* K12 and *Bacillus subtilis*) were downloaded from NCBI Microbial Genomes FTP site. For each of the master genomes, Basic Local Alignment Search Tool (BLAST) search was performed against all representative genomes and in the opposite direction with a permissive threshold (*e* value 0.01, no low-complexity filtering, or composition-based statistics adjustment); BBH were recorded for each of the genome pairs. Score for each BBH was normalized by the self-hit score in the master genome and converted into distance using the *distance* = –ln(*score*) relationship. The distance between the compared genomes was estimated as the median distance between BBH pairs. The number of BBH pairs was normalized by the geometric mean of the number of protein-coding genes in the master and the target genome. The BLAST hits between the master and the target genomes with scores ≥99% of the top-scoring hit was classified as "best hit," whereas hits with scores ≥1% of the top-scoring hit was classified as "significant hits." The 99% bracket was used because the ranking of hits by BLAST has a non-negligible margin of error. Benchmarking with the *H. marismortui* protein sequences showed that using the strict best hit instead led to the loss of 22 of the 15,639 BBH in the middle of syntenic gene triplets. The best hits were tested for bidirectionality by using sequences from the target genomes as BLAST queries against the master genome.

In agreement with the BBHO conjecture, we found that the overwhelming majority of the middle genes from syntenic triplets indeed were best hits, most of these bidirectional (fig. 2 and table 1). For pairs of genomes from relatively close organisms, such as *Proteobacteria* in the case of *E. coli* or *Halobacteria* in the case of *H. marismortui*, the BBH accounted for more than 95% of the *x* genes; among the remaining ones, the majority did not show significant similarity to the gene from the master genome, and only a small fraction (<2%) represented genes with significant similarity to the query that, however, were not the best hits (table 1; see supplementary file S1, Supplementary Material online, for details). Although the number of syntenic gene triplets dramatically

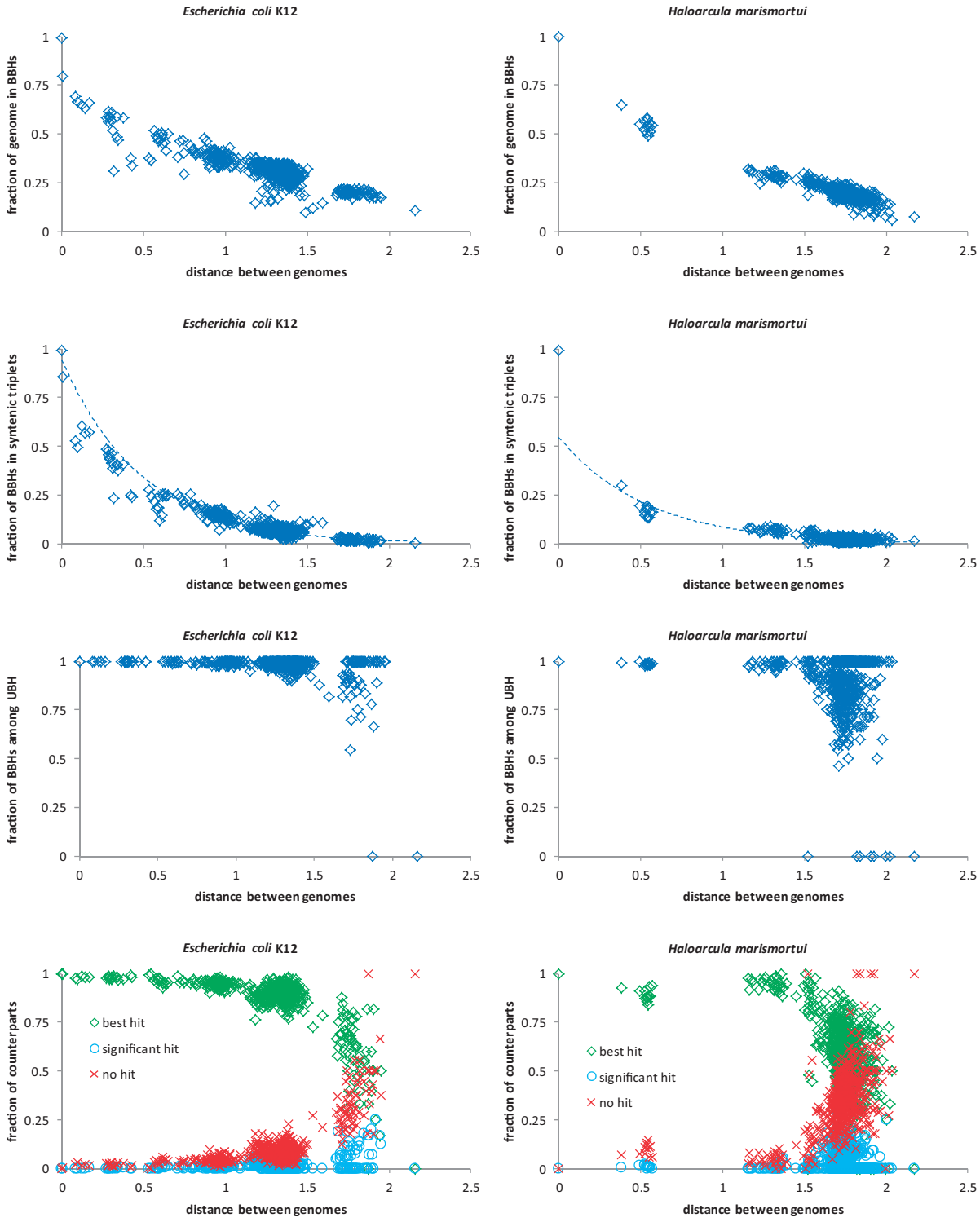

Fig. 2.—Dependency of the relationship between the middle genes in syntenic gene triplets on the distance between the compared genomes.

drops in comparisons of more distant organisms, the fraction of non-BBH significant hits remains low; it only increases to approximately 5% for the most distant comparisons, namely archaea against bacteria, in which the number of syntenic gene triplets is in the low tens (fig. 2 and table 1). The nonhomologous *y* genes are irrelevant in the context of the BBHO conjecture because, obviously, not all syntenic gene triplets are operons. By contrast, the genes with similar

Table 1

The Status of Middle Genes (*x*) in Syntenic Gene Triplets (*ixj*) Depending on the Distance between Compared Genomes

Taxa	BBH d[a]	*f* BH[b]	*f* SH[c]	*f* NH[d]	BBH/G[e]	T/BBH[f]	fBBH[g]
Escherichia coli K12							
Enterobacteria	0.3042	0.9810	0.0010	0.0180	0.5555	0.4487	0.9978
Gamma-proteobacteria	0.8509	0.9451	0.0060	0.0490	0.4058	0.1717	0.9885
Proteobacteria-	1.1818	0.9082	0.0150	0.0768	0.3346	0.0936	0.9886
Bacteria	1.3591	0.8956	0.0169	0.0875	0.2890	0.0649	0.9816
Archaea	1.7915	0.5918	0.0542	0.3540	0.1973	0.0206	0.9075
Haloarcula marismortui							
Halobacteria	0.4881	0.8987	0.0084	0.0929	0.5804	0.2327	0.9849
Methanomicrobia	1.2498	0.9434	0.0000	0.0566	0.2910	0.0754	0.9815
Euryarchaeota	1.3500	0.9441	0.0103	0.0456	0.2815	0.0646	0.9863
Archaea	1.5352	0.8133	0.0135	0.1732	0.2518	0.0452	0.9202
Bacteria	1.7606	0.5900	0.0449	0.3631	0.1915	0.0217	0.8815

[a]Mean distance between BBH for the master genome and other genomes in the respective group.
[b]Fraction of best hits among the counterparts of the middle genes in syntenic triplets.
[c]Fraction of other significant hits among the counterparts.
[d]Fraction of nonhomologous genes among the counterparts.
[e]Fraction of genes in BBH.
[f]Fraction of BBH in syntenic triplets.
[g]Fraction of BBH among best hits.

sequences that do not rank first in the list of homologs of the query are suspected violators of the BBHO conjecture.

For eight selected pairs of genomes, we performed an exhaustive, case by case examination of these putative violations of the BBHO conjecture (table 2). Excluding the cases when one of the *x* genes was represented by a fragment (most likely due to sequencing errors and/or misannotation) or the middle genes belonged to a complex family of paralogous genes (e.g., two-component signal transduction systems), phylogenetic analysis using the FastTree approximate maximum likelihood method (Price et al. 2010; Liu et al. 2011) with the WAG evolutionary model was performed for the suspect genes. Multiple alignments for the phylogenetic analysis were constructed using MUSCLE (Edgar 2004); and sites containing >50% of gaps were removed. The rationale was that, when in a phylogenetic tree the non-BBH hits from the compared pair of genomes appear as neighbor branches (the closest branches for the species in question), the BBHO conjecture is violated: orthologs are not BBH. In contrast, when neighbor branches are the query and a homologous gene outside the syntenic triplet, whereas the *x* gene is reliably separated in the tree, the BBHO conjecture holds: the ortholog is not the gene in the syntenic triplet (or else the homologous genes within and outside of the syntenic triplet are co-orthologs; see later). Strikingly, in eight genome pairs, we identified only three apparent violations of the BBHO conjecture (table 2).

Figure 3 shows examples of phylogenetic trees for non-BBH significant hits of middle genes in syntenic gene triplets (see supplementary file S2, Supplementary Material online, for multiple sequence alignments that were used for tree construction). The cases in figure 3*A* and *B* present typical violations of the BBHO conjecture where the in situ counterparts confidently cluster together even though they are not BBH. This topology of the phylogenetic trees implies that the homologs in syntenic gene triplets are the actual orthologs. The anomalous ranking of sequence similarity then could be attributed to acceleration of evolution in one or both of the compared genomes that is likely to accompany functional change. Figure 3*C* shows a more complicated case where multiple paralogs of the query gene are present both in the master genome and in the target genome, but both the most similar gene in the target genome and the syntenic homolog appear to be distant paralogs of the query; the true ortholog probably has been lost in the target lineage, a special case of the BBHO conjecture violation. Figure 3*D*–*F* illustrates situations where the BBHO conjecture appears to hold even though the in situ counterparts are not BBH. In each of these trees, the query gene confidently clusters with the most similar gene in the target genome even though the latter is not the in situ counterpart. Conceivably, in some of these cases, the ortholog has been replaced with a paralog in situ, that is, without disruption of the operon (fig. 3*D*). Previously, we have observed several occasions of in situ displacement of a gene in an operon with a nonhomologous gene indicating that displacement events, however rare, do occur during the evolution of bacteria and archaea (Omelchenko et al. 2003). In other, probably more common cases, there is a less exotic explanation for the observed anomalous ranking of sequences similarity. As shown in figure 3*E* and *F*, the target genomes encompass two closely related in-paralogs one of which is located within the syntenic gene triplet and the other one elsewhere in the genome. In these situations, the best hit may not be the in situ counterpart,

Table 2
Test of the BBH–Orthology Conjecture for Selected Pairs of Genomes

Gene	Tree Analysis and Status of the BBHO Conjecture
Escherichia coli K12–*Cronobacter turicensis*	
yfeD DUF1323	Large family of paralogs (MerR-like HTH-containing transcription regulators)
dkgB COG0656	Fragment or different architecture
fimF COG3539	Fragment or different architecture
ompF COG3203	Violated
Escherichia coli K12–*Ralstonia eutropha*	
kdpD COG2205	Large family of paralogs (two-component regulatory system)
dedD COG3147	Large family of paralogs (periplasmic binding proteins)
yehY COG1174	Fragment or different architecture
rfbA COG1209	Fragment or different architecture
tolA	Fragment or different architecture
nuoE COG1905	Violated
hisC COG0079	Supported
ccmG COG0526	Supported
ccmF COG1138	Supported
narJ COG2180	Supported
narG COG5013	Supported
Escherichia coli K12–*Bacillus subtilis*	
ycjO COG1175	Large family of paralogs (permeases)
rpsN COG0199	fragment or different architecture
chbC COG1455	Supported
Haloarcula marismortui–*Haloterrigena turkmenica*	
rrnAC2237 arCOG02980	Large family of paralogs (uncharacterized proteins)
rrnAC1023	Large family of paralogs (uncharacterized proteins)
rrnAC1507 COG0581	Fragment or different architecture
rrnAC2884 arCOG06342	Fragment or different architecture
rrnAC1790 COG1228	Supported
Haloarcula marismortui–*Pyrococcus furiosus*	
rrnAC3533 COG2111	Fragment or different architecture
rrnAC0069 COG0148	Violated
rrnAC2679 COG0189	Supported

essentially by chance, because the difference between the closely related sequences of the in-paralogs is beyond the resolution of BLASTP. Formally at least, the BBHO conjecture holds, with the in-paralogs classified as co-orthologs of the query gene (Koonin 2005) although one could argue that the syntenic location is ancestral and accordingly the orthologs that share this location ("toporthologs" [Dewey 2011]) reflect the "true" orthologous lineage.

We restricted our analysis to prokaryotic genomes for the following reasons: domain architecture of eukaryotic proteins is considerably more variable even within orthologous families and the propensity of eukaryotic genomes for duplications and annotation errors due to inaccurate intron–exon recognition increase the variation among published sets of genes. Because of these problems, the concept of one-on-one orthology is inapplicable to a large fraction of the eukaryotic genes, particular in complex multicellular organisms. Even more pertinent in the specific context of the present analysis, because there is much less selection pressure on the gene order in eukaryotes (Koonin 2009), the utility of synteny analysis would be limited to short evolutionary distances.

Conclusions

The results of this analysis reveal remarkable consistency of the BBH between genomes of bacteria and archaea. Whenever the flanking genes in a gene triplet are BBH, it is almost certain that, if the middle genes are homologous at all, they are also BBH. This consistency of the BBH is strong evidence in support of the BBHO conjecture. Assuming that the middle genes in syntenic triplets represent an unbiased sample of orthologous genes, we are justified to conclude that the majority of orthologs from BBH. The only caveat we are aware of is that this statement could become technically incorrect for genomes that encompass numerous in-paralogs. In the present analysis, this situation was not observed for any of the analyzed genome pairs. Although the analysis presented here does not directly address gene functions, the results are best

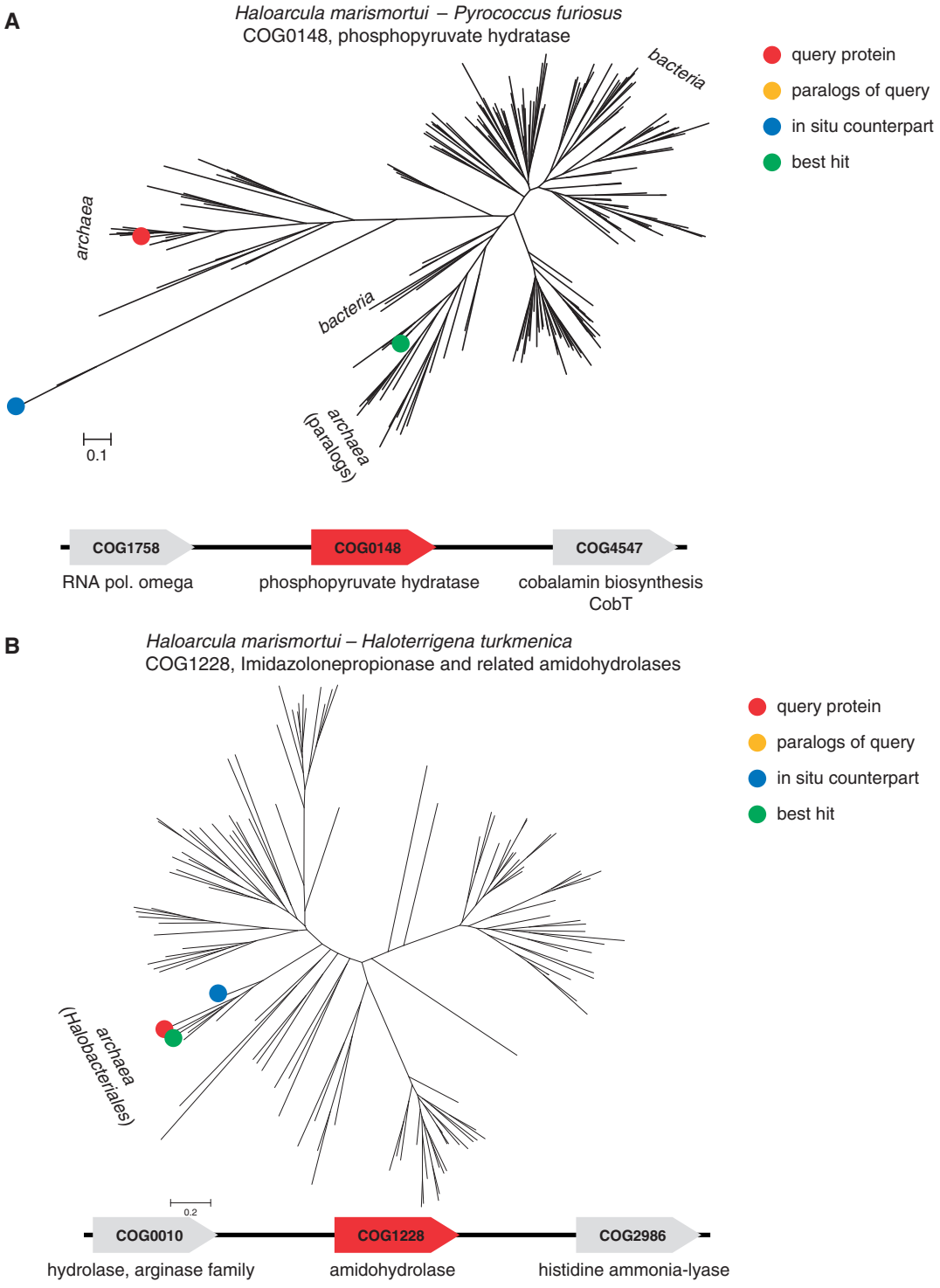

Fig. 3.—Examples of deviations from the predictions of the BBH–orthology conjecture. (*A*) *Haloarcula marismortui–Pyrococcus furiosus,* phosphopyruvate hydratase (COG0148). BBH–orthology conjecture violated due to an acceleration of evolution of one of the in situ homologs. (*B*) *Escherichia coli* K12–*Ralstonia eutropha,* NADH:ubiquinone oxidoreductase NuoE (COG1905). BBH–orthology conjecture violated, probably due to an acceleration of evolution of one of the in situ homologs.

(continued)

compatible with functional conservation between orthologous genes. Indeed, the apparent BBH–orthology equivalence is a direct consequence of the molecular clock principle as formulated by Kimura (1983): the rate of evolution of a gene remains approximately constant as long as its function does not change. We found that major deviations from this principle resulting in violations of the BBHO conjecture are extremely rare. Recently, it has been shown that orthologs, on average, share a greater structural similarity than paralogs at the same level of sequence divergence (Peterson et al.

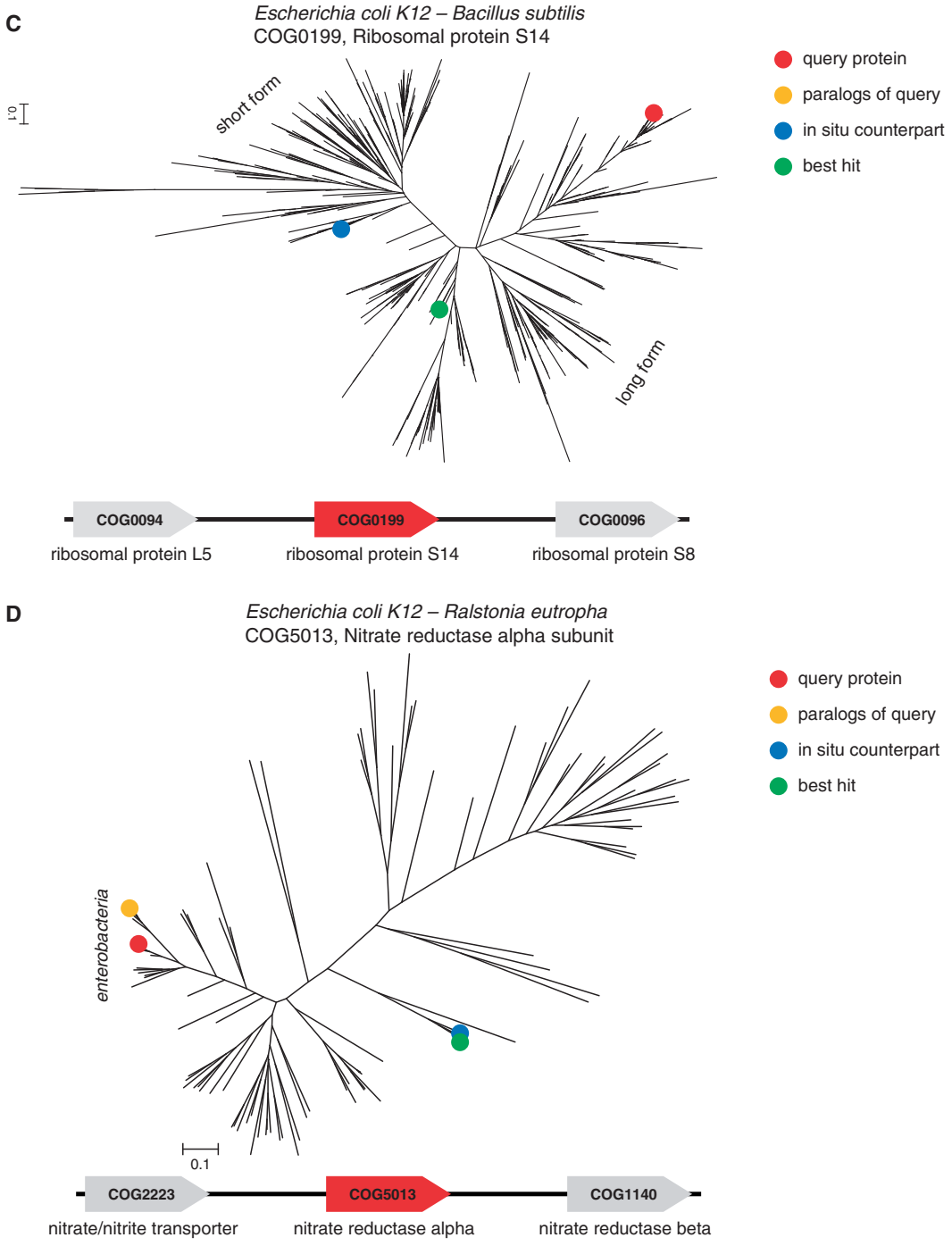

Fig. 3.— (*C*) *Escherichia coli* K12–*Cronobacter turicensis*. Outer membrane porin OmpC (COG3203). BBH–orthology conjecture violated, complex evolutionary relationships between multiple paralogs. (*D*) *Escherichia coli* K12–*Bacillus subtilis*, ribosomal protein S14 (COG0199). Compatible with the BBH–orthology conjecture. The in situ homolog is markedly more distant from the query gene in sequence and domain architecture as well as the position in the tree.

(continued)

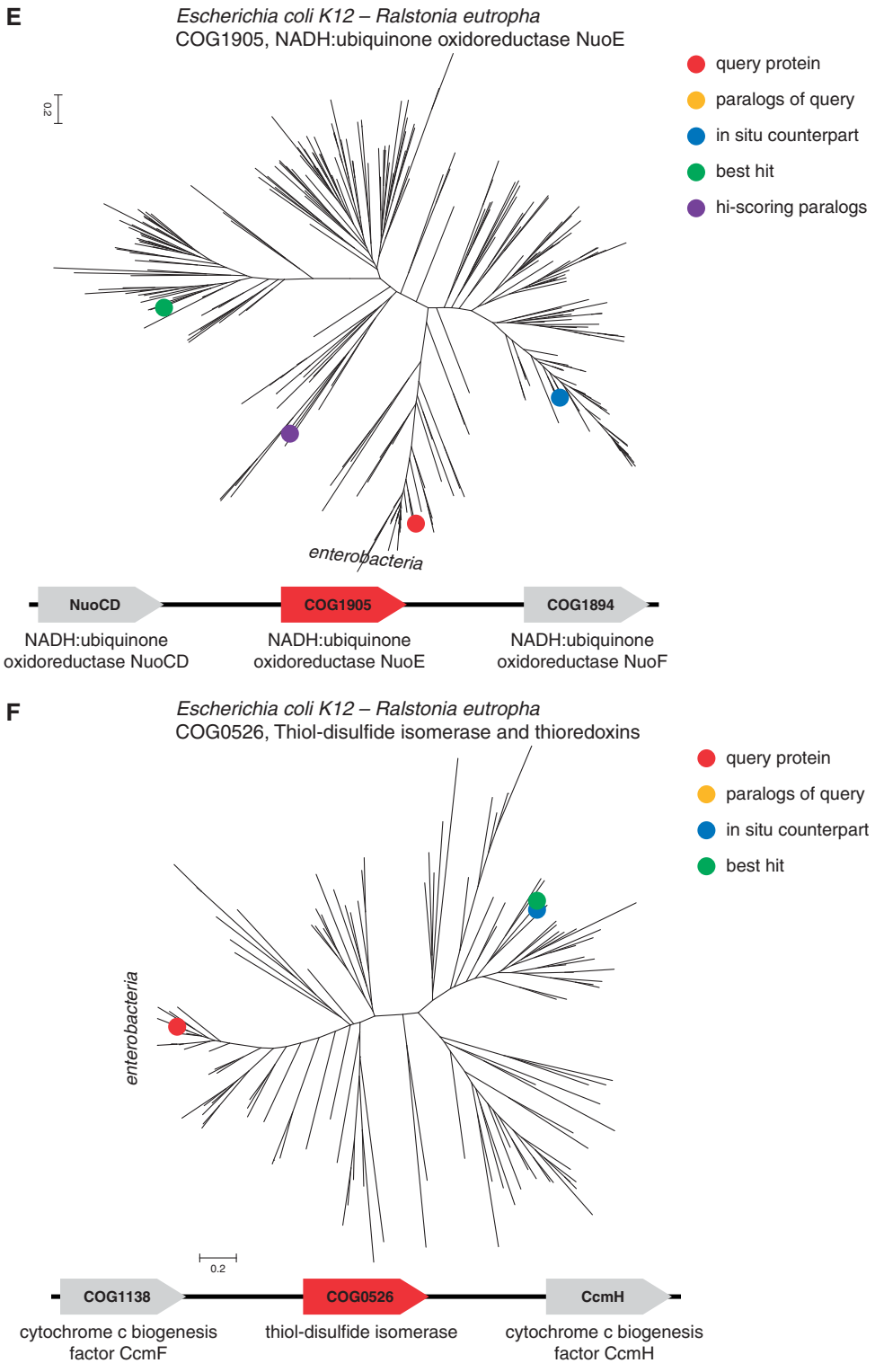

Fig. 3.— (*E*) *Escherichia coli* K12–*R. eutropha,* Thiol-disulfide isomerase and thioredoxins (COG0526). Compatible with BBH–orthology conjecture; the best hit and the in situ homolog are closely related in-paralogs in *R. eutropha*. (*F*) *Escherichia coli* K12–*R. eutropha,* Nitrate reductase alpha subunit (COG5013). Compatible with BBH–orthology conjecture; the best hit and the in situ homolog are closely related in-paralogs in *R. eutropha*.

2009), a greater similarity of domain architectures (Forslund et al. 2011) and more similar tissue expression profiles (Huerta-Cepas et al. 2011). Taken together, all these findings imply that orthology is a valid and powerful concept that captures congruent evolutionary trajectories of genes derived from the same ancestral gene. The more practical implication is that, all the caveats notwithstanding, orthology is a solid basis for information transfer in genome annotation.

Supplementary Material

Supplementary files S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This research was supported by intramural funds of the U.S. Department of Health and Human Services to the National Library of Medicine.

Literature Cited

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

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol. 8:e1002514.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Dewey CN. 2011. Positional orthology: putting genomic evolutionary relationships into context. Brief Bioinform. 12:401–412.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Zool. 19:99–106.

Fitch WM. 2000. Homology: a personal view on some of the problems. Trends Genet. 16:227–231.

Forslund K, Pekkari I, Sonnhammer EL. 2011. Domain architecture conservation in orthologs. BMC Bioinformatics 12:326.

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldon T. 2011. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. Brief Bioinform. 12:442–448.

Hulsen T, Huynen MA, de Vlieg J, Groenen PM. 2006. Benchmarking ortholog identification methods using functional genomics data. Genome Biol. 7:R31.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge (United Kingdom): Cambridge University Press.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 39:309–338.

Koonin EV. 2009. Evolution of genome architecture. Int J Biochem Cell Biol. 41:298–306.

Koski LB, Golding GB. 2001. The closest BLAST hit is often not the nearest neighbor. J Mol Evol. 52:540–542.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. 2011. Computational methods for gene orthology inference. Brief Bioinform. 12: 379–391.

Lemoine F, Lespinet O, Labedan B. 2007. Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. BMC Evol Biol. 7:237.

Liu K, Linder CR, Warnow T. 2011. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. PLoS One 6:e27731.

Lomax J. 2005. Get ready to GO! A biologist's guide to the gene ontology. Brief Bioinform. 6:298–304.

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

Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV. 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biol. 4:R55.

Peterson ME, et al. 2009. Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Sci. 18:1306–1315.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

Skunca N, Altenhoff A, Dessimoz C. 2012. Quality of computationally inferred gene ontology annotations. PLoS Comput Biol. 8: e1002533.

Sonnhammer EL, Koonin EV. 2002. Orthology, paralogy, and proposed classification for paralog subtypes. Trends Genet. 18:619–620.

Thomas PD, Wood V, Mungall CJ, Lewis SE, Blake JA. 2012. On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: a short report. PLoS Comput Biol. 8: e1002386.

Associate editor: George Zhang