Vol 465 | 13 May 2010 | doi:10.1038/nature09014

A formal test of the theory of universal common ancestry

Douglas L. Theobald[1]

Universal common ancestry (UCA) is a central pillar of modern evolutionary theory[1]. As first suggested by Darwin[2], the theory of UCA posits that all extant terrestrial organisms share a common genetic heritage, each being the genealogical descendant of a single species from the distant past[3–6]. The classic evidence for UCA, although massive, is largely restricted to 'local' common ancestry—for example, of specific phyla rather than the entirety of life—and has yet to fully integrate the recent advances from modern phylogenetics and probability theory. Although UCA is widely assumed, it has rarely been subjected to formal quantitative testing[7–10], and this has led to critical commentary emphasizing the intrinsic technical difficulties in empirically evaluating a theory of such broad scope[1,5,8,9,11–15]. Furthermore, several researchers have proposed that early life was characterized by rampant horizontal gene transfer, leading some to question the monophyly of life[11,14,15]. Here I provide the first, to my knowledge, formal, fundamental test of UCA, without assuming that sequence similarity implies genetic kinship. I test UCA by applying model selection theory[5,16,17] to molecular phylogenies, focusing on a set of ubiquitously conserved proteins that are proposed to be orthologous. Among a wide range of biological models involving the independent ancestry of major taxonomic groups, the model selection tests are found to overwhelmingly support UCA irrespective of the presence of horizontal gene transfer and symbiotic fusion events. These results provide powerful statistical evidence corroborating the monophyly of all known life.

In the conclusion of *On the Origin of Species*, Darwin proposed that "all the organic beings which have ever lived on this earth have descended from some one primordial form"[2]. This theory of UCA—the proposition that all extant life is genetically related—is perhaps the most fundamental premise of modern evolutionary theory, providing a unifying foundation for all life sciences. UCA is now supported by a wealth of evidence from many independent sources[18], including: (1) the agreement between phylogeny and biogeography; (2) the correspondence between phylogeny and the palaeontological record; (3) the existence of numerous predicted transitional fossils; (4) the hierarchical classification of morphological characteristics; (5) the marked similarities of biological structures with different functions (that is, homologies); and (6) the congruence of morphological and molecular phylogenies[9,10]. Although the consilience of these classic arguments provides strong evidence for the common ancestry of higher taxa such as the chordates or metazoans, none expressly address questions such as whether bacteria, yeast and humans are all genetically related. However, the 'universal' in universal common ancestry is primarily supported by two further lines of evidence: various key commonalities at the molecular level[6] (including fundamental biological polymers, nucleic acid genetic material, L-amino acids, and core metabolism) and the near universality of the genetic code[4,7]. Notably, these two traditional arguments for UCA are largely qualitative, and typical presentations of the evidence do not assess quantitative measures of support for competing hypotheses, such as the probability of evolution from multiple, independent ancestors.

The inference from biological similarities to evolutionary homology is a feature shared by several of the lines of evidence for common ancestry. For instance, it is widely assumed that high sequence resemblance, often gauged by an E value from a BLAST search, indicates genetic kinship[19]. However, a small E value directly demonstrates only that two biological sequences are more similar than would be expected by chance[20]. A Karlin–Altschul E value is a Fisherian null-hypothesis significance test in which the null hypothesis is that two random sequences have been aligned[20]. Therefore, an E value in principle cannot provide evidence for or against the hypothesis that two sequences share a common ancestor. (In fact, an E value cannot even provide evidence for the random null hypothesis.[21]) Sequence similarity is an empirical observation, whereas the conclusion of homology is a hypothesis proposed to explain the similarity[22]. Statistically significant sequence similarity can arise from factors other than common ancestry, such as convergent evolution due to selection, structural constraints on sequence identity, mutation bias, chance, or artefact manufacture[19]. For these reasons, a sceptic who rejects the common ancestry of all life might nevertheless accept that universally conserved proteins have similar sequences and are 'homologous' in the original pre-Darwinian sense of the term (homology here being similarity of structure due to "fidelity to archetype")[23]. Consequently, it would be advantageous to have a method that is able to objectively quantify the support from sequence data for common-ancestry versus competing multiple-ancestry hypotheses.

Here I report tests of the theory of UCA using model selection theory, without assuming that sequence similarity indicates a genealogical relationship. By accounting for the trade-off between data prediction and simplicity, model selection theory provides methods for identifying the candidate hypothesis that is closest to reality[16,17]. When choosing among several competing scientific models, two opposing factors must be taken into account: the goodness of fit and parsimony. The fit of a model to data can be improved arbitrarily by increasing the number of free parameters. On the other hand, simple hypotheses (those with as few ad hoc parameters as possible) are preferred. Model selection methods weigh these two factors statistically to find the hypothesis that is both the most accurate and the most precise. Because model selection tests directly quantify the evidence for and against competing models, these tests overcome many of the well-known logical problems with Fisherian null-hypothesis significance tests (such as BLAST-style E values)[16,21]. To quantify the evidence supporting the various ancestry hypotheses, I applied three of the most widely used model selection criteria from all major statistical schools: the log likelihood ratio (LLR), the Akaike information criterion (AIC) and the log Bayes factor (LBF)[16,17].

Using these model selection criteria, I specifically asked whether the three domains of life (Eukarya, Bacteria and Archaea) are best

[1]Department of Biochemistry, Brandeis University, Waltham, Massachusetts 01778, USA.

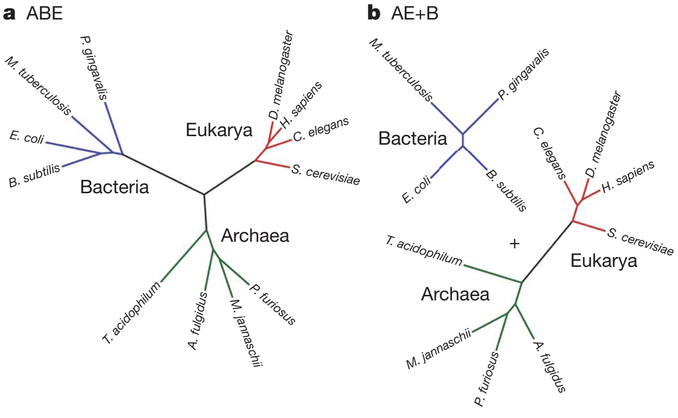

Figure 1 | Selected class I evolutionary hypotheses, excluding HGT. a, The model ABE, representing UCA of all taxa in the three domains of life. b, A competing multiple-ancestry model, AE+B, representing common ancestry of Archaea and Eukarya, but an independent ancestry for Bacteria. Trees shown are actual maximum likelihood estimates, with branch lengths proportional to the number of sequence substitutions.

described by a unified, common genetic relationship (that is, UCA) or by multiple groups of genetically unrelated taxa that arose independently and in parallel. As one example, a simplified model was considered for the hypothesis that Archaea and Eukarya share a common ancestor but do not share a common ancestor with Bacteria. This model (indicated by 'AE+B' in Fig. 1 and Table 1) comprises two independent trees—one containing Archaea and Eukarya and another containing only Bacteria. In these models the primary assumptions are: (1) that sequences change over time by a gradual, time-reversible Markovian process of residue substitution, described by a 20 × 20 instantaneous rate matrix defined by certain amino acid equilibrium frequencies and a symmetric matrix of amino acid exchangeabilities; (2) that new genetically related genes are generated by duplication during bifurcating speciation or gene duplication events; and (3) that residue substitutions are uncorrelated along different lineages and at different sites. The model selection tests evaluate how well these assumptions explain the given data set when various subsets of taxa and proteins are postulated to share ancestry, without any recourse to measures of sequence similarity.

The theory of UCA allows for the possibility of multiple independent origins of life[1–6]. If life began multiple times, UCA requires a 'bottleneck' in evolution in which descendants of only one of the independent origins have survived exclusively until the present (and the rest have become extinct), or, multiple populations with independent, separate origins convergently gained the ability to exchange essential genetic material (in effect, to become one species). All of the models examined here are compatible with multiple origins in both the above schemes, and therefore the tests reported here are designed to discriminate specifically between UCA and multiple ancestry, rather than between single and multiple origins of life. Furthermore, UCA does not demand that the last universal common ancestor was a single organism[24,25], in accord with the traditional evolutionary view that common ancestors of species are groups, not individuals[26]. Rather, the last universal common ancestor may have comprised a population of organisms with different genotypes that lived in different places at different times[25].

Table 1 | Class I hypotheses of single versus multiple ancestries

Hypothesis	$-\Delta K$	LLR	ΔAIC	LBF	ML evolutionary model
ABE	0	0	0	0	R-IGF
AE+B	17	6,569	6,586	6,889	(AE) R-IGF; (B) R-GF
AB+E	17	7,805	7,822	8,031	(AB) W-IGF; (E) R-GF
BE+A	18	8,192	8,210	8,488	(BE) R-IGF; (A) W-IGF
A+B+E	34	13,350	13,384	13,865	(E) R-GF; (B) R-GF; (A) W-IGF
ABE$_{-M}$+M	16	12,104	12,120	12,186	(ABE$_{-M}$) W-IF; (M) R-GF
ABE$_{-H}$+H	59	14,040	14,057	14,001	(ABE$_{-H}$) R-IGF; (H) empirical

Shown are the model section scores for class I hypotheses of single ancestry versus multiple ancestries, excluding HGT events. A, Archaea; B, Bacteria; E, Eukarya; H, Homo sapiens; M, Metazoa; ABE$_{-M}$, ABE without Metazoa; ABE$_{-H}$, ABE without H. sapiens. AE+B denotes a hypothesis of two independent ancestries, one tree for A and E together, and another separate tree for B. K denotes the total number of parameters in the model. All criteria are given as differences from ABE, so that larger values indicate less support for that model relative to ABE. LLR and ΔAIC scores correspond to the maximum likelihood (ML) estimates. For the ML evolutionary model, the first letter refers to the rate matrix: R, RtREV; W, WAG. The following letters denote models with additional parameters: I, invariant positions; G, gamma rate variation; F, empirical amino acid frequencies. The raw log likelihood for ABE is −126,299, and the marginal log likelihood is −126,713.

The data set consists of a subset of the protein alignment data from ref. 27, containing 23 universally conserved proteins for 12 taxa from all three domains of life, including nine proteins thought to have been horizontally transferred early in evolution[27]. The conserved proteins in this data set were identified based on significant sequence similarity using BLAST searches, and they have consequently been postulated to be orthologues. The first class of models I considered (presented in Table 1 and Fig. 1) constrains all the universally conserved proteins in a given set of taxa to evolve by the same tree, and hence these models do not account for possible horizontal gene transfer (HGT) or symbiotic fusion events during the evolution of the three domains of life. Hereafter I refer to this set of models as 'class I'. The class I model ABE, representing universal common ancestry of all taxa in the three domains of life and shown in Fig. 1a, can be considered to represent the classic three-domain 'tree of life' model of evolution[28].

Among the class I models, all criteria select the UCA tree by an extremely large margin (score differences ranging from 6,569 to 14,057), even though nearly half of the proteins in the analysis probably have evolutionary histories complicated by HGT. For all model selection criteria, by statistical convention a score difference of 5 or greater is viewed as very strong empirical evidence for the hypothesis with the better score (in this work higher scores are better)[16,17]. All scores shown are also highly statistically significant (the estimated variance for each score is approximately 2–3). According to a standard objective Bayesian interpretation of the model selection criteria, the scores are the log odds of the hypotheses[16,17]. Therefore, UCA is at least $10^{2,860}$ times more probable than the closest competing hypothesis. Notably, UCA is the most accurate and the most parsimonious hypothesis. Compared to the multiple-ancestry hypotheses, UCA provides a much better fit to the data (as seen from its higher likelihood), and it is also the least complex (as judged by the number of parameters).

The extraordinary strength of these results in the face of suspected HGT events suggests that the preference for the UCA model is robust to the extent of HGT. To test this possibility, the analysis was expanded to include models that allow each protein to have a distinct, independent evolutionary history. I refer to this set of models, which rejects a single tree metaphor for genealogically related taxa, as 'class II'. Representative class II models are shown in Fig. 2. Within each set of genealogically related taxa, each of the 23 universally conserved proteins is allowed to evolve on its own separate phylogeny, in which both branch lengths and tree topology are free parameters. For example, the multiple-ancestry model $[\mathrm{AE+B}]^{\mathrm{II}}$ comprises two clusters of protein trees, one cluster (AE) in which Archaea and Eukarya share a common ancestor but are genetically unrelated to another cluster (B) consisting only of Bacteria. Class II models are highly reticulate, phylogenetic networks that can represent very complex evolutionary mechanisms, including unrestricted HGT, symbiotic fusion events and independent ancestry of various taxa. Overall, the model selection tests show that the class II models are greatly preferred to the class I models. For instance, the class II UCA hypothesis ($[\mathrm{ABE}]^{\mathrm{II}}$) versus the class I UCA hypothesis (ABE) gives a highly significant LLR of 3,557, a ΔAIC of 2,633 and an LBF of 2,875. The optimal class II models represent an upper limit to the degree of HGT, as many of the apparent reticulations are probably due to incomplete lineage sorting, hidden paralogy, recombination, or inaccuracies in the evolutionary models. Nonetheless, as with the class I non-HGT hypotheses, all model selection criteria unequivocally support a single common genetic ancestry for all taxa. Also similar to the class I models, the class II UCA model has the greatest explanatory power and is the most parsimonious.

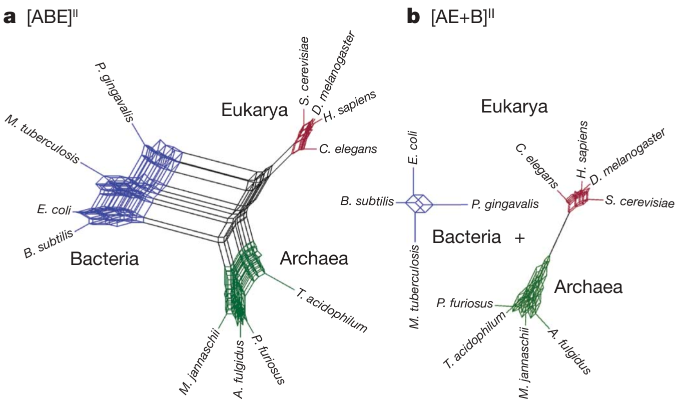

Figure 2 | Selected class II evolutionary hypotheses, including HGT. a, The reticulated model [ABE]$^{\mathrm{II}}$, representing UCA. **b**, A competing network model of multiple ancestry, [AE+B]$^{\mathrm{II}}$, representing common ancestry of Archaea and Eukarya, but a separate ancestry for Bacteria. Models are shown as phylogenetic networks (reticulate trees). The phylogenetic networks are derived from the maximum likelihood estimates of the 23 individual protein phylogenies using the evolutionary model parameters shown for ABE and AE+B in Table 1.

Several hypotheses have been proposed to explain the origin of eukaryotes and the early evolution of life by endosymbiotic fusion of an early archaeon and bacterium[29]. A key commonality of these hypotheses is the rejection of a single, bifurcating tree as a proper model for the ancestry of Eukarya. For instance, in these biological hypotheses certain eukaryotic genes are derived from Archaea whereas others are derived from Bacteria. The class II models freely allow eukaryotic genes to be either archaeal-derived or bacterial-derived, as the data dictate, and hence class II hypotheses can model several endosymbiotic 'rings' and HGT events. Because specific endosymbiotic fusion schemes can be represented by constrained versions of the unrestricted class II models, the endosymbiotic fusion hypotheses are nested within the class II hypotheses shown in Table 2. For nested hypotheses, the constrained versions necessarily have equal or lower likelihoods than the unconstrained versions. As a result, strict bounds can be placed on the LLR and ΔAIC scores for the constrained class II network models that represent specific endosymbiotic fusion or HGT hypotheses (see Methods and Supplementary Information). In all cases, these bounds show that multiple-ancestry versions of the constrained class II models are overwhelmingly rejected by the tests (model selection scores of several thousands), indicating that common ancestry is also preferred for all specific HGT and endosymbiotic fusion models. In terms of a fusion hypothesis for the origin of Eukarya, the data conclusively support a UCA model in which Eukarya share an ancestor with Bacteria and another independently with Archaea, and in which Bacteria and Archaea are also genetically related independently of Eukarya (see Table 3).

The proteins in this data set were postulated to be orthologous on the basis of significant sequence similarity[27]. Because the proteins are universally conserved, all of the taxa have their own specific versions of each of the proteins. It would be of interest to know how the tests respond to the inclusion of proteins that are not universally conserved, as omitting independently evolved proteins could perhaps bias the results towards common ancestry. Nevertheless, the inclusion of bona fide independently evolved genes has no effect on the likelihoods of the winning class II models, except in certain cases to strengthen the conclusion of common ancestry (for a formal proof, see the Supplementary Information). Many proteins probably do exist that have independent origins. For instance, in the Metazoa certain protein domains have probably evolved *de novo* that are not found in either Bacteria or Archaea[30]. However, the independent evolution of unique Metazoan proteins, by itself, is not evidence for or against UCA. The probability that the Metazoa would evolve a new protein domain is the same whether or not the Metazoa are related to Bacteria and Archaea. Therefore, omitting proteins with independent origins from the data set does not affect support for the UCA hypothesis versus multiple-ancestry hypotheses. In fact, including independently evolved proteins is expected to increase support for common ancestry for the subsets of taxa that share them (in this example, to increase support for common ancestry of the Metazoa).

Table 2 | Class II hypotheses of single versus multiple ancestries

Hypothesis	$-\Delta K$	LLR	ΔAIC	LBF
[ABE]$^{\mathrm{II}}$	0	0	0	0
[AE+B]$^{\mathrm{II}}$	391	7,642	8,033	8,124
[AB+E]$^{\mathrm{II}}$	391	8,473	8,864	8,864
[BE+A]$^{\mathrm{II}}$	414	8,829	9,243	9,333
[A+B+E]$^{\mathrm{II}}$	782	14,481	15,263	15,369
[ABE$_{-\mathrm{M}}$+M]$^{\mathrm{II}}$	391	12,061	12,452	12,512
[ABE$_{-\mathrm{H}}$+H]$^{\mathrm{II}}$	391	14,141	14,532	14,126

Shown are model selection scores for class II hypotheses of single ancestry versus multiple ancestries, allowing for unlimited HGT and/or endosymbiotic fusion events. Abbreviations are as in the Table 1 legend. All criteria are listed as differences from [ABE]$^{\mathrm{II}}$. All scores shown are highly statistically significant (the estimated variance for each score is approximately 3−6). The raw log likelihood for [ABE]$^{\mathrm{II}}$ is −122,742, and the marginal log likelihood is −123,838.

Table 3 | Class I and class II hypotheses for selected subsets

Hypotheses	$-\Delta K$	LLR	ΔAIC	LBF
AB versus A+B	17	5,545	5,562	5,837
BE versus B+E	16	5,157	5,173	5,380
AE versus A+E	17	6,782	6,899	6,979
[AB]$^{\mathrm{II}}$ versus [A+B]$^{\mathrm{II}}$	391	6,008	6,399	6,505
[BE]$^{\mathrm{II}}$ versus [B+E]$^{\mathrm{II}}$	368	5,652	6,020	6,036
[AE]$^{\mathrm{II}}$ versus [A+E]$^{\mathrm{II}}$	391	6,839	7,230	7,245

Shown are model selection scores for class I and II hypotheses for selected subsets of the taxa. Single ancestry hypotheses are listed left, multiple-ancestry hypotheses right. Terms are as in Table 1.

As is common in phylogenetic practice, most gaps and poorly aligned regions were removed from the original data set used in this analysis[27], leaving only those sites that were thought to be homologous with high confidence. To explore the effect of these omitted sites, the model selection tests were performed on a similar data set, with the same proteins and species, in which all gaps were kept in the final alignment (see Supplementary Methods and Supplementary Tables 5–8). The inclusion of these gapped and poorly aligned regions in the analyses greatly increases the support for UCA in all cases (for instance, with the ABE versus AE+B test, the class I ΔAIC is 10,323 and the class II ΔAIC is 11,072).

What property of the sequence data supports common ancestry so decisively? When two related taxa are separated into two trees, the strong correlations that exist between the sequences are no longer modelled, which results in a large decrease in the likelihood. Consequently, when comparing a common-ancestry model to a multiple-ancestry model, the large test scores are a direct measure of the increase in our ability to accurately predict the sequence of a genealogically related protein relative to an unrelated protein. The sequence correlations between a given clade of taxa and the rest of the tree would be eliminated if the columns in the sequence alignment for that clade were randomly shuffled. In such a case, these model-based selection tests should prefer the multiple-ancestry model. In fact, in actual tests with randomly shuffled data, the optimal estimate of the unified tree (for both maximum likelihood and Bayesian analyses) contains an extremely large internal branch separating the shuffled taxa from the rest. In all cases tried, with a wide variety of evolutionary models (from the simplest to the most parameter rich), the multiple-ancestry models for shuffled data sets are preferred by a large margin over common ancestry models (LLR on the order of a thousand), even with the large internal branches. Hence, the large test scores in favour of UCA models reflect the immense power of a tree structure, coupled with a gradual Markovian mechanism of residue substitution, to accurately and precisely explain the particular patterns of sequence correlations found among genealogically related biological macromolecules.

METHODS SUMMARY

All analyses were performed with 12 taxa, four from each domain of life, from the previously described data set comprising 23 ubiquitous proteins[27]. Archaea: *Methanococcus jannaschii*, *Archaeoglobus fulgidus*, *Pyrococcus furiosus* and *Thermoplasma acidophilum*; Eukarya: *Drosophila melanogaster*, *Homo sapiens*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*; Bacteria: *Escherichia coli*, *Bacillus subtilis*, *Mycobacterium tuberculosis* and *Porphyromonas gingivalis*. Optimal models were determined using both maximum likelihood and Bayesian phylogenetic methods. For a hypothesis involving several independent trees, such as model AE+B, each tree in the model was allowed to have its own independent evolutionary model parameters (such as amino acid substitution matrix, shape parameter for the gamma rate distribution, fraction of invariant sites, and empirical amino acid background frequencies), if it improved the likelihood. For a multiple-tree model such as AE+B, the total likelihood is simply the product of the individual likelihoods from each independent tree. Similarly, in a Bayesian analysis the total marginal likelihood is the product of marginal likelihoods from each independent tree. The AIC was calculated as $\mathrm{AIC} = L - K$, where L is the log likelihood and K is the total number of parameters in the model. Note that this differs from some common versions of the AIC by a factor of -2, and thus a maximum is preferred; this version was chosen for ease of comparison with the other test scores. No assumptions were made about the positions of the roots of the trees, as all inferred trees are unrooted. For the class II models involving HGT, each protein was given its own branch length and topology parameters; all other parameters were identical to the analogous class I model. The class II models thus implicitly assume that HGT involves the exchange of entire protein-coding genes. All phylogenetic input files are available by request.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 28 August 2009; accepted 17 March 2010.

1. Sober, E. *Evidence and Evolution* Ch. 4 (Cambridge University Press, 2008).
2. Darwin, C. *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life* Ch. 14 (J. Murray, 1859).
3. Raup, D. M. & Valentine, J. W. Multiple origins of life. *Proc. Natl Acad. Sci. USA* **80,** 2981–2984 (1983).
4. Crick, F. H. C. The origin of the genetic code. *J. Mol. Biol.* **38,** 367–379 (1968).
5. Sober, E. & Steel, M. Testing the hypothesis of common ancestry. *J. Theor. Biol.* **218,** 395–408 (2002).
6. Dobzhansky, T. Nothing in biology makes sense except in the light of evolution. *Am. Biol. Teach.* **35,** 125–129 (1973).
7. Hinegardner, R. T. & Engelberg, J. Rationale for a universal genetic code. *Science* **142,** 1083–1085 (1963).
8. Penny, D., Hendy, M. D. & Poole, A. M. Testing fundamental evolutionary hypotheses. *J. Theor. Biol.* **223,** 377–385 (2003).
9. Penny, D., Foulds, L. R. & Hendy, M. D. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature* **297,** 197–200 (1982).
10. Zuckerkandl, E. & Pauling, L. in *Evolving Genes and Proteins* (eds Bryson, V. & Vogel, H. J.) 97–166 (Academic Press, 1965).
11. Doolittle, W. F. The nature of the universal ancestor and the evolution of the proteome. *Curr. Opin. Struct. Biol.* **10,** 355–358 (2000).
12. How true is the theory of evolution? *Nature* **290** (Editorial), 75–76 (1981).
13. Popper, K. R. *Unended Quest: An Intellectual Autobiography* revised edn (Fontana, 1976).
14. Syvanen, M. On the occurrence of horizontal gene transfer among an arbitrarily chosen group of 26 genes. *J. Mol. Evol.* **54,** 258–266 (2002).
15. Woese, C. R. On the evolution of cells. *Proc. Natl Acad. Sci. USA* **99,** 8742–8747 (2002).
16. Burnham, K. P. & Anderson, D. R. *Model Selection and Inference: A Practical Information-Theoretic Approach* (Springer, 1998).
17. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90,** 773–795 (1995).
18. Futuyma, D. J. *Evolutionary Biology* 3rd edn (Sinauer Associates, 1998).
19. Murzin, A. G. How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8,** 380–387 (1998).
20. Karlin, S. & Altschul, S. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA* **87,** 2264–2268 (1990).
21. Harlow, L. L., Mulaik, S. A. & Steiger, J. H. *What If There Were No Significance Tests? (Multivariate Applications)* (Lawrence Erlbaum, 1997).
22. Reeck, G. *et al.* ''Homology'' in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell* **50,** 667 (1987).
23. Mindell, D. & Meyer, A. Homology evolving. *Trends Ecol. Evol.* **16,** 434–440 (2001).
24. Crick, F. H. C. in *Progress in Nucleic Acid Research* (eds Davidson, J. N. & Cohn, W. E.) 163–217 (Academic Press, 1963).
25. Doolittle, W. F. The practice of classification and the theory of evolution, and what the demise of Charles Darwin's tree of life hypothesis means for both of them. *Phil. Trans. R. Soc. Lond. B* **364,** 2221–2228 (2009).
26. Huxley, J. S. *Evolution: The Modern Synthesis* 2nd edn, 397–399 (G. Allen & Unwin, 1943).
27. Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. Universal trees based on large combined protein sequence data sets. *Nature Genet.* **28,** 281–285 (2001).
28. Woese, C. & Fox, G. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74,** 5088–5090 (1977).
29. Poole, A. & Penny, D. Evaluating hypotheses for the origin of eukaryotes. *Bioessays* **29,** 74–84 (2007).
30. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. Evolution of the protein repertoire. *Science* **300,** 1701–1703 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements I thank J. Felsenstein, P. Garrity, N. Matzke, C. Miller, C. Theobald and J. Wilkins for critical commentary.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence and requests for materials should be addressed to D.L.T. (dtheobald@brandeis.edu).

METHODS

Data sets. The original data set comprises 6,591 aligned amino acids from 23 ubiquitous proteins[27]: alanyl-tRNA synthetase, aspartyl-tRNA synthetase, glutamyl-tRNA synthetase, histidyl-tRNA synthetase, isoleucyl-tRNA synthetase, leucyl-tRNA synthetase, methionyl-tRNA synthetase, phenylalanyl-tRNA synthetase β subunit, threonyl-tRNA synthetase, valyl-tRNA synthetase, initiation factor 2, elongation factor G, elongation factor Tu, ribosomal protein L2, ribosomal protein S5, ribosomal protein S8, ribosomal protein S11, aminopeptidase P, DNA-directed RNA polymerase β chain, DNA topoisomerase I, DNA polymerase III γ subunit, signal recognition particle protein and rRNA dimethylase. The original data set was constructed by removing poorly aligned regions and most gapped columns from the CLUSTALW alignment[27]. I constructed a similar data set, using the same proteins from the same taxa, which retained the entire protein sequences. The proteins in this data set were independently aligned with ProbCons[31]. The resulting complete unmodified alignment comprised 25,411 columns, including gaps.

Likelihood phylogenetics. For the LLR and AIC tests, more than 1,800 competing biological models were fit to this data using the method of maximum likelihood and the program ProtTest 1.4 (ref. 32) (defaults) supplemented by independent runs with PhyML 2.4.5 (ref. 33). ProtTest calculates the maximum likelihood for 72 evolutionary models for each tree in each model: B, B-F, B-G, B-GF, B-I, B-IF, B-IG, B-IGF, C, C-F, C-G, C-GF, C-I, C-IF, C-IG, C-IGF, D, D-F, D-G, D-GF, D-I, D-IF, D-IG, D-IGF, J, J-F, J-G, J-GF, J-I, J-IF, J-IG, J-IGF, MM, MM-F, MM-G, MM-GF, MM-I, MM-IF, MM-IG, MM-IGF, MR, MR-F, MR-G, MR-GF, MR-I, MR-IF, MR-IG, MR-IGF, R, R-F, R-G, R-GF, R-I, R-IF, R-IG, R-IGF, V, V-F, V-G, V-GF, V-I, V-IF, V-IG, V-IGF, W, W-F, W-G, W-GF, W-I, W-IF, W-IG, and W-IGF, where the substitution matrices are coded as B = Blosum62, C = CtREV, D = Dayhoff, J = JTT, MM = MtMam, MR = MtREV, R = RtREV, V = VT, and W = WAG. The following letters denote models with further parameters: I = invariant positions, G = gamma distributed rate variation, F = empirical amino acid frequencies. For the class II HGT models, 23 different protein trees were calculated for each cluster of taxa proposed to be genealogically related. For example, the model $[AE+B]^{II}$ comprises 46 different trees—23 different protein trees for Archaea and Eukarya, and another 23 trees for Bacteria. The total log likelihood for a particular class II model is the sum of the log likelihoods for all the protein trees in the model.

Bayesian phylogenetics. All Bayesian analyses were calculated with the parallel version of MrBayes 3.1.2 (ref. 34) and used mixed-rate matrices and gamma-distributed rate variation across sites (16 categories). A uniform (0.0, 200.0) prior was assumed for the shape parameter of the gamma distribution, an unconstrained exponential prior (mean = 0.1) was assumed for the branch lengths, and a uniform prior was assumed for all topologies. Two independent Markov chain Monte Carlo (MCMC) analyses were performed (each with one cold and three heated chains), with all other parameters set to defaults. Convergence was inferred after the cold chain topologies had reached a standard deviation of split frequencies of less than 0.01 (generally never more than 10,000,000 generations). After convergence, the first half of the chain was discarded as 'burn in'. For the class II HGT models, the data were partitioned by protein, and all parameters (topology, branch lengths, state frequencies, amino acid substitution model and gamma shape) were unlinked across partitions.

Phylogenetic networks. Phylogenetic networks were computed and displayed with SplitsTree 4.10 (ref. 35), using the equal angle, consensus network algorithm (threshold = 0, to show all reticulations). The phylogenetic networks shown in Fig. 2 are derived from the maximum likelihood estimates of the 23 individual protein phylogenies using the evolutionary model parameters shown in Table 1.

Model selection test scores. LLR values were calculated directly from the likelihoods output by ProtTest and PhyML. The LLR test for non-nested hypotheses was used as previously described[36], which involves estimating the variance of a centred log likelihood using the per site likelihoods as output by PhyML. The number of parameters K was calculated as follows: one parameter per branch length for all trees in the model, where the number of branch lengths per tree is given by $2T-3$ (T is the number of taxa in a given tree); one parameter per tree if the number of invariant sites was estimated; one parameter per tree if the gamma-distribution shape parameter was estimated; 19 parameters per tree if the empirical amino acid frequencies were estimated. Marginal likelihoods for the Bayes factors were calculated with MrBayes[34] using the harmonic-mean estimator[17]. The LBF was calculated as the difference in the marginal-log likelihoods for each model.

Bounds on model selection scores. Consider three hypotheses: H_A, H_B and H_C. If H_B is a partially constrained hypothesis nested within H_C, then the following inequalities necessarily hold:

$$\mathrm{LLR}_{A-B} \geq L_A - L_C \tag{1}$$

$$\Delta\mathrm{AIC}_{A-B} \geq \mathrm{AIC}_A - L_C \tag{2}$$

where $\mathrm{LLR}_{A-B} = L_A - L_B$, $\Delta\mathrm{AIC}_{A-B} = \mathrm{AIC}_A - \mathrm{AIC}_B$, and L_X is the log likelihood for hypothesis H_X. These inequalities follow directly from the definitions of the model-selection scores and the fact that the likelihood for a nested, constrained hypothesis is always less than or equal to the likelihood of the unconstrained hypothesis[16]. Derivations and discussion are provided in the Supplementary Materials. The inequalities are especially useful for the purposes of this work, where H_A is a UCA hypothesis and H_B and H_C are multiple-ancestry hypotheses.

31. Do, C. B., Mahabhashyam, M. S., Brudno, M. & Batzoglou, S. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15**, 330–340 (2005).
32. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).
33. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).
34. Altekar, G. Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20**, 407–415 (2004).
35. Huson, D. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
36. Vuong, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–333 (1989).

1 Supplementary Tables

Table S1:

Raw likelihoods (PhyML) and marginal likelihoods (MrBayes) for optimal Class I models (original Brown *et al.* dataset, no gaps).

Hypothesis	K	log likelihood	marginal log likelihood
ABE	42	-126,299	-126,713
$AE + B$	59	-132,868	-133,602
$AB + E$	59	-134,104	-134,744
$BE + A$	60	-134,491	-135,201
$A + B + E$	76	-139,649	-140,578
$ABE^{-M} + M$	59	-138,403	-138,899
$ABE^{-H} + H$	59	-140,339	-140,713
AE	34	-82,765	-83,192
AB	34	-93,825	-94,168
BE	34	-85,224	-85,606
ABE^{-M}	35	-106,282	-106,552
ABE^{-H}	40	-121,200	-121,526
$A + B$	51	-99,370	-100,005
$A + E$	51	-89,547	-90,171
$B + E$	50	-90,381	-90,986
A	26	-49,268	-49,595
B	25	-50,102	-50,410
E	25	-40,279	-40,576
M	23	-32,121	-32,347
H	19	-19,139	-19,187

Table S2:

Raw likelihoods (PhyML) and marginal likelihoods (MrBayes) for optimal Class II models (original Brown *et al.* dataset, no gaps).

Hypothesis	K	log likelihood	marginal log likelihood
$[ABE]^{II}$	966	-122,742	-123,838
$[AE+B]^{II}$	1357	-130,384	-131,962
$[AB+E]^{II}$	1357	-131,215	-132,702
$[BE+A]^{II}$	1380	-131,571	-133,171
$[A+B+E]^{II}$	1748	-137,223	-139,207
$[ABE^{-M}+M]^{II}$	1357	-134,803	-136,350
$[ABE^{-H}+H]^{II}$	1357	-136,883	-137,964
$[AE]^{II}$	782	-81,254	-82,137
$[AB]^{II}$	782	-91,551	-92,407
$[BE]^{II}$	782	-83,142	-84,084
$[ABE^{-M}]^{II}$	805	-103,176	-104,139
$[ABE^{-H}]^{II}$	920	-117,744	-118,777
$[A+B]^{II}$	1173	-97,559	-98,912
$[A+E]^{II}$	1173	-88,093	-89,382
$[B+E]^{II}$	1150	-88,794	-90,120
$[A]^{II}$	598	-48,429	-49,087
$[B]^{II}$	575	-49,130	-49,825
$[E]^{II}$	575	-39,664	-40,295
$[M]^{II}$	529	-31,627	-32,211
$[H]^{II}$	19	-19,139	-19,187

Table S3:

Class II minimum model selection scores for constrained multiple-origin hypotheses, relative to the unconstrained single-origin hypothesis $[ABE]^{II}$ (i.e., reticulated UCA). The hypotheses listed in the lefthand column represent all constrained versions of the model. Original Brown *et al.* dataset, no gaps.

Hypothesis	min LLR	min ΔAIC	min ΔBIC
$[AE+B]^{II}$	7,642	6,676	3,395
$[AB+E]^{II}$	8,473	7,507	4,226
$[BE+A]^{II}$	8,829	7,863	4,582
$[A+B+E]^{II}$	14,481	13,515	10,234
$[ABE^{-M}+M]^{II}$	12,061	11,095	7,814
$[ABE^{-H}+H]^{II}$	14,141	13,175	9,894

Table S4:

Class II minimum model selection scores for single origin vs constrained multiple-origin hypotheses (original Brown *et al.* dataset, no gaps). The single-origin hypothesis (left) is unconstrained, whereas the multiple-origin hypotheses (right) are constrained.

Hypothesis	min LLR	min ΔAIC	min ΔBIC
$[AB]_{II}$ vs $[A+B]_{II}$	6,008	5,226	2,570
$[BE]_{II}$ vs $[B+E]_{II}$	5,652	4,870	2,214
$[AE]_{II}$ vs $[A+E]_{II}$	6,839	6,057	3,401

Table S5:

Raw likelihoods (PhyML) and marginal likelihoods (MrBayes) for optimal Class I models, all gaps included.

Hypothesis	K	log likelihood	marginal log likelihood
ABE	41	-338,997.1	-340,618
$AE+B$	59	-349,301.6	-351,472
$AB+E$	59	-351,171.6	-353,218
$BE+A$	59	-351,079.7	-353,230
$A+B+E$	75	-358,970.0	-361,491
$ABE^{-M}+M$	59	-358,941.1	-360,937
$ABE^{-H}+H$	59	-365,748.0	-367,495
AE	34	-231,313.2	-232,813
AB	34	-215,757.3	-217,295
BE	34	-245,512.4	-246,321
ABE^{-M}	35	-251,200.9	-252,831
ABE^{-H}	39	-311,184.5	-312,769
$A+B$	50	-223,555.7	-225,568
$A+E$	50	-240,981.6	-242,832
$B+E$	50	-253,402.7	-254,582
A	25	-105,567.3	-106,909
B	25	-117,988.4	-118,659
E	25	-135,414.3	-135,923
M	23	-107,740.1	-108,106
H	20	-54,563.5	-54,726

Table S6:

Hypothesis	ΔK	LLR	ΔAIC	LBF
$[ABE]$	0	0	0	0
$[AE+B]$	18	10,304.5	10,322.5	10,854
$[AB+E]$	18	12,174.5	12,192.5	12,600
$[BE+A]$	18	12,082.6	12,100.6	12,612
$[A+B+E]$	34	19,972.9	20,006.9	20,873
$[ABE^{-M}+M]$	18	19,944.0	19,962.0	20,319
$[ABE^{-H}+H]$	18	26,750.9	26,768.9	26,877

Table S7:

Raw likelihoods (PhyML) and marginal likelihoods (MrBayes) for optimal Class II models, all gaps included.

Hypothesis	K	log likelihood	marginal log likelihood
$[ABE]^{II}$	943	-334,075.8	-335,753
$[AE+B]^{II}$	1311	-344,780.0	-347,843
$[AB+E]^{II}$	1311	-346,704.8	-348,967
$[BE+A]^{II}$	1311	-346,442.7	-349,237
$[A+B+E]^{II}$	1725	-354,246.7	-358,166
$[ABE^{-M}+M]^{II}$	1357	-353,144.7	-356,391
$[ABE^{-H}+H]^{II}$	1357	-362,417.7	-363,862
$[AE]^{II}$	759	-227,875.1	-230,058
$[AB]^{II}$	759	-213,785.1	-214,758
$[BE]^{II}$	759	-242,020.6	-243,061
$[ABE^{-M}]^{II}$	805	-247,631.0	-249,661
$[ABE^{-H}]^{II}$	897	-307,854.2	-309,006
$[A+B]^{II}$	1150	-221,327.0	-223,961
$[A+E]^{II}$	1150	-237,341.8	-240,385
$[B+E]^{II}$	1150	-249,824.6	-251,994
$[A]^{II}$	575	-104,422.1	-106,176
$[B]^{II}$	575	-116,904.9	-117,785
$[E]^{II}$	575	-132,919.7	-134,209
$[M]^{II}$	529	-105,513.7	-106,730
$[H]^{II}$	20	-54,563.5	-54,855

Table S8:

Hypothesis	ΔK	LLR	ΔAIC	LBF
$[ABE]^{II}$	0	0	0	0
$[AE+B]^{II}$	368	10,704.2	11,072	12,090
$[AB+E]^{II}$	368	12,629.0	12,997	13,214
$[BE+A]^{II}$	368	12,366.9	12,735	13,484
$[A+B+E]^{II}$	782	20,170.9	20,953	22,413
$[ABE^{-M}+M]^{II}$	414	19,068.9	19,483	20,638
$[ABE^{-H}+H]^{II}$	414	28,341.9	28,756	28,109

2 Supplementary Equations and Discussion

2.1 Specific reticulate hypotheses are implicitly considered in these tests

Endosymbiotic fusion hypotheses, like the Hydrogen hypothesis[1] or the “ring of life”[2], propose a reticulated network, rather than a tree, to explain the history of genes. In both of these fusion hypotheses some eukaryotic genes are archaeal-derived while others are eubacterial-derived. The Class II models represent general hypotheses of this type, since the Class II models freely allow eukaryotic proteins to be either archaeal-derived or eubacterial-derived if the data so dictate. The Class II models are fully reticulate, phylogenetic networks that possibly contain both symbiotic “rings” and HGT events.

An example of a possible ring in one of the models is shown in Figure S1. This example is pulled directly from the Class II model $[\mathrm{ABE}]^{II}$. Two of the 23 protein trees are shown: one for Glu-tRNA synthetase (Figure S1A) and another for rRNA dimethylase (Figure S1B). The synthetase is an informational protein, and as one might expect the tree shows the eukaryotic version (red) branching from within Archaea (green), i.e. the eukaryotic protein is archaeal-derived. In contrast, the dimethylase is a metabolic protein, and, again as one might expect, the tree shows the eukaryotic proteins branching from within Bacteria (blue), i.e. the eukaryotic methylase is bacterial-derived. These two trees, taken at face value, can be reconciled only by invoking either a bona fide HGT event or a fusion event, where a ring can be created by drawing a line from the Eukarya to both Bacteria and Archaea.

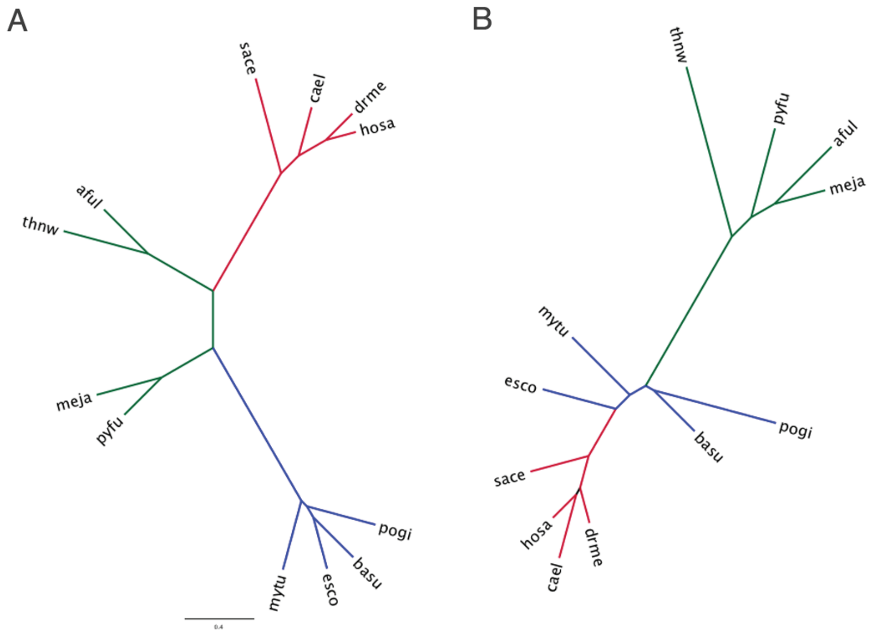

Figure S 1: A possible “ring” in Class II model $[\mathrm{ABE}]^{II}$. A. The individual unconstrained protein tree for Glu-tRNA synthetase. B. The protein tree for rRNA dimethylase. Archaeal proteins are in green, bacterial in blue, eukaryotic in red.

Importantly, the Class II models used in these tests are completely unconstrained, in the sense that every protein is free to have a different phylogeny with its own topology, branch lengths, background residue frequencies, invariant sites parameter, and gamma rate distribution. However, a particular Class II hypothesis, like the Hydrogen hypothesis, can be specified within this modeling framework by constraining certain proteins to evolve on specified phylogenies. For example, the hydrogen hypothesis might require that certain eukaryotic metabolic proteins, like the eukaryotic signal recognition protein, should be most closely related to the proteobacterium *Escherichia coli*. Specific Class II hypotheses, like the ring of life and hydrogen hypotheses, can be represented as constrained versions of the unrestricted Class II models shown in Table II of the main text. In this sense the endosymbiotic fusion hypotheses are nested within the unconstrained Class II hypotheses. Therefore, we can put strict bounds on the

LLR and ΔAIC scores for endosymbiotic fusion hypotheses (and other reticulated scenarios involving HGT). These bounds are a trivial consequence of the fact that nested hypotheses necessarily have a lower maximum likelihood (and fewer parameters) than the unconstrained hypothesis[3]. A derivation of these bounds is given below.

2.2 Bounds on model selection scores for nested models

Let H_R represent a Class II hypothesis that has not been tested directly, say, a model representing the "ring of life" hypothesis of Rivera and Lake. Let L_R represent the log likelihood for this "ring of life" hypothesis. Then, the AIC for this H_R model is given by:

$$\text{AIC}_R = L_R - K_R \tag{1}$$

where K_R is the number of parameters in the "ring of life" hypothesis. Note that this version of the AIC differs from some common versions be a factor of -2. Similarly, let L_U be the log likelihood for the corresponding unconstrained Class II hypothesis H_U:

$$\text{AIC}_U = L_U - K_U \tag{2}$$

Finally, let L_A be the log likelihood for a single origin Class I hypothesis (H_A). Its AIC is:

$$\text{AIC}_A = L_A - K_A \tag{3}$$

As explained above, if H_R is nested within the H_U hypothesis, then necessarily both $K_R \leq K_U$ and $L_R \leq L_U$ (both of which follow directly from the properties of nested hypotheses).

For testing the single-origin hypothesis H_A vs H_R, we have

$$\begin{aligned} \Delta\text{AIC}_{A-R} &= \text{AIC}_A - \text{AIC}_R && (4)\\ &= \text{AIC}_A - (L_R - K_R) && (5)\\ &= \text{AIC}_A - L_R + K_R && (6) \end{aligned}$$

where a positive ΔAIC_{A-R} will favor H_A, and a negative ΔAIC_{A-R} will favor H_R.

In general, unless we specify H_R exactly and calculate its maximum likelihood, we cannot know either L_R or K_R. However, because H_R is nested within H_U, we do know that $L_R \leq L_U$, so

$$\Delta\text{AIC}_{A-R} \geq \text{AIC}_A - L_U + K_R \tag{7}$$

And we also know that $K_R \geq 0$, so

$$\Delta\text{AIC}_{A-R} \geq \text{AIC}_A - L_U \tag{8}$$

So now we have an expression for the minimum ΔAIC_{A-R} in terms of known values. Analogous arguments prove that

$$\text{LLR}_{A-R} \geq L_A - L_U \tag{9}$$

Therefore, the AIC can possibly favor the nested H_R hypothesis only when $L_U > \text{AIC}_A$. These results always hold for any three hypotheses H_A, H_R, and H_U, whenever H_R is nested within H_U.

Inspection of Supplementary Tables S1-2, S5, and S7 reveals that all likelihoods for the unconstrained multiple-origin Class II hypotheses are substantially less than the scores for the corresponding single-origin hypotheses. Therefore, the minimum AIC scores are on the order of several thousand for a test of single-ancestry vs any of the multiple-ancestry Class II endosymbiotic fusion or HGT hypotheses. For instance, in a comparison of the single-origin model [ABE]^{II} with, say, a constrained hydrogen hypothesis version of the [AE+B]^{II} multiple-origin model, the ΔAIC must be greater than 6,676. In all cases, these bounds show that multiple origin hypotheses are rejected for constrained Class II models (results are summarized in Tables S3 and S4).

Bounds for the Bayesian marginal likelihood (and the log Bayes factor) are not as easily derived. However, the results of these tests clearly suggest that the Bayes factor will behave similarly to the likelihood scores.

2.3 Nested intermediate models show that the large ABE scores are due to common ancestry, rather than evolutionary processes

This analysis in fact considers a very wide variety of related evolutionary models, due to the nesting of simpler models within the general ones. For example, model ABE assumes a single tree with a single evolutionary process (substitution matrix, rate variation, etc.) along the entire tree. On the other hand, model AB+E assumes two trees and two evolutionary processes. The model selection tests highly prefer model ABE, but perhaps this has nothing to do with one versus two trees and is simply because one evolutionary process, rather than two, is most appropriate. Another possible test, then, would consider an intermediate model (AB+E)*, which assumes two trees but one model. However, this analysis already considers the intermediate model (AB+E)*, along with myriad similar variations. The intermediate model (AB+E)* is nested within (AB+E), and thus the intermediate model must have a likelihood less than or equal to (AB+E). All multiple origin models have likelihoods that are much lower than the single origin models, and they also have more parameters (see Tables S1, S2, S5, and S7). Therefore, the intermediate model (AB+E)* necessarily has both a much lower likelihood and a much lower AIC than the common ancestry hypothesis (ABE). This result follows directly from the inequalities derived above for model selection scores (Equations 8 and 9). For example, in this particular case, comparing ABE vs (AB+E)* yields a LLR $\geqslant 7805$ and a ΔAIC $\geqslant 7763$. Thus, a single evolutionary process cannot explain the advantage of the single origin model ABE. Rather, the increased explanatory power of ABE vs AB+E must be due to the unified, single origin tree.

3 Supplementary Methods and Results

3.1 Tests with randomly shuffled taxa

Multiple tests, similar to those reported for the real datasets, were performed on datasets in which the alignment columns of one or more taxa were randomly shuffled. For instance, I randomly shuffled the alignment columns of the eukaryotic taxa in the AE dataset (the eight archaeal and eukaryotic proteins). The maximum likelihood tree was found, and also a Bayesian phylogenetic analysis was performed on the same shuffled dataset with MrBayes (as described for the Class I methods). For simplicity the likelihood analyses all used the WAG substitution matrix. To test the effects of the complexity of the model, I performed two different types of likelihood analyses, one called "complex" that used gamma rate variation, invariant sites parameter, and empirical amino acid frequencies, and another called "simple" that used only a fixed gamma rate variation parameter. Multiple independently shuffled datasets were analyzed for each set of taxa. Similar analyses were performed with the AB dataset (the eight archaeal and bacterial proteins). Representative results are summarized in Table S9. In all of the likelihood analyses, the multiple-origin models were preferred by a very large margin by the model selection criteria. A typical tree, with branch lengths, is shown in Figure S2.

Table S9:

Results for analyses in which a portion of the taxa had randomly shuffled alignment columns. The model in parentheses contains shuffled taxa. “LLR” is the log likelihood ratio, “LBF” is the log Bayes factor, “std dev” is the standard deviation for the replicates, and “n” is the number of replicates.

Hypothesis	LLR/LBF	std dev	n
(AB) vs $A + B$, simple ML	-934	19	6
(AB) vs $A + B$, complex ML	-505	23	6
(AE) vs $A + E$, complex ML	-707	38	7
(AB) vs $A + B$, Bayes	-316	19	3

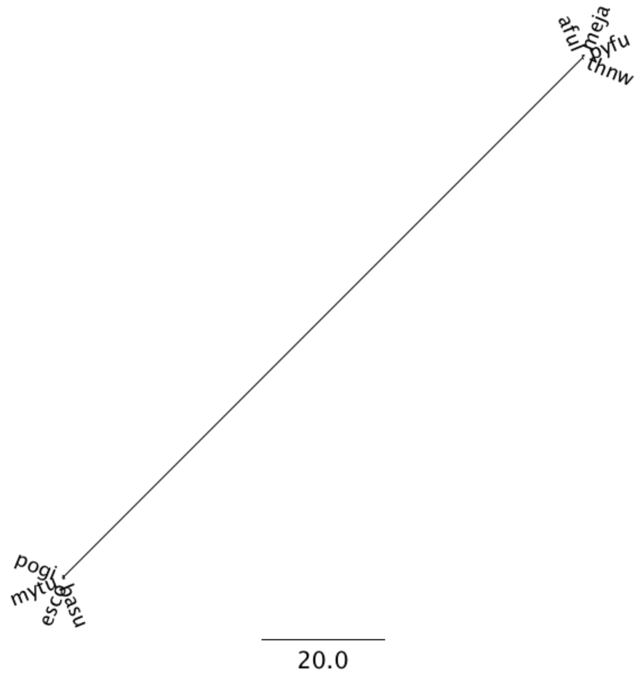

Figure S 2: A typical tree resulting from shuffling a subset of taxa. This tree represents the maximum likelihood estimate for a shuffled AB hypothesis (Archaea and Bacteria), in which the the alignment columns of the bacterial proteins were shuffled. The large internal branch is of length 121, whereas the remaining branches, too small to be displayed, average 0.25.

3.2 Omission of proteins with independent origins does not bias the model selection tests towards common ancestry

Consider two groups of taxa, A and B, each of which share a set of orthologous proteins ω. The data in ω is ordered in a sequence alignment, which can be represented as a matrix with N rows for each of the N taxa and S columns for each of the S sites in the alignment. The sequence data in ω also comprises multiple proteins (of total number P), so that ω can also be represented as a vector of proteins ($\omega = \{\omega_1, \omega_2, ..., \omega_P\}$). Now suppose that taxa A has an additional set of independently evolved proteins (set α), which is not found in taxa B. We would like to know how omitting or including this protein set α affects the model selection tests for the competing hypotheses of AB (taxa A and B sharing common ancestry) versus A+B (taxa A and B having independent origins). I will show that, for the Class II models (the winning models), there is no effect on the model selection tests.

Recall that the log-likelihood for a hypothesis H conditional on an observed data set D is defined as the natural logarithm of the probability of D given H:

$$\ell\left(H|D\right) = \ln \mathrm{p}\left(D|H\right) \tag{10}$$

For multiple pieces of independently distributed data (where $D = \{d_1, d_2, ..., d_m\}$ for m data points), the log-likelihood is simply the sum of the individual log-likelihoods:

$$\ell\left(H|D\right) = \sum_{i}^{m} \ell\left(H|d_i\right) \tag{11}$$

Thus, making the usual site independence assumption for the sequence data, the log-likelihood for the hypothesis AB and model parameters ϕ given the universally shared protein data set ω can be written as the sum:

$$\ell\left(AB, \phi|\omega\right) = \sum_{i}^{P} \ell\left(AB, \phi|\omega_i\right) \tag{12}$$

with an analogous expression for the hypothesis A+B. For the combined data set of both the universally shared proteins ω and the independently evolved proteins α, the log-likelihoods can also be expressed as sums:

$$\ell\left(AB, \phi|\omega, \alpha\right) = \ell\left(AB, \phi|\omega\right) + \ell\left(AB, \phi|\alpha\right) \tag{13}$$

$$\ell\left(A{+}B, \psi|\omega, \alpha\right) = \ell\left(A{+}B, \psi|\omega\right) + \ell\left(A{+}B, \psi|\alpha\right) \tag{14}$$

In the Class II models, each protein has its own independent set of parameters (tree topology, branch lengths, rate variation, substitution matrix, etc.). Each protein (each ω_i and α) is described by independent submodels, and so each protein set ω_i has a corresponding independent set of parameters ϕ_i for each of the P proteins in ω. The protein set α also has its own independent set of parameters ϕ_α:

$$\begin{aligned}\ell\left(AB, \phi|\omega, \alpha\right) &= \ell\left(AB, \phi_\omega|\omega\right) + \ell\left(AB, \phi_\alpha|\alpha\right) &(15)\\ &= \sum_{i}^{P} \ell\left(AB, \phi_i|\omega_i\right) + \ell\left(AB, \phi_\alpha|\alpha\right) &(16)\end{aligned}$$

$$\begin{aligned}\ell\left(A{+}B, \psi|\omega, \alpha\right) &= \ell\left(A{+}B, \psi_\omega|\omega\right) + \ell\left(A{+}B, \psi_\alpha|\alpha\right) &(17)\\ &= \sum_{i}^{P} \ell\left(A{+}B, \psi_i|\omega_i\right) + \ell\left(A{+}B, \psi_\alpha|\alpha\right) &(18)\end{aligned}$$

The log-likelihood ratio for hypotheses AB vs A+B given data set ω can be written:

$$\begin{aligned}\mathrm{LLR}(AB, A{+}B|\omega) &= \ell\left(AB, \hat{\phi}_\omega|\omega\right) - \ell\left(A{+}B, \hat{\psi}_\omega|\omega\right) &(19)\\ &= \sum_i^P \ell\left(AB, \hat{\phi}_i|\omega_i\right) - \sum_i^P \ell\left(A{+}B, \hat{\psi}_i|\omega_i\right) &(20)\end{aligned}$$

where $\hat{\phi}_\omega$ represents the ML estimate of ϕ based on data ω. Similarly, the log-likelihood ratio for hypotheses AB vs A+B given the full data set $\{\omega, \alpha\}$ is:

$$\begin{aligned}\mathrm{LLR}(AB, A{+}B|\omega, \alpha) = &\sum_i^P \ell\left(AB, \hat{\phi}_i|\omega_i\right) + \ell\left(AB, \hat{\phi}_\alpha|\alpha\right) &(21)\\ &- \sum_i^P \ell\left(A{+}B, \hat{\psi}_i|\omega_i\right) - \ell\left(A{+}B, \hat{\psi}_\alpha|\alpha\right) &(22)\end{aligned}$$

We can express the difference in the log-likelihood ratios (ΔLLR) for including versus omitting the data set α with independent origins by substitution:

$$\begin{aligned}\Delta\mathrm{LLR} &= \mathrm{LLR}(AB, A{+}B|\omega, \alpha) - \mathrm{LLR}(AB, A{+}B|\omega) &(23)\\ &= \ell\left(AB, \hat{\phi}_\alpha|\alpha\right) - \ell\left(A{+}B, \hat{\psi}_\alpha|\alpha\right) &(24)\end{aligned}$$

Now, the only difference between the two likelihoods above is that taxa A are attached to B in the first, but they are separated in the second. Since taxa B lacks the protein set α, all the B sites corresponding to α are gaps that represent "missing data", and these sites do not contribute to the likelihoods in either case. In both cases, the likelihood is maximized conditional on the data in taxa A alone. Hence, $\ell\left(AB, \hat{\phi}_\alpha|\alpha\right) = \ell\left(A, \hat{\phi}_\alpha|\alpha\right)$ and $\ell\left(A{+}B, \hat{\psi}_\alpha|\alpha\right) = \ell\left(A, \hat{\psi}_\alpha|\alpha\right)$. Therefore,

$$\Delta\mathrm{LLR} = \ell\left(A, \hat{\phi}_\alpha|\alpha\right) - \ell\left(A, \hat{\psi}_\alpha|\alpha\right) \tag{25}$$

However, both of the likelihoods above are also maximized over the same parameter space using the same data, i.e. $\hat{\phi}_\alpha$ and $\hat{\psi}_\alpha$ are identical. Thus, for Class II models, the two likelihoods above are equal:

$$\ell\left(A, \hat{\phi}_\alpha|\alpha\right) = \ell\left(A, \hat{\psi}_\alpha|\alpha\right) \tag{26}$$

and

$$\Delta\mathrm{LLR} = \ell\left(A, \hat{\phi}_\alpha|\alpha\right) - \ell\left(A, \hat{\psi}_\alpha|\alpha\right) = 0 \tag{27}$$

Hence, for the winning Class II models, the omission of proteins with independent origins has no effect on the log-likelihood ratios, nor, for that matter, on any of the model selection tests (since the number of parameters is also equivalent for ϕ and ψ). The only exception (not considered above) is when testing independent origin models that split up the taxa within set A. In that case, if the proteins are truly homologous within A and the models are approximately valid, including the α protein set is expected to favor common ancestry for A.

Equation (27) has been verified empirically, and those results are described in the next section.

3.3 Empirical tests with independently evolved proteins have no significant effect on the model selection scores

I took the original Brown et al. dataset of 23 proteins from 12 taxa, and included an additional 502 aa protein found only in the Eukarya. I then performed both the likelihood and Bayesian tests for two Class II analyses, one in which the eukaryotic protein was omitted, and one where it was included. The results are shown in Table S10.

Table S10

	Hypothesis	LL	marginal LL	LLR (relative to ABE)	Bayes factor
omit protein	$[ABE]^{II}$	-122,742	-123,838	0	0
	$[AB+E]^{II}$	-131,215	-132,702	8473	8864
include protein	$[ABE]^{II}$	-125,992	-127,071	0	0
	$[AB+E]^{II}$	-134,465	-135,933	8473	8862

As can be clearly seen, the likelihood ratios are unchanged whether the independently evolved protein is omitted from or included in the analysis (the small difference in the Bayes factor is well within the error of the estimates).

3.4 Inclusion of potential orthologs identified from structural and functional considerations alone

One may wonder how this analysis would fare if sequence similarity were not used as the criterion for including proteins in the analysis, since it is likely I have excluded true orthologs that have diverged so greatly that they no longer share significant sequence similarity. This method can be modified easily to include potential orthologs based only on structural and functional characteristics, rather than on sequence. For instance, the ribosome is shared among all three domains of life, and there are representative structures of each, including several large RNAs and many ribosomal proteins. One could choose to include in the analysis only ribosomal proteins that have similar biochemical functions, protein folds, and quaternary position in the ribosomal subunits, irrespective of sequence. Of course, such a structure/function-based method would be expected to primarily identify proteins with high sequence similarity (since the majority of these hypothesized orthologs in reality do have high sequence similarity). I have performed studies along these lines, by finding orthologs of the Archaeal ribosomal proteins from *Haloarcula marismortui* using only structural considerations and structure-based sequence alignments. As expected, most of the hypothesized orthologs have high sequence similarity. Even those proteins that have no statistically significant sequence similarity (according to BLAST) still support common ancestry in the tests (though to a lower extent than highly similar sequences), and none contradict common ancestry.

Here I provide one representative example from the ribosome in which sequence similarity is negligible. I used the 246 aa L4 ribosomal protein sequence from the Archaeal *Haloarcula marismortui* large subunit (X-ray crystal, 1jj2, chain C), for which there are several high resolution crystal structures. I searched the RCSB PDB structure database using Dali (http://ekhidna.biocenter.helsinki.fi/dali_server/) to find potential orthologs based on structure alone, regardless of any sequence similarity. The highest scoring hit in Eubacteria was the *Escherichia coli* L4 ribosomal protein (cryo-EM, 2gya, chain C). The highest scoring eukaryotic hit was the L4 ribosomal protein from *Saccharomyces cerevisiae* (cryo-EM, 1s1i, chain D). Only the yeast and archaeal proteins show significant sequence similarity (BLAST E-value = 1e-30), whereas the other have E-values > 10. These three L4 proteins were structurally aligned using MATT (http://groups.csail.mit.edu/cb/matt/), and the resulting sequence alignment (based on structural information alone), was used in the tests of common ancestry. The results are shown below in Table S11 for the contribution from L4 specifically.

Table S11

Hypothesis	LL	LLR (relative to ABE)
ABE	-1848.0	0
AE+B	-1878.4	30.4
AB+E	-1950.8	102.8
BE+A	-1957.1	109.1
A+B+E	-1977.3	129.3

As can be seen from the log-likelihood ratios, all the tests support the common ancestry of ribosomal protein L4, with the closest independent origin hypothesis 30.4 logs less likely (likelihood ratio $\approx 1e13$). Because L4 supports common ancestry individually, it will also support common ancestry when incorporated in the Class II models. Again, while 30.4 logs may seem somewhat small relative to the log-likelihood ratios found using the Brown et al. dataset, the LLR per site is 0.0897, which is analogous to 591 logs for a dataset as large as the Brown et al. dataset. The remaining ribosomal protein orthologs that could be identified from structural considerations behave similarly or provide even stronger support for common ancestry (largely because they have higher sequence similarity, such as is the case with L2). Hence, the omission of potential orthologs with negligible sequence similarity has only a marginal effect on the results.

4 Supplementary Notes

4.1 Problems with BLAST-style null-hypothesis tests

As alluded to in the main text of the paper, there are deep theoretical difficulties with inferring homology from sequence similarity detected by small E-values from pairwise BLAST searches. A BLAST E-value is a Fisherian null hypothesis significance test[4, 5]. According to the logic of null hypothesis testing, a small E-value allows us to reject the null hypothesis at some specified "level of significance"[3, 6, 7, 8]. With BLAST searches, the null-hypothesis holds that the observed alignment score was generated by the optimal alignment of two random sequences. However, rejecting this random null hypothesis is not logically equivalent to accepting common ancestry. This reasoning could be valid only if 'randomness' and 'common ancestry' were mutually exclusive hypotheses, but they are not. As discussed in the main text, significant sequence similarity (greater than random) can be due to many other factors besides common ancestry. For these reasons, the conclusions from model selection tests may be importantly different from conclusions based on E-value null hypothesis tests.

For instance, a small E-value strictly indicates only that the two sequences being compared have an alignment score that is improbable if they are random. By itself, however, a small probability is meaningless unless it can be shown that the observed alignment score is more probable if the two sequences are related. In some cases, the alignment score may be even *less* probable under the common ancestry hypothesis. If so, one should logically prefer the null hypothesis instead of homology, even in the face of a "statistically significant" E-value. An example of such a case is provided below (section 4.2).

In a rigorous objective comparison with the null hypothesis, common ancestry hypotheses also usually have more parameters (such as a branch length between the two sequences) that must be accounted for as explained in the introduction. These are well-known faults of frequentist null hypothesis tests that model selection tests overcome[3, 6, 7, 8]. Consequently, this is a great advantage of the model selection view taken in this analysis.

For the sake of argument, let us accept, *a priori* and contrary to the purpose of this analysis, the assumption that significant pairwise BLAST similarities suggest common ancestry for two proteins. Even given this assumption, there are still compelling reasons to suspect that these phylogenetic analyses may favor independent origins for certain sets of sequences with significant similarity. BLAST E-values apply solely to pairwise comparisons, where one query sequence is compared to one subject sequence. As such, BLAST E-values cannot directly establish the common ancestry of multiple proteins (such as all the conserved proteins from the three domains of life in this analysis). Measures of sequence similarity, like BLAST alignment scores, are not true distance metrics, because they do not satisfy the triangle inequality. Therefore, similarities among three or more sequences may conflict – two sequences A and B may be similar, and B and C may be similar, yet A and C can be quite dissimilar (even less similar than expected by pure chance). Thus, a more stringent and rigorous way to test the common ancestry for more than two proteins would be to model the possible relationships among the proteins, which may conflict. It is plausible, then, that the extra information imparted by a phylogenetic model may favor independent origins over common ancestry, even for a set of proteins with low BLAST E-values. A simple example of exactly this situation is given in section 4.3 below.

All these considerations demonstrate that one should not expect low pairwise E-values to automatically guarantee a conclusion of common ancestry in these phylogenetic model selection tests.

4.2 The "null hypothesis" may be favored even with significant BLAST similarity

The following two artificial protein sequences have high similarity as gauged by BLAST E-value (E = 2e-40, using the bl2seq utility with the database size of the current non-redundant database, 2,165,046,588). The BLAST output is shown below. Using the BLOSUM62 background frequencies, the log-likelihood for the "null hypothesis", that these two sequences are unrelated and the similarities between them are entirely due to chance, is -3667.4. A competing common ancestry hypothesis, using the BLOSUM62 transition matrix as implemented in PhyML and a branch length of 0.65 gives a log-likelihood of -3684.5 – which is 17.1 logs lower. Thus, the AIC is 16.1, and the random null hypothesis is roughly 10 million times more likely.

```
>SEQ1
KSTAEIRQSKGHQYSRATTSQRCVRAGGKHNDLENVGGDHDAIQFAWYESDDEAYIIRIG
DNKGAPYASDNFWQMGDTGPCGPTEIFYDHGMEPLTGMGLERIAAVLQHVNSNYDIRTLI
QAVAKVTGATDLSNKSLRVIADHIRSCAFLIADGVMPSNENRGYVRAVRHGNMAKETFFY
KLVGPLIDVMGSAGEDLKRQQAQVEQVLKTEEEQFARTLERGLALLDEELATAFRLYDTY
GFPVTADVCRERNIKVDEAGFEAAMEEQRRRAREASGFGADYNAMIRVDSASEFKGYDHL
ELNGKVALFVDGKNAGQEAVVVLDQTPFYAESGGQVGDKGELKSFAVEDTQKYGQAIGHI
GKLAAGSLKVGDAVQADVDEARRARIRLNHSATHLMHAALRQVLGTHVSQKGSLVNDKVS
HNEAMKPERAVEDLVNTQIRRNLNIMDLEAAKAKGAMALFGEKYDERVRELCGGTHASRT
GDIGLFRIISAGVRRIEAVTGEGAIATVHADSDRLVAHLLKNLADKVRSVLERTRQLQLK
EQAAAQESANLSSKAIDVNGVKLLVSEEPKMLRTMVDDLKNQLGSTIIVLATVVEGKVSL
IAGVSKDVTDRVKAGELIGMVAQQVGGKGGGRP

>SEQ2
MPINEIRVQKGQRYRDAINDKKAVKKGDKFLDLDEIAGEDEAVQPVYLDSEDGLYILRHS
AAHLLANAVTNLFPLPNTGPVVENGFYYDFDMKPITEEDLSKIEEEMKRVKENVPIRRMI
YSKDELLKIFSKNPYKIRIINENVEGKSSVYQQGNFPHVPSTGYIKAFKLLSIASAVYKY
DESKNLVRIYGTAFPDEKSLRRYLENLEEAKKRDHRKIIEMDLAVFNSEWAPGFPMYTPN
GQIIRKELIKYMDYVNGKNGWTDVWTPHKDTIWKQSGHYAKYNMYLFVLPDGDSYGIKPM
NCPGHIAIFARRKYSYRDLPVKYSETVYRYEKSGEVGGLTRPRAFTQDDGHEFDQIVGEI
KTLLGMVRETFTTVFGNIEMADLSVIDKEHPENYLLSYVCKDCGNRVEGLRGTDIECPVC
HSHNMDPDFSTWDNATEQLRQAMDSMGITYKEYPGEAAFYGPKIDVHVKDALGRMWQLST
IQVDFFMPINFGLTYTNSEGKEERVVIIHGSYERFMAILLEHFAGKLPTWLTPIQTYVIP
VGTANAEYARKVNKSLLDAGIRSVVDDGPDTVSKKIKMIHDQRPSYIVVVGAKEEQDNTV
TVRNRAGKSKTYGMNEFLEIIKNEIEKRSVGQA
```

```
Query= SEQ1
         (633 letters)



>SEQ2
          Length = 633

 Score =  172 bits (435), Expect = 2e-40
 Identities = 127/591 (21%), Positives = 255/591 (43%)

Query: 5   EIRQSKGHQYSRATTSQRCVRAGGKHNDLENVGGDHDAIQFAWYESDDEAYIIRIGDNKG 64
           EIR  KG +Y  A   ++ V+ G K  DL+ + G+ +A+Q  + +S+D  YI+R
Sbjct: 5   EIRVQKGQRYRDAINDKKAVKKGDKFLDLDEIAGEDEAVQPVYLDSEDGLYILRHSAAHL 64

Query: 65  APYASDNFWQMGDTGPCGPTEIFYDHGMEPLTGMGLERIAAVLQHVNSNYDIRTLIQAVA 124
              A  N + + +TGP      +YD  M+P+T   L +I   ++ V  N  IR +I +
Sbjct: 65  LANAVTNLFPLPNTGPVVENGFYYDFDMKPITEEDLSKIEEEMKRVKENVPIRRMIYSKD 124

Query: 125 KVTGATDLSNKSLRVIADHIRSCAFLIADGVMPSNENRGYVRAVRHGNMAKETFFYKLVG 184
           ++      +   +R+I +++   + +   G  P   + GY++A +  ++A   + Y
Sbjct: 125 ELLKIFSKNPYKIRIINENVEGKSSVYQQGNFPHVPSTGYIKAFKLLSIASAVYKYDESK 184

Query: 185 PLIDVMGSAGEDLKRQQAQVEQVLKTEEEQFARTLERGLALLDEELATAFRLYDTYGFPV 244
            L+ + G+A  D K  +  +E + + ++    + +E  LA+ + E A  F +Y   G  +
Sbjct: 185 NLVRIYGTAFPDEKSLRRYLENLEEAKKRDHRKIIEMDLAVFNSEWAPGFPMYTPNGQII 244

Query: 245 TADVCRERNIKVDEAGFXXXXXXXXXXXXXXSGFGADYNAMIRVDSASEFKGYDHLELNG 304
             ++ +  +    + G+              SG  A YN  + V    +  G   +   G
Sbjct: 245 RKELIKYMDYVNGKNGWTDVWTPHKDTIWKQSGHYAKYNMYLFVLPDGDSYGIKPMNCPG 304

Query: 305 KVALFVDGKNAGQEAVVVLDQTPFYAESGGQVGDKGELKSFAVEDTQKYGQAIGHIGKLA 364
            +A+F   K + ++  V   +T +  E  G+VG     ++F  +D  ++ Q +G I  L
Sbjct: 305 HIAIFARRKYSYRDLPVKYSETVYRYEKSGEVGGLTRPRAFTQDDGHEFDQIVGEIKTLL 364

Query: 365 AGSLKVGDAVQADVDEARRARIRLNHSATHLMHAALRQVLGTHVSQKGSLVNDKVSHNEA 424
               +    V  +++ A  + I   H   +L+    +         +G+ +   V H+
Sbjct: 365 GMVRETFTTVFGNIEMADLSVIDKEHPENYLLSYVCKDCGNRVEGLRGTDIECPVCHSHN 424

Query: 425 MKPERAVEDLVNTQIRRNLNIMDLEAAKAKGAMALFGEKYDERVRELCGGTHASRTGDIG 484
           M P+ +  D    Q+R+ ++ M +   +  G  A +G K D  V++  G      T  +
Sbjct: 425 MDPDFSTWDNATEQLRQAMDSMGITYKEYPGEAAFYGPKIDVHVKDALGRMWQLSTIQVD 484

Query: 485 LFRIISAGVRRIEAVTGEGAIATVHADSDRLVAHLLKNLADKVRSVLERTRXXXXXXXXX 544
            F  I+ G+    +   E  +  +H   +R +A LL++ A K+ + L   +
Sbjct: 485 FFMPINFGLTYTNSEGKEERVVIIHGSYERFMAILLEHFAGKLPTWLTPIQTYVIPVGTA 544

Query: 545 XXXXXNLSSKAIDVNGVKLLVSEEPKMLRTMVDDLKNQLGSTIIVLATVVE 595
                   +K++   G++ +V + P  +   +  + +Q  S I+V+    E
Sbjct: 545 NAEYARKVNKSLLDAGIRSVVDDGPDTVSKKIKMIHDQRPSYIVVVGAKEE 595


Lambda     K      H
   0.317    0.134    0.377

Gapped
Lambda     K      H
   0.267   0.0410    0.140


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 1
Number of Hits to DB: 771
Number of extensions: 33
Number of successful extensions: 3
Number of sequences better than 10.0: 1
Number of HSP’s gapped: 1
Number of HSP’s successfully gapped: 1
Length of query: 633
Length of database: 2,165,046,588
Length adjustment: 144
Effective length of query: 489
Effective length of database: 2,165,046,588
Effective search space: 1058707781532
Effective search space used: 1058707781532
Neighboring words threshold: 11
Window for multiple hits: 40
X1: 16 ( 7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (21.6 bits)
S2: 84 (37.0 bits)
```

4.3 Independent ancestry favored in a phylogenetic test for three highly similar proteins due to the "triangle inequality"

The following three artificial (i.e. intelligently designed) 480 aa protein sequences have high sequence similarity as gauged by BLAST E-values (again using the bl2seq utility with the database size of the current non-redundant database, 2,165,046,588). The 'mytu1' sequence is the query sequence, and both the 'esco1' and 'pogi1' sequences are highly significant 'hits' with very low E-values (1e-116 and 1e-122 respectively). In contrast, 'esco1' and 'pogi1' have undetectable sequence similarity between them. The optimal alignment from MUSCLE and the BLAST results are shown below.

Let the common ancestry hypothesis, that all three proteins are related, be MEP. A competing independent ancestry hypothesis ME+P represents the case where mytu1 and esco1 share a common ancestor to the exclusion of pogi1 (with analogous definitions for MP+E and EP+M). The log-likelihoods are shown in Table S12, using the WAG matrix with gamma distributed rate variation. The best hypothesis (with the highest likelihood) is MP+E. The common ancestry hypothesis, MEP, is relatively disfavored by 15.9 logs, which means it is about 8 million times less likely than the best independent origins hypothesis. While 15.9 logs may appear small compared to the log-likelihood ratios for the tests reported in the main text, the LLR per site is 0.0331, which is equivalent to 218 logs for a dataset with 6591 sites, such as the Brown et al. dataset used in these analyses.

Table S12

Hypothesis	LL	LLR (relative to MEP)
MEP	-4170.1	0
MP+E	-4154.2	-15.9
EP+M	-4612.8	442.7
ME+P	-4161.2	-8.9

```
>esco1
AWHPWDPAAAWNADHHAHHWWWPDPNWAAHHADWNADDAADAPWHWDAHWAAHWWWHPWWAHAHPPWAPA
NAPAAAHHAHPWWAWNPPPAPDAHHNWAADPWNANADPDNWAPWNDWAAWWPNWHDDDNHHANWNDPHWP
PAWWNDWHANNDAANWNPDNNAAHADAWAWNAAWHHNWAPDHPWAWAAAAAWPWHAWNWWWAANWNDWAH
WWWDHWWAHADWWHWAAWWNDPPPAPWAWWDHHDWHWPPAAPWAHDNANDWPWNPDPPDAWADPWNAWWH
PNWNPWWHWDWWWDWWHDDAPHWNAHAWDDDPDNDNPANADDWWNANHWWANWNWAWAHDAAPADWHAPA
HWHWDNWPHDPAADWWANANAWAWWHPPWHWPHWNNWPAPPAANNPAWWDDAAAWWNPHAWWDWADAPDW
NNADHHAHDNAHWAWAWAPHAAHDAWPWANAWAWWNAHADPAAHAHWDANNDNPWANNWN
>mytu1
CDHPDDPRRCPNRDHHRHHPDPPDPNPCCHHRDDNRDDRRDCPPHPDCHDCRHDDDHPDPRHRHPPPCPR
NCPCCCHHRHPDPRDNPPPRPDCHHNDRRDPDNCNRDPDNDCPPNDPCCDPPNDHDDDNHHRNDNDPHDP
PRDDNDPHRNNDRCNDNPDNNCRHCDRDCDNCRPHHNDRPDHPPRPCRRRRPPDHCDNPPDRCNDNDPRH
DDPDHPDCHRDPDHDRCDDNDPPPCPPCDPDHHDPHPPPRCPPCHDNRNDDPPNPDPPDCDRDPDNRPDH
PNPNPPPHPDDDDDDPHDDRPHPNCHRPDDDPDNDNPCNRDDPDNCNHPDCNPNPRPCHDCRPCDPHRPR
HPHPDNPPHDPCCDDPCNRNCPCPDHPPDHPPHPNNPPRPPCCNNPCPDDDRRCDDNPHRDPDDRDRPDP
NNRDHHCHDNCHPCDCDCPHRCHDRDPDCNCDRPPNRHCDPRCHCHDDRNNDNPPCNNPN
>pogi1
CDGYDYYRRCPGRYGGRGGPDPYYYGPCCGGRYDGRYYRRYCYPGPYCGDCRGDDDGYDPRGRGYYPCYR
GCYCCCGGRGYDPRDGYYYRYYCGGGDRRYYDGCGRYYYGDCYPGYPCCDPYGDGYYYGGGRGDGYYGDY
YRDDGYPGRGGYRCGDGYYGGCRGCYRDCDGCRPGGGDRYYGYPRPCRRRRPYDGCDGPPDRCGDGYPRG
DDPYGPDCGRYPDGDRCDDGYYYYCYPCDPYGGYPGPYYRCYPCGYGRGYDYPGYYYYYCDRYYDGRPDG
YGPGYPPGPYDDDYDPGYYRYGPGCGRPYYYYYGYGYCGRYYPDGCGGPDCGPGPRPCGYCRYCYPGRYR
GPGPYGPYGYYCCYDPCGRGCPCPDGYYDGPYGPGGPYRYYCCGGYCPDYYRRCDDGYGRDPYDRYRYYP
GGRYGGCGYGCGPCDCDCYGRCGYRDYDCGCDRPPGRGCYYRCGCGDYRGGYGYPCGGPG
```

```
CLUSTAL W (1.83) multiple sequence alignment


esco1           AWHPWDPAAAWNADHHAHHWWWPDPNWAAHHADWNADDAADAPWHWDAHWAAHWWWHPWW
mytu1           CDHPDDPRRCPNRDHHRHHPDPPDPNPCCHHRDDNRDDRRDCPPHPDCHDCRHDDDHPDP
pogi1           CDGYDYYRRCPGRYGGRGGPDPYYYGPCCGGRYDGRYYRRYCYPGPYCGDCRGDDDGYDP

esco1           AHAHPPWAPANAPAAAHHAHPWWAWNPPPAPDAHHNWAADPWNANADPDNWAPWNDWAAW
mytu1           RHRHPPPCPRNCPCCCHHRHPDPRDNPPPRPDCHHNDRRDPDNCNRDPDNDCPPNDPCCD
pogi1           RGRGYYPCYRGCYCCCGGRGYDPRDGYYYRYYCGGGDRRYYDGCGRYYYGDCYPGYPCCD

esco1           WPNWHDDDNHHANWNDPHWPPAWWNDWHANNDAANWNPDNNAAHADAWAWNAAWHHNWAP
mytu1           PPNDHDDDNHHRNDNDPHDPPRDDNDPHRNNDRCNDNPDNNCRHCDRDCDNCRPHHNDRP
pogi1           PYGDGYYYGGGRGDGYYGDYYRDDGYPGRGGYRCGDGYYGGCRGCYRDCDGCRPGGGDRY

esco1           DHPWAWAAAAAWPWHAWNWWWAANWNDWAHWWWDHWWAHADWWHWAAWWNDPPPAPWAWW
mytu1           DHPPRPCRRRRPPDHCDNPPDRCNDNDPRHDDPDHPDCHRDPDHDRCDDNDPPPCPPCDP
pogi1           YGYPRPCRRRRPYDGCDGPPDRCGDGYPRGDDPYGPDCGRYPDGDRCDDGYYYYCYPCDP

esco1           DHHDWHWPPAAPWAHDNANDWPWNPDPPDAWADPWNAWWHPNWNPWWHWDWWWDWWHDDA
mytu1           DHHDPHPPPRCPPCHDNRNDDPPNPDPPDCDRDPDNRPDHPNPNPPPHPDDDDDDPHDDR
pogi1           YGGYPGPYYRCYPCGYGRGYDYPGYYYYYCDRYYDGRPDGYGPGYPPGPYDDDYDPGYYR

esco1           PHWNAHAWDDDPDNDNPANADDWWNANHWWANWNWAWAHDAAPADWHAPAHWHWDNWPHD
mytu1           PHPNCHRPDDDPDNDNPCNRDDPDNCNHPDCNPNPRPCHDCRPCDPHRPRHPHPDNPPHD
pogi1           YGPGCGRPYYYYYGYGYCGRYYPDGCGGPDCGPGPRPCGYCRYCYPGRYRGPGPYGPYGY

esco1           PAADWWANANAWAWWHPPWHWPHWNNWPAPPAANNPAWWDDAAAWWNPHAWWDWADAPDW
mytu1           PCCDDPCNRNCPCPDHPPDHPPHPNNPPRPPCCNNPCPDDDRRCDDNPHRDPDDRDRPDP
pogi1           YCCYDPCGRGCPCPDGYYDGPYGPGGPYRYYCCGGYCPDYYRRCDDGYGRDPYDRYRYYP

esco1           NNADHHAHDNAHWAWAWAPHAAHDAWPWANAWAWWNAHADPAAHAHWDANNDNPWANNWN
mytu1           NNRDHHCHDNCHPCDCDCPHRCHDRDPDCNCDRPPNRHCDPRCHCHDDRNNDNPPCNNPN
pogi1           GGRYGGCGYGCGPCDCDCYGRCGYRDYDCGCDRPPGRGCYYRCGCGDYRGGYGYPCGGPG
```

```
Query= mytu1
         (480 letters)



>esco1
          Length = 480

 Score =  422 bits (1084), Expect = e-116
 Identities = 240/478 (50%), Positives = 240/478 (50%)

Query: 3   HPDDPRRCPNRDHHRHHPDPPDPNPCCHHRDDNRDDRRDCPPHPDCHDCRHDDDHPDPRH 62
           HP DP    N DHH HH   PDPN   HH D N DD  D P H D H   H   HP   H
Sbjct: 3   HPWDPAAAWNADHHAHHWWWPDPNWAAHHADWNADDAADAPWHWDAHWAAHWWWHPWWAH 62

Query: 63  RHPPPCPRNCPCCCHHRHPDPRDNPPPRPDCHHNDRRDPDNCNRDPDNDCPPNDPCCDPP 122
            HPP  P N P   HH HP    NPPP PD HHN   DP N N DPDN  P ND     P
Sbjct: 63  AHPPWAPANAPAAAHHAHPWWAWNPPPAPDAHHNWAADPWNANADPDNWAPWNDWAAWWP 122

Query: 123 NDHDDDNHHRNDNDPHDPPRDDNDPHRNNDRCNDNPDNNCRHCDRDCDNCRPHHNDRPDH 182
           N HDDDNHH N NDPH PP   ND H NND  N NPDNN  H D    N   HHN  PDH
Sbjct: 123 NWHDDDNHHANWNDPHWPPAWWNDWHANNDAANWNPDNNAAHADAWAWNAAWHHNWAPDH 182

Query: 183 PPRPCRRRRPPDHCDNPPDRCNDNDPRHDDPDHPDCHRDPDHDRCDDNDPPPCPPCDPDH 242
           P         P H  N     N ND  H   DH   H D  H     NDPPP P    DH
Sbjct: 183 PWAWAAAAAWPWHAWNWWWAANWNDWAHWWWDHWWAHADWWHWAAWWNDPPPAPWAWWDH 242

Query: 243 HDPHPPPRCPPCHDNRNDDPPNPDPPDCDRDPDNRPDHPNPNPPPHPDDDDDDPHDDRPH 302
           HD H PP  P  HDN ND P NPDPPD   DP N   HPN NP  H D   D  HDD PH
Sbjct: 243 HDWHWPPAAPWAHDNANDWPWNPDPPDAWADPWNAWWHPNWNPWWHWDWWWDWWHDDAPH 302

Query: 303 PNCHRPDDDPDNDNPCNRDDPDNCNHPDCNPNPRPCHDCRPCDPHRPRHPHPDNPPHDPC 362
            N H  DDDPDNDNP N DD  N NH   N N    HD  P D H P H H DN PHDP
Sbjct: 303 WNAHAWDDDPDNDNPANADDWWNANHWWANWNWAWAHDAAPADWHAPAHWHWDNWPHDPA 362

Query: 363 CDDPCNRNCPCPDHPPDHPPHPNNPPRPPCCNNPCPDDDRRCDDNPHRDPDDRDRPDPNN 422
            D   N N     HPP H PH NN P PP  NNP   DD     NPH   D  D PD NN
Sbjct: 363 ADWWANANAWAWWHPPWHWPHWNNWPAPPAANNPAWWDDAAAWWNPHAWWDWADAPDWNN 422

Query: 423 RDHHCHDNCHPCDCDCPHRCHDRDPDCNCDRPPNRHCDPRCHCHDDRNNDNPPCNNPN 480
            DHH HDN H      PH  HD  P  N     N H DP  H H D NNDNP  NN N
Sbjct: 423 ADHHAHDNAHWAWAWAPHAAHDAWPWANAWAWWNAHADPAAHAHWDANNDNPWANNWN 480


Lambda     K      H
   0.320    0.149    0.595

Gapped
Lambda     K      H
   0.267   0.0410    0.140


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 1
Number of Hits to DB: 4338
Number of extensions: 804
Number of successful extensions: 5
Number of sequences better than 10.0: 1
Number of HSP's gapped: 1
Number of HSP's successfully gapped: 1
Length of query: 480
Length of database: 2,165,046,588
Length adjustment: 141
Effective length of query: 339
Effective length of database: 2,165,046,588
Effective search space: 733950793332
Effective search space used: 733950793332
Neighboring words threshold: 11
Window for multiple hits: 40
X1: 16 ( 7.4 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (21.7 bits)
S2: 82 (36.2 bits)
```

```
Query= mytu1
         (480 letters)


>pogi1
          Length = 480

 Score =  443 bits (1140), Expect = e-122
 Identities = 240/479 (50%), Positives = 240/479 (50%)

Query: 1   CDHPDDPRRCPNRDHHRHHPDPPDPNPCCHHRDDNRDDRRDCPPHPDCHDCRHDDDHPDP 60
           CD  D  RRCP R   R  PDP    PCC  R D R  RR C P P C DCR DDD  DP
Sbjct: 1   CDGYDYYRRCPGRYGGRGGPDPYYYGPCCGGRYDGRYYRRYCYPGPYCGDCRGDDDGYDP 60

Query: 61  RHRHPPPCPRNCPCCCHHRHPDPRDNPPPRPDCHHNDRRDPDNCNRDPDNDCPPNDPCCD 120
           R R   PC R C CCC  R  DPRD    R  C   DRR  D C R    DC P  PCCD
Sbjct: 61  RGRGYYPCYRGCYCCCGGRGYDPRDGYYYRYYCGGGDRRYYDGCGRYYYGDCYPGYPCCD 120

Query: 121 PPNDHDDDNHHRNDNDPHDPPRDDNDPHRNNDRCNDNPDNNCRHCDRDCDNCRPHHNDRP 180
           P  D       R D    D  RDD  P R   RC D     CR C RDCD CRP   DR
Sbjct: 121 PYGDGYYYGGGRGDGYYGDYYRDDGYPGRGGYRCGDGYYGGCRGCYRDCDGCRPGGGDRY 180

Query: 181 DHPPRPCRRRRPPDHCDNPPDRCNDNDPRHDDPDHPDCHRDPDHDRCDDNDPPPCPPCDP 240
              PRPCRRRRP D CD PPDRC D  PR DDP  PDC R PD DRCDD     C PCDP
Sbjct: 181 YGYPRPCRRRRPYDGCDGPPDRCGDGYPRGDDPYGPDCGRYPDGDRCDDGYYYYCYPCDP 240

Query: 241 DHHDPHPPPRCPPCHDNRNDDPPNPDPPDCDRDPDNRPDHPNPNPPPHPDDDDDDPHDDR 300
               P P  RC PC   R  D P      CDR  D RPD   P  PP P DDD DP   R
Sbjct: 241 YGGYPGPYYRCYPCGYGRGYDYPGYYYYYCDRYYDGRPDGYGPGYPPGPYDDDYDPGYYR 300

Query: 301 PHPNCHRPDDDPDNDNPCNRDDPDNCNHPDCNPNPRPCHDCRPCDPHRPRHPHPDNPPHD 360
             P C RP         C R  PD C  PDC P PRPC  CR C P R R P P  P
Sbjct: 301 YGPGCGRPYYYYYGYGYCGRYYPDGCGGPDCGPGPRPCGYCRYCYPGRYRGPGPYGPYGY 360

Query: 361 PCCDDPCNRNCPCPDHPPDHPPHPNNPPRPPCCNNPCPDDDRRCDDNPHRDPDDRDRPDP 420
            CC DPC R CPCPD   D P  P  P R  CC   CPD  RRCDD   RDP DR R  P
Sbjct: 361 YCCYDPCGRGCPCPDGYYDGPYGPGGPYRYYCCGGYCPDYYRRCDDGYGRDPYDRYRYYP 420

Query: 421 NNRDHHCHDNCHPCDCDCPHRCHDRDPDCNCDRPPNRHCDPRCHCHDDRNNDNPPCNNP 479
             R   C   C PCDCDC  RC  RD DC CDRPP R C  RC C D R     PC  P
Sbjct: 421 GGRYGGCGYGCGPCDCDCYGRCGYRDYDCGCDRPPGRGCYYRCGCGDYRGGYGYPCGGP 479


Lambda     K      H
   0.320    0.149    0.595

Gapped
Lambda     K      H
   0.267   0.0410    0.140


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 1
Number of Hits to DB: 3636
Number of extensions: 655
Number of successful extensions: 7
Number of sequences better than 10.0: 1
Number of HSP’s gapped: 1
Number of HSP’s successfully gapped: 1
Length of query: 480
Length of database: 2,165,046,588
Length adjustment: 141
Effective length of query: 339
Effective length of database: 2,165,046,588
Effective search space: 733950793332
Effective search space used: 733950793332
Neighboring words threshold: 11
Window for multiple hits: 40
X1: 16 ( 7.4 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (21.7 bits)
S2: 82 (36.2 bits)
```

```
Query= mytu1

         (480 letters)

 ***** No hits found ******


Lambda     K      H
   0.320    0.149    0.595

Gapped
Lambda     K      H
   0.267   0.0410    0.140


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 1
Number of Hits to DB: 3903
Number of extensions: 719
Number of successful extensions: 3
Number of sequences better than 10.0: 0
Number of HSP’s gapped: 3
Number of HSP’s successfully gapped: 0
Length of query: 480
Length of database: 2,165,046,588
Length adjustment: 141
Effective length of query: 339
Effective length of database: 2,165,046,588
Effective search space: 733950793332
Effective search space used: 733950793332
Neighboring words threshold: 11
Window for multiple hits: 40
X1: 16 ( 7.4 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (21.7 bits)
S2: 82 (36.2 bits)
```

References

[1] Martin, W. & Muller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).

[2] Rivera, M. & Lake, J. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**, 152–155 (2004).

[3] Burnham, K. P. & Anderson, D. R. *Model selection and inference: a practical information-theoretic approach* (Springer, New York, 1998).

[4] Altschul, S. & Gish, W. Local alignment statistics. *Methods Enzymol* **266**, 460–80 (1996).

[5] Karlin, S. & Altschul, S. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A* **87**, 2264–8 (1990).

[6] Harlow, L. L., Mulaik, S. A. & Steiger, J. H. *What If There Were No Significance Tests? (Multivariate Applications)* (Lawrence Erlbaum, 1997).

[7] Morrison, D. E. & Henkel, R. E. *The significance test controversy: A reader*, vol. Methodological perspectives (Butterworths, London, 1970). Edited by Denton E. Morrison and Ramon E. Henkel.

[8] Royall, R. *Statistical evidence: A Likelihood Paradigm.*, vol. 71 of *Monographs on Statistics and Applied Probability* (Chapman and Hall, London New York, 1997). Times cited: 3.