

# Conditions under which distributions of edge length ratios on phylogenetic trees can be used to order evolutionary events

EDWARD SUSKO<sup>1,2</sup>, MIKE STEEL<sup>3</sup> AND ANDREW J. ROGER<sup>1,4</sup>

<sup>1</sup> *Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University*

<sup>2</sup> *Department of Mathematics and Statistics, Dalhousie University,*

*Halifax, Nova Scotia, Canada B3H 4R2*

<sup>3</sup> *Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand 8041*

<sup>4</sup> *Department of Biochemistry and Molecular Biology, Dalhousie University,*

*Halifax, Nova Scotia, Canada B3H 4R2*

*Corresponding Author: Edward Susko, Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5; Phone: (902) 494-8865; Fax: (902) 494-5130; E-mail: edward.susko@gmail.com*

## Abstract

Two recent high profile studies have attempted to use edge (branch) length ratios from large sets of phylogenetic trees to determine the relative ages of genes of different origins in the evolution of eukaryotic cells. This approach can be straightforwardly justified if substitution rates are constant over the tree for a given protein. However, such strict molecular clock assumptions are not expected to hold on the billion-year timescale. Here we propose an alternative set of conditions under which comparisons of edge length distributions from multiple sets of phylogenies of proteins with different origins can be validly used to discern the order of their origins. We also point out scenarios where these conditions are not expected to hold and caution is warranted.

## Main

The origin of eukaryotic cells from prokaryotic precursors - eukaryogenesis - remains one of the more mysterious major evolutionary transitions in the history of life on Earth. This transition involved a host cellular lineage related to asgard Archaea (Eme *et al.* 2017) that, at some point prior to the last eukaryotic common ancestor (LECA), took up an endosymbiotic alphaproteobacterium that became the mitochondrion, an integrated energy-producing organelle within eukaryotic cells (Dacks *et al.* 2016; Roger *et al.* 2017; Porter 2020). Genes in LECA, therefore, have multiple possible origins: either they were inherited from the host lineage, acquired from the mitochondrial symbiont by endosymbiotic gene transfer, transferred from potentially many other prokaryotic donors by lateral gene transfer (Rochette *et al.* 2014; Pittis and Gabaldón 2016a), or arose *de novo* during eukaryogenesis. Regardless of their origin, many genes were extensively duplicated during this period, as many new cellular traits including the cytoskeleton, nucleus, endomembrane system evolved

29 in the proto-eukaryote lineage. Determining the order of these events remains a major  
30 roadblock in our understanding of eukaryogenesis.

31 In 2016, Pittis and Gabaldón introduced a novel approach to approximating the relative  
32 ages of genes of different origins that were acquired during eukaryogenesis (Pittis and Ga-  
33 baldón 2016a). The approach relies on the notion that edge lengths on phylogenetic trees  
34 estimated from aligned genes or proteins, represent expected numbers of amino acid substi-  
35 tutions along the edge and are proportional to the product of rates of substitution along the  
36 edge and the time span of the edge. Under a strict molecular clock assumption the relative  
37 lengths of edges are proportional to time spans of the edges.

38 To characterize the relative timespan a gene has been resident in the proto-eukaryote  
39 lineage and LECA, Pittis and Gabaldón focused on the edge in each gene tree between the  
40 LECA node and the node representing the common ancestor of the closest prokaryotic sister  
41 group and the eukaryote lineage (Fig. 1), an edge they call the stem the length of which is  
42 denoted here as  $L_s$ . All genes from the same origin,  $O$ , are expected to have a stem edge  
43 that corresponds to the same time span ( $T_{s*}^g$  is constant for  $g \in O$ ). A serious complication  
44 arises here almost immediately. For genes in the proto-eukaryote genome that were inherited  
45 from its common ancestor with the closest sampled asgard archaeon, the time span of the  
46 stem edge in the protein tree  $T_s^g$  is the same as the timespan it has been resident during  
47 eukaryogenesis,  $T_{s*}^g$ . However, for genes that were laterally acquired during eukaryogenesis  
48 either *via* the mitochondrial symbiont or from other prokaryotic sources,  $T_s^g$  is expected  
49 to be larger than  $T_{s*}^g$  (Fig. 1). This is because the sampled taxa are unlikely to include  
50 representatives from the actual immediate prokaryotic sister group of the donor lineage of  
51 the gene(s). Reasons for this include inadequacy in sampling of living prokaryote lineages  
52 but, more likely, it is because the actual sister group, as opposed to the sampled one, went  
53 extinct. In what follows, we assume that for all comparisons  $T_s^g \approx T_{s*}^g$ , but it is important to  
54 recognize the caveats accompanying conclusions coming from stem-length methods applied  
55 to comparisons amongst acquired genes. For instance, a claim that the time of acquisition  
56 of a group  $B$  is earlier than that of an acquired group  $A$ , is more directly an inference that  
57 the closest sampled sister lineage of group  $B$  diverged earlier than that of group  $A$ .

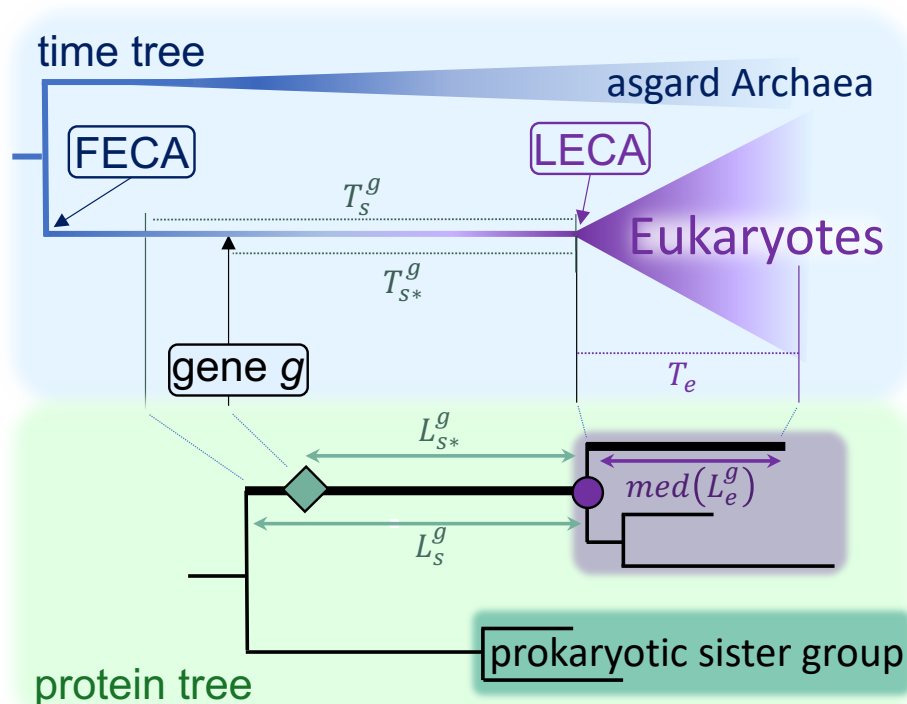


Figure 1: **The correspondence between edges on the phylogeny of a gene acquired during eukaryogenesis and edges on the geological time tree of life.** The top tree (blue box) shows the part of the geological tree of life depicting the closest asgard archaeal sister group relationship with the 'host' lineage of eukaryotes. FECA represents the *first eukaryotic common ancestor* and LECA is the *last eukaryotic common ancestor*. At some point along the eukaryogenesis edge between FECA and LECA, gene  $g$  was acquired by the proto-eukaryote genome from a prokaryotic lineage. The timespan that gene  $g$  was present during eukaryogenesis was  $T_{s*}^g$  and from LECA to the present is  $T_e$ . Below (green box) is the estimated phylogeny of protein  $g$  and its orthologs. The length of the stem edge is  $L_s^g$  and corresponds to timespan  $T_s^g$  in the time tree. The length of the segment of the latter edge post-acquisition by the proto-eukaryote lineage (green diamond to purple circle) is  $L_{s*}^g$  and corresponds to timespan  $T_{s*}^g$ . Note that  $T_s^g$  is an upper bound on the timespan of the desired stem edge post-acquisition,  $T_{s*}^g$ , because any gene transfer to the proto-eukaryote from a prokaryotic source must have occurred after speciation of the donor lineage from its closest sampled extant sister group. This discrepancy between the time of origin of a gene and timespan of its stem edge only occurs for genes acquired during eukaryogenesis (e.g. mitochondrial and other laterally acquired genes). The median length of all possible paths between the LECA node (purple circle) and eukaryote leaf node is  $med(L_e^g)$ . The normalized stem length,  $sl^g = L_s^g / med(L_e^g)$ .

58 To remove the effect of different overall rates of evolution in different genes,  $R_g$ , Pittis and  
59 Gabaldón normalize the original (raw) stem edge length, by a path length,  $L_e$ , for a path,  
60  $e$  ( $e$  for eukaryote), that corresponds to a constant time span over genes, denoted  $T_e$ , giving  
61 the time from LECA to the present (Fig. 1). This path length is the sum of consecutive edge  
62 lengths over all edges  $j$  in the path,  $L_e = \sum_{j \in e} L_j$ . Because for a given protein tree there  
63 are multiple paths from LECA to the present, and to exclude potential outliers, the path of  
64 median length was chosen as the normalization factor so that the normalized stem length is:  
65  $sl_g = L_s^g / \text{med}(L_e^g)$ . Following Pittis and Gabaldón, we refer to  $sl_g$  as a 'stem length' even  
66 though it is actually a normalized edge length.

67 Under the molecular clock model, for any path  $p$ ,  $L_p = R_g T_p$ , where  $T_p$  is the accumulated  
68 time for the path and  $R_g$  is the rate of substitution that is constant over time but may vary  
69 across genes. Thus, under the molecular clock model, the stem length satisfies that

$$70 \quad sl_g = R_g T_s / \text{med}(R_g T_e) = T_s^g / T_e$$

71 Pittis and Gabaldón (2016a) compared the distributions of estimated stem lengths,  $\hat{sl}_g$  for  
72 proteins of different origins. They found that  $\hat{sl}_g$  distributions from archaeal origin proteins  
73 ( $g \in R$ ) were, based on Mann-Whitney U tests, significantly shifted to be larger than those of  
74 bacterial ( $g \in C$ ) origin which were, in turn, significantly greater than alphaproteobacterial  
75 proteins ( $g \in M$ ) (the latter are assumed to correspond to genes that originated with the  
76 mitochondrial symbiont). Since the times are constant within groups (for instance,  $T_s^g = T_s^R$   
77 for  $g \in R$ ), they interpret this as evidence that  $T_s^R > T_s^C > T_s^M$ . An important conclusion  
78 of their study was, therefore, that the mitochondrial symbiosis took place much later in eu-  
79 karyogenesis than suggested in mitochondria early hypotheses (e.g., Lane and Martin 2010).  
80 More recently, Vosseberg and colleagues (Vosseberg *et al.* 2020) have extended this approach  
81 to address the relative timings of the mitochondrial symbiosis and gene duplication events  
82 for a variety of functional classes of protein families that expanded during eukaryogenesis.

83 Pittis and Gabaldón's approach was strongly criticized by Martin and colleagues (Martin  
84 *et al.* 2017) who argued that the results were meaningless because the method depends  
85 on the assumption that a molecular clock should hold over evolutionary time spans on  
86 the billion-year time scale. They investigated a number of the individual phylogenies from  
87 Pittis and Gabaldón study and showed that variation in edge lengths within gene trees  
88 were substantial and not consistent with a molecular clock. Pittis and Gabaldón have since  
89 countered by arguing that their approach does not assume a molecular clock and demonstrate  
90 its ability to successfully recover correct orderings of more recent evolutionary divergences in  
91 eukaryotes (Pittis and Gabaldón 2016b). However, they did not provide a detailed theoretical  
92 justification for why the method should give reliable evolutionary orderings in the absence

93 of a molecular clock

94 Here we show that assuming a molecular clock is not necessary for the method to work. In  
95 what follows, we will show that if proteins  $a$  and  $b$  in two groups  $a \in A$  and  $b \in B$  of different  
96 origins evolve independently according to the same time-dependent stochastic substitution  
97 rate process, the distributions of the normalized stem lengths of phylogenies of proteins  $B$   
98 will be systematically larger than  $A$ , i.e.  $\mathbb{P}[sl_b - sl_a > 0] > 1/2$ , if and only if  $T_s^B > T_s^A$ . We  
99 show that restrictions on the stochastic substitution rate process and the data set required  
100 for the foregoing result to hold are surprisingly few, but do include a requirement that there  
101 are no systematic differences between groups  $A$  and  $B$  at any given time point. This result  
102 is shown to hold even when estimation of edge lengths is taken into account, although a  
103 modified version of the Mann-Whitney U test will be required for testing when the variances  
104 in edge length estimates are systematically different between the groups. We also show that  
105 these methods will work in cases in which the proteins within the groups being compared  
106 have ranges of different ages. In the latter cases, however, we suggest that statistical test  
107 rejection is difficult to meaningfully interpret. Finally, we outline scenarios in which the  
108 required assumptions of the method will not hold and caution is warranted.

109 Borrowing on relaxed molecular clock theory that assumes that the rate of substitution  
110 varies stochastically over the tree (Bromham *et al.* 2019), suppose that the rate of substitu-  
111 tion at any point along a path  $p$  can be represented as  $R_g r_p^g(t)$ , where  $\{r_p^g(t)\}$  is a continuous  
112 time stochastic process ( $R_g$  is the overall fixed rate of gene  $g$  as before). Assuming a con-  
113 ventional Markov substitution model, the probability of substituting state  $i$  with state  $j$  in  
114  $(t, t+h]$  is  $q_{ij} R_g r_p(t)h + o(h)$ , for some state transition rate  $q_{ij}$ . Some constraint is required to  
115 identify parameters and we assume, without loss of generality, that  $E[R_g] = 1$  as well as the  
116 conventional constraint,  $\sum_i \sum_{j \neq i} \pi_i q_{ij} = 1$ . Since the chance of two or more substitutions is  
117 small relative to  $h$ ,  $o(h)$ , the expected number of substitutions in  $(t, t+h]$ ,  $E[N(t, t+h)]$ , is  
118  $\sum_i \sum_{j \neq i} \pi_i q_{ij} R_g r_p(t)h + o(h) = R_g r_p(t)h + o(h)$ . Taking  $u_0 = 0, u_1 = t_0/N, \dots, u_N = t_0$ , the  
119 number of substitutions in the time period  $(0, t_0]$  is the sum,  $\sum_{k=1}^N N(u_{k-1}, u_k)$  of the substi-  
120 tutions over the intervals  $(0, u_1], \dots, (u_{N-1}, u_N]$ . Since this is true for any  $N$ , the expected  
121 number of substitutions along path  $p$  and over time period  $(0, t_0]$  is

$$\begin{aligned} E[N(0, t_0)] &= \lim_N \sum_{k=1}^N E[N(u_{k-1}, u_k)] \\ &= \lim_N \left\{ \sum_{k=1}^N R_g r_p(u_{k-1}) t_0/N + N o(1/N) \right\} = R_g \int_0^{t_0} r_p(t) dt. \end{aligned}$$

122

123 Thus the stem length for protein  $g$  is

$$124 \quad sl_g = \int_{T_e}^{T_s+T_e} r_s^g(t) dt / \text{med} \left( \int_0^{T_e} r_e^g(t) dt \right)$$

125 Suppose that comparison is between group  $A$  and  $B$  and that  $T_s^B > T_s^A$ . Let

$$126 \quad Y_g = \text{med} \left( \int_0^{T_e} r_e^g(t) dt \right), \quad X_g = \int_{T_e}^{T_s^A+T_e} r_s^g(t) dt \quad \text{and} \quad V_g = \int_{T_s^A+T_e}^{T_s^B+T_e} r_s^g(t) dt. \quad (1)$$

127 Then  $sl_g = X_g/Y_g$  for  $g \in A$  and  $sl_g = (X_g + V_g)/Y_g$  for  $g \in B$ .

128 We now make the assumption that for any given path  $p$  and any two proteins  $g$  and  $g'$   
129 in  $A \cup B$ , the rate processes are probabilistically equivalent on  $[0, T_s^A]$ :

130 For any  $0 \leq t_1 < \dots < t_m \leq T_s^A$ , the joint probability distribution of  $[r_p^g(t_1), \dots, r_p^g(t_m)]$  and  
131  $[r_p^{g'}(t_1), \dots, r_p^{g'}(t_m)]$  are the same.

132 Note that this assumption allows possibly radical rate changes throughout the tree and  
133 across proteins. Moreover, the rate processes need not be stationary nor Markov processes.  
134 What is required, however, is that there be no systematic differences between the two groups.  
135 Thus, for instance, the model allows that as a consequence of a radical environmental change  
136 or change in population size at time  $t$ , for some particular lineage  $l$ , the distribution of  $r_l^g(t)$   
137 is skewed to the right of the distribution of the rate  $r_l^g(t')$  at some other time  $t'$ . But that  
138 difference in distributions is expected to apply whether  $g \in A$  or  $g \in B$ . For simplicity, we  
139 also make the assumption that the rate process is bounded such that  $r_p^g(t) \in [\beta, \gamma]$  for  
140 some  $\beta > 0$  and  $\gamma < \infty$ . This assumption is reasonable as we do not expect substitution  
141 rates to go to 0 or to increase without bound. In any case, this assumption can be loosened  
142 but some sort of assumption is required to avoid having a stem lengths that are almost 0  
143 due entirely to having extremely low average rates along the stem or extremely high average  
144 rates from LECA to present. Finally, we assume that the rate processes are independent  
145 over genes.

146 With the assumptions above, if the eukaryotic taxa sampled are the same for groups  
147  $A$  and  $B$ , then  $Y_g$  will have the same distribution whether  $g \in A$  or  $g \in B$ . Note that in  
148 Pittis and Gabaldón the same taxa were not necessarily present in any two groups of proteins  
149 being compared. We argue below that, for the approach to work, it is best if  $Y_g$  has the same  
150 distribution for  $g \in A$  or  $g \in B$ , otherwise differences in normalized stem lengths between  
151 the groups may be due to unusual rates for eukaryote taxa present in one group but not the  
152 other. Thus, it may be desirable to take means or medians over the set of eukaryote taxa  
153 present in both groups. Nevertheless, medians are not as likely to be affected by outlying  
154 rates (which was one of the original motivations for using them), so if taxon sampling is

comparable for the two groups, the distributions of  $Y_g$  are expected to be approximately the same for the two groups under the assumptions above.

Consider now  $\mathbb{P}[sl_b - sl_a > u]$  for fixed  $u$ ,  $a \in A$  and  $b \in B$ . In terms of the random variables above this can be expressed as  $\mathbb{P}[U + V_b/Y_b > u]$  where  $U = X_b/Y_b - X_a/Y_a$ . Since  $r_p^g(t) \in [\beta, \gamma]$ , the smallest  $V_b/Y_b$  could be is  $w := \beta(T_s^B - T_s^A)/(\gamma T_e) > 0$ . Thus

$$\mathbb{P}[sl_b - sl_a > u] \geq \mathbb{P}[U + w > u] = \mathbb{P}[U > u - w] \quad (2)$$

Similarly

$$\mathbb{P}[sl_a - sl_b \geq u] = \mathbb{P}[-U - V_b/Y_b \geq u] \leq \mathbb{P}[-U - w \geq u] = \mathbb{P}[-U \geq u + w] \quad (3)$$

With the assumptions above  $X_b/Y_b$  and  $X_a/Y_a$  have the same distribution. Thus  $U = X_b/Y_b - X_a/Y_a$  has a symmetric distribution around 0. Consequently,

$$\mathbb{P}[sl_a - sl_b \geq u] \leq \mathbb{P}[U \geq u + w] \leq \mathbb{P}[U > u - w] \leq \mathbb{P}[sl_b - sl_a > u] \quad (4)$$

where the first inequality is from (3) and the third from (2). The inequalities are strict unless  $U$  does not have mass in  $(u - w, u + w]$ . Since  $X_b/Y_b$  and  $X_a/Y_a$  have the same distribution then  $U$  is sure to have positive probability in  $(-w, w)$ . Thus with  $u = 0$  and since  $\mathbb{P}[sl_a - sl_b \geq 0] = \mathbb{P}[sl_b - sl_a \leq 0] = 1 - \mathbb{P}[sl_b - sl_a > 0]$ , then we have

$$0 < \mathbb{P}[sl_b - sl_a > 0] - \mathbb{P}[sl_a - sl_b \geq 0] = 2\mathbb{P}[sl_b - sl_a > 0] - 1 \quad (5)$$

or  $\mathbb{P}[sl_b - sl_a > 0] > 1/2$ .

We have shown that under the alternative hypothesis that  $T_s^B > T_s^A$  then we have that  $\mathbb{P}[sl_b - sl_a > 0] > 1/2$ . Under the null hypothesis that  $T_s^B = T_s^A$ , the distributions of  $sl_b$  and  $sl_a$  are the same. Thus if the actual normalized stem lengths were used for the two groups, the null and alternative hypotheses of interest imply the null and alternative hypotheses of the Mann-Whitney U test. However, the actual stem lengths are not known for the two groups; only estimates of these quantities from sequence data are available. This raises the question: Will the null and alternative hypotheses  $T_s^B = T_s^A$  and  $T_s^B > T_s^A$  correspond to Mann-Whitney U test null and alternative hypotheses if estimated stem length distributions are used?

Assume that the number of sites is sufficiently large for each gene that asymptotic likelihood theory gives a good approximation to the sampling distributions of the stem lengths. That theory implies that  $\hat{L}_p^g$  is approximately normal with mean  $L_p^g$ . It follows from delta-method arguments (cf. §5.3.2 of Bickel and Doksum 2007) that  $\hat{L}_s^g/\hat{L}_e^g$  is approximately normal with mean  $L_s^g/L_e^g$  for any path  $e$  from LECA to a eukaryotic taxon. Because there

186 are finitely many paths, for relatively large sequence lengths, the path,  $e^*$  say, corresponding  
 187 to the median  $\hat{L}_e^g$  should coincide with the path corresponding the median  $L_e^g$ . Thus  $\hat{sl}_g$  will  
 188 be  $\hat{L}_s^g/\hat{L}_{e^*}^g$  for the path  $e^*$  corresponding to the median  $L_e^g$ . Since  $\hat{L}_s^g/\hat{L}_{e^*}^g$  is approximately  
 189 normal with mean  $L_s^g/L_{e^*}^g$ ,  $\hat{sl}_g = sl_g + \epsilon_g$  where  $\epsilon_g$  has a normal distribution that is sym-  
 190 metric around 0. Consequently  $\mathbb{P}[\hat{sl}_b - \hat{sl}_a > 0] = \mathbb{P}[sl_b - sl_a > \epsilon_b - \epsilon_a]$ . As a difference  
 191 of independent, symmetric normals,  $\epsilon_b - \epsilon_a$  is symmetric normal. Denote the probability  
 192 density function of the latter as  $p(u)$ . Then

$$\begin{aligned}
 193 \quad \mathbb{P}[\hat{sl}_b - \hat{sl}_a > 0] &= \int_{-\infty}^{\infty} \mathbb{P}[sl_b - sl_a > u] p(u) du \\
 194 &= \int_0^{\infty} \mathbb{P}[sl_b - sl_a > u] p(u) du + \int_{-\infty}^0 \mathbb{P}[sl_b - sl_a > u] p(u) du \\
 195 &= \int_0^{\infty} \mathbb{P}[sl_b - sl_a > u] p(u) du + \int_{-\infty}^0 \mathbb{P}[sl_a - sl_b < -u] p(u) du \\
 196 &= \int_0^{\infty} \{\mathbb{P}[sl_b - sl_a > u] + \mathbb{P}[sl_a - sl_b < u]\} p(u) du \\
 197 &= \int_0^{\infty} \{\mathbb{P}[sl_b - sl_a > u] + 1 - \mathbb{P}[sl_a - sl_b \geq u]\} p(u) du \quad (6)
 \end{aligned}$$

198 By (4), under the alternative hypothesis,  $\mathbb{P}[sl_b - sl_a > u] - \mathbb{P}[sl_a - sl_b \geq u] \geq 0$  with strict  
 199 inequality in a neighbourhood of 0. Thus

$$200 \quad \mathbb{P}[\hat{sl}_b - \hat{sl}_a > 0] > \int_0^{\infty} p(u) du = 1/2 \quad (7)$$

201 We see that the alternative hypothesis of interest corresponds to  $\mathbb{P}[\hat{sl}_b - \hat{sl}_a > 0] > 1/2$   
 202 as required for the Mann-Whitney U test. However, the situation under the null is a little  
 203 more problematic. Under the null hypothesis,  $\mathbb{P}[sl_b - sl_a > u] - \mathbb{P}[sl_a - sl_b \geq u] \leq 0$ , so (6)  
 204 gives that

$$205 \quad \mathbb{P}[\hat{sl}_b - \hat{sl}_a > 0] \leq \int_0^{\infty} p(u) du = 1/2$$

206 However, the Mann-Whitney U test requires that the distributions of  $\hat{sl}_b$  and  $\hat{sl}_a$  be the  
 207 same. Although the distributions of the  $sl_a$  and  $sl_b$  are the same and the distributions of  
 208  $\epsilon_a$  and  $\epsilon_b$  are both symmetrically normal, their variances need not be comparable. These  
 209 variances reflect precision of estimation and reasons that they might differ include that  
 210 numbers of sites in alignments tend to differ substantially for one group versus the other.  
 211 The null distribution used by the Mann-Whitney U test is not correct in such settings.  
 212 Indeed, Kyusa (2000) shows that if the distributions for the two groups considered by the  
 213 Mann-Whitney test are normal but with differing variances, the type I error of the test can  
 214 be inflated. Nevertheless, Chung and Romano (2015) provide an alternative test that can



215 be used under the null hypothesis,  $\mathbb{P}[\widehat{sl}_b - \widehat{sl}_a > 0] \leq 1/2$  but  $\widehat{sl}_b$  and  $\widehat{sl}_a$  do not have the  
216 same distribution and we recommend use of this test as a safeguard. That being said, if the  
217 variability of estimation is comparable for the two groups, a Mann-Whitney U test should  
218 give reasonable results.

219 Much of the preceding discussion considers a case arising in both the analyses of Pittis and  
220 Gabaldón (2016a), and Vosseberg and colleagues (2020) where the times of origin associated  
221 with a stem length are constant for proteins within a group (for instance because they all  
222 derived from a mitochondrial symbiont or were all inherited from the archaeal host). Pittis  
223 and Gabaldón, and Vosseberg and colleagues, however, also compared groups made up of  
224 proteins of different bacterial origins and considered functional classes of proteins as groups.  
225 In these cases, proteins within a group are not expected to have a single time of origin (i.e.,  
226  $T_s^g$  will vary within a group).

227 To allow for stem times that are not constant within groups, we assume a model in which  
228  $T_s^g$  are independent across genes and independent of the rate variation processes  $r_p^g(t)$ . With  
229 the previous assumptions, the null hypothesis (that for  $a \in A$  and  $b \in B$ ,  $T_s^a$  and  $T_s^b$  have  
230 the same distribution) implies that  $sl_a$  has the same distribution as  $sl_b$ . The alternative  
231 hypothesis of greatest interest is that there is no overlap in the  $T_s^a$  and  $T_s^b$  distributions: that  
232  $\mathbb{P}[T_s^a < T_s^b] = 1$ . With this assumption, the arguments assuming fixed  $T_s^A < T_s^B$ , apply for  
233 the conditional distribution of  $sl_s^a - sl_s^b$ , given  $T_s^a$  and  $T_s^b$ . Averaging over  $T_s^a$  and  $T_s^b$  give  
234 that  $\mathbb{P}[sl_b - sl_a > 0] \leq 1/2$  under the null hypothesis and  $\mathbb{P}[sl_b - sl_a > 0] > 1/2$  under the  
235 alternative hypothesis.

236 One difficulty with analyses when the  $T_s^g$  vary within groups is that we have no control  
237 over the alternative hypothesis. The desired alternative conclusion is that  $\mathbb{P}[T_s^a < T_s^b] = 1$   
238 and we have argued above that such an alternative relationship leads to a Mann-Whitney U  
239 test null and alternative hypotheses for  $\widehat{sl}_b - \widehat{sl}_a$ . But suppose now that

$$240 \quad \mathbb{P}[T_s^b - T_s^a > z] > \mathbb{P}[T_s^a - T_s^b > z], \text{ all } z \quad (8)$$

241 This condition is related to the hypothesis of interest but might not be very meaningful. For  
242 instance, if  $\log(T_s^b) \sim N(\mu_B, \sigma^2)$  and  $\log(T_s^a) \sim N(\mu_A, \sigma^2)$  with  $\mu_B > \mu_A$ , (8) holds but if  
243  $\sigma^2$  is large then there is a substantial chance that any given  $T_s^a$  is larger than a given  $T_s^b$ .  
244 In other words, if substantial numbers of proteins in a given group  $A$  have an older origin  
245 than many of the proteins in group  $B$ , then what should we conclude from rejecting the null  
246 hypothesis that the group  $A$  distribution is shifted to be older than the group  $B$  distribution?  
247 More broadly, the rationale for grouping proteins together to test hypotheses about timings  
248 of origin is unclear if the age ranges across proteins in the groups heavily overlap and there  
249 is large variation amongst them.

250 We now show that (8) can lead to  $\mathbb{P}[\widehat{sl}_b > \widehat{sl}_a] > 1/2$ . Assume for simplicity that  
 251  $\mathbb{P}[r_p(t) = 1] = 1$ . Then the arguments leading to (4) apply with random  $W = (T_s^b - T_s^a)/T_e$   
 252 and exact equality:

$$253 \quad \mathbb{P}[sl_b - sl_a > u] = \mathbb{P}[W > u - U] = \mathbb{P}[T_s^b - T_s^a > T_e(u - U)]$$

$$254 \quad = \mathbb{P}[T_s^b - T_s^a > T_e(u + U)] \quad (9)$$

$$255 \quad \mathbb{P}[sl_a - sl_b > u] = \mathbb{P}[T_s^a - T_s^b > T_e(u + U)] \quad (10)$$

256 where the last equality in (9) and the equality in (10) follows from independence and the  
 257 symmetric distribution of  $U$ . Letting  $Z = T_e(u - U)$ , then

$$258 \quad \mathbb{P}[sl_b - sl_a > u] - \mathbb{P}[sl_a - sl_b > u] = \int_{-\infty}^{\infty} \{\mathbb{P}[T_s^b - T_s^a > z] - \mathbb{P}[T_s^a - T_s^b > z]\} p(z) dz > 0$$

259 It follows as before that  $\mathbb{P}[\widehat{sl}_b > \widehat{sl}_a] > 1/2$ . Since the Mann-Whitney U test or the Chung and  
 260 Romano robust alternative are designed to detect  $\mathbb{P}[\widehat{sl}_b > \widehat{sl}_a] > 1/2$  vs  $\mathbb{P}[\widehat{sl}_b > \widehat{sl}_a] \leq 1/2$ ,  
 261 whatever the cause, rejection could correspond to less meaningful alternatives like those  
 262 discussed above.

263 Another concern arises specifically for groups of genes that were laterally acquired during  
 264 eukaryogenesis from a prokaryotic lineage. As discussed above and shown in Fig. 1, the actual  
 265 stem-length time  $T_{s^*}$  for these genes is less than the stem-length time  $T_s$  for the observed  
 266 tree. Throughout the preceding discussion we have assumed that  $T_{s^*}^g \approx T_s^g$ . Suppose now  
 267 that the two groups  $A$  and  $B$  have different prokaryotic origins and that  $\mathbb{P}[T_{s^*}^B > T_{s^*}^A] = 1$ .  
 268 Will the Mann-Whitney U test be likely to reject in this case? Let  $K_g = T_s^g/T_{s^*}^g \geq 1$ .  
 269 Recall that  $T_s^g > T_{s^*}^g$  is expected because an immediate extant sister group to the actual  
 270 prokaryotic transfer lineage is unlikely to be among the sampled taxa because of extinction.  
 271 If the extinction processes are roughly the same for the two prokaryotic origin groups, then  
 272 it is plausible that  $K_g$  will have the same distribution for the two groups. We thus make the  
 273 additional assumption that the  $K_g$  have the same distribution for the two groups and are  
 274 independent of the rate process below.

275 We now argue that the Mann-Whitney U test is indeed likely to reject when two such  
 276 acquired groups of genes  $A$  and  $B$  have different single prokaryotic origins and origin times  
 277 are well separated:  $\mathbb{P}[T_{s^*}^B > T_{s^*}^A] = 1$ . We condition on fixed  $T_{s^*}^A$  and  $T_{s^*}^B$  in what follows.  
 278 Because the result below holds for all fixed  $T_{s^*}^B$  and  $T_{s^*}^A$ , then averaging with respect to the  
 279 distribution of  $[T_{s^*}^A, T_{s^*}^B]$  gives the result for random  $T_{s^*}^A$  and  $T_{s^*}^B$ .

280 Similarly as when  $T_{s^*} = T_s$  was assumed,  $sl_g = X_g/Y_g$  for  $g \in A$  and  $sl_g = (X_g + V_g)/Y_g$   
 281 for  $g \in B$ , where  $Y_g$  is as in (1) but now

$$282 \quad X_g = \int_{T_e}^{K_g T_{s^*}^A + T_e} r_s^g(t) dt \text{ and } V_g = \int_{K_g T_{s^*}^A + T_e}^{K_g T_{s^*}^B + T_e} r_s^g(t) dt.$$

283 For  $a \in A$  and  $b \in B$ , observe first that, because the  $\{r_b(t)\}$  and  $\{r_a(t)\}$  processes are  
284 probabilistically equivalent, and because  $K_a$  and  $K_b$  have the same distributions, then  $X_b/Y_b$   
285 and  $X_a/Y_a$  have the same distribution. Second, because  $K_g \geq 1$ , the smallest  $V_b/Y_b$  can be  
286 is then  $w = \beta(T_{s^*}^B - T_{s^*}^A)/(\gamma T_e)$ . These two properties were what was used in the arguments  
287 leading to (2)–(5) and so those results hold in this setting. Here (5) gives the conclusion  
288 required for the Mann-Whitney U test, that  $\mathbb{P}[sl_b - sl_a > 0] > 1/2$  and (2) is the key  
289 inequality that can be used, exactly as before, to show (7), that, even with estimation,  
290  $\mathbb{P}[\widehat{sl}_b - \widehat{sl}_a > 0] > 1/2$ .

291 Note that the above assumption that  $K_g$  has the same distribution across groups is  
292 violated for some comparisons. For instance, comparisons of distributions from groups of  
293 acquired genes (where  $T_{s^*}^g \leq T_s^g$  and, possibly, broad ranges of  $T_s^g$  values within the group)  
294 with genes inherited from the asgard archaeon-eukaryote common ancestor or genes that  
295 originate by duplication (for which  $T_{s^*}^g = T_s^g$  in both cases). In the simplest case of fixed  $T_s^g$   
296 values for the genes in an acquired group, the inferred age of that group will be biased to  
297 be younger than its true age in comparison with genes inherited from the asgard-eukaryote  
298 common ancestor or groups of duplicated genes.

299 We have shown that validity of the edge length ratio methods introduced by Pittis and  
300 Gabaldón (2016a), and extended by Vosseberg and colleagues (2020) do not require a molec-  
301 ular clock. They can be justified in much more general settings where substitution rates in a  
302 protein stochastically vary over the tree. Indeed, the only restrictions are that the stochastic  
303 substitution rate process is bounded away from 0 and infinity and that genes in groups of  
304 different origins (or functional classes) in a genome are all independently evolving according  
305 to this same process (i.e. the rate process for different genes are probabilistically equiva-  
306 lent). In terms of biological realism, it is this latter assumption that may not always hold.  
307 For example, it is well known that proteins may periodically experience episodes of rapid  
308 adaptive evolution due to acquisition of novel functions and/or loss of ancestral functions  
309 (Studer, Dessailly and Orengo, 2013). If this functional divergence differentially affected  
310 proteins within groups of different origins or functional classes, then stem length distribu-  
311 tions of one group of proteins versus another will likely reflect this episodic shift in rates in  
312 one group, and cannot be used to test if they originated at different times. When applying  
313 these methods, it is therefore important to investigate evidence for systematic differences in  
314 the evolutionary dynamics of one group of proteins versus another.

315 It is also important to select edge lengths from phylogenies for different gene groups to be  
316 compared in the same manner to ensure that no biases are introduced. For example, when  
317 extending the approach to genes duplicated during eukaryogenesis, Vosseberg and colleagues  
318 (2020) were faced with deciding how to deal with the multiple possible edges or paths to

319 LECA nodes created by duplications (Fig. 2). Their approach was to always select the  
 320 minimum of all possible edge or path lengths for calculations of stem lengths or duplication  
 321 lengths (all duplication lengths were calculated as  $dl_g = L_d^g / \text{med}(L_e^g)$ ). For two groups  $D$   
 322 and  $S$  of duplicated or acquired genes (for which stem lengths can include duplicated edges),  
 323 the alternative hypothesis of greatest interest is  $H_{DS}$ :  $\mathbb{P}[T_{di} > T_{sj}] = 1$  for  $d \in D$  and  $s \in S$ ,  
 324 regardless of  $i$  and  $j \in \{1, 2\}$ .

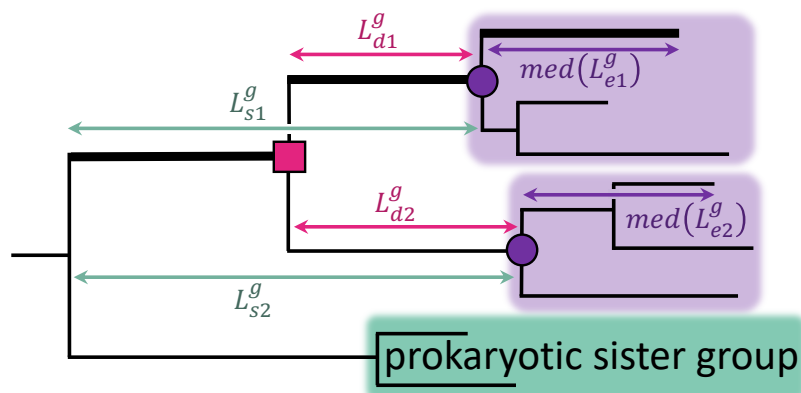


Figure 2: **Example of the multiple possible stem paths and duplication edges in the phylogeny of a protein that has been duplicated during eukaryogenesis.** This gene was acquired by the proto-eukaryote genome from a prokaryotic lineage (green box) and then duplicated (magenta box) prior to LECA (the LECA nodes of each duplicate are shown as purple circles). As a result there are two possible stem paths with lengths  $L_{s1}^g$  and  $L_{s2}^g$  (green arrows), two possible duplication edges with lengths  $L_{d1}^g$  and  $L_{d2}^g$  and two possible median paths within the eukaryote subtrees (purple boxes) with lengths  $\text{med}(L_{e1}^g)$  and  $\text{med}(L_{e2}^g)$ . To calculate stem lengths and duplication lengths, Vosseberg and colleagues (2020) used the edges/paths with minimum values (minimums shown as thicker edges).

325 Assume that evolution is independent post duplication. Then, if all duplicated lengths  
 326 are included in a comparison involving minima, the arguments above apply and the Mann-  
 327 Whitney U test would be likely to reject when  $H_{DS}$  holds. One possible motivation for  
 328 using minima is that it could potentially alleviate bias due to the functional divergence  
 329 phenomenon alluded to above because functional divergence is more likely to have occurred  
 330 in the duplicate with the longer  $L_{di}^g$  over  $i \in \{1, 2\}$ . Unfortunately, taking minimums leads to  
 331 some loss of information and could lead to a bias which we illustrate assuming no functional  
 332 divergence. Similar arguments as those above related to the median path lengths, imply

333 that the Mann-Whitney U test would have substantial probability of rejection whenever  
334  $\mathbb{P}[\min(T_{d1}, T_{d2}) > \min(T_{s1}, T_{s2})] = 1$ . If  $H_{DS}$  holds and evolution is independent, then this  
335 hypothesis will hold, so the approach should work. In less ideal alternative hypothesis  
336 scenarios where there is overlap in the distributions of  $T_{di}$  and  $T_{sj}$ , biases can occur. As an  
337 illustrative example, suppose that  $S$  is a stem-length group of acquired genes, that  $T_e = 1.5$   
338 billion years ago, that  $T_{sj} = 0.4$  billion years and that, independently,  $T_{di} \sim U(0, 1)$ . Then

339 
$$\mathbb{P}[T_{di} > T_{sj}] = \mathbb{P}[T_{di} > 0.4] = 1 - 0.4 = 0.6,$$

340 consistent with longer duplication lengths. But

341 
$$\mathbb{P}[\min(T_{d1}, T_{d2}) > \min(T_{s1}, T_{s2})] = \mathbb{P}[T_{d1} > 0.4, T_{d2} > 0.4] = 0.6^2 = 0.36$$

342 Consequently, based on the Mann-Whitney U test, one would conclude that the age of the  
343 duplicated group of genes  $D$  tended to be less than the acquired genes  $S$  when, in fact, 60%  
344 of the duplication group genes,  $D$ , duplicated prior to their acquisition.

345 Another complication arises in the comparison of duplication groups and stem-length  
346 groups. For duplication groups, duplication lengths are all minima. The stem-length groups,  
347 however, are usually a mix of stem lengths, some of which are chosen to be minima (Fig. 2)  
348 and some of which did not require a minimum (Fig. 1) (Vosseberg *et al.* 2020). In this case  
349 biases might arise even under the strong hypothesis,  $\mathbb{P}[\min(T_{d1}, T_{d2}) > \min(T_{s1}, T_{s2})] = 1$  for  
350  $d \in D$  and  $s \in S$ , making it difficult to reject when  $H_{DS}$  holds. Such problems can be averted  
351 by including all duplication and stem lengths rather than minima. There are potentially  
352 other ways of addressing functional divergence that may be less likely to introduce bias  
353 (e.g., identifying functionally divergent sites using methods reviewed in Studer, Dessailly  
354 and Orengo (2013) and removing them prior to analysis).

355 In summary, although there are a number of caveats, if the assumptions of the methods  
356 we have elaborated above are met by the data, these edge length ratio methods have the  
357 potential to provide important new insights into the roles of gene duplication and gene  
358 invention in different cellular systems and clarify the relative contributions of host, symbiont  
359 and lateral transfers to a lineage of interest.

## 360 Acknowledgements

361 This research was supported a grant (Award ID: 735923LPI) awarded to A.J.R. and E.S.  
362 as part of the Moore-Simons Project on the Origin of the Eukaryotic Cell. E.S. and A.J.R.  
363 also acknowledge partial support for this work from Discovery grants awarded to them by  
364 the Natural Sciences and Engineering Research Council of Canada. A.J.R. acknowledges the  
365 Allan Wilson Centre in New Zealand for sabbatical support received in 2006 when some of  
366 these ideas were first discussed.

367 **References**

- 368 Bickel, P.J. and Doksum, K.A. (2007). Mathematical Statistics: Basic ideas and selected  
369 topics. Pearson, New Jersey.
- 370 Bromham, L., Duchne, S., Hua, X., Ritchie, A.M., Duchne, D.A., and Ho, S.Y.W. (2018).  
371 Bayesian molecular dating: opening up the black box. *Biol. Rev. Camb. Philos. Soc.*  
372 *93*, 11651191.
- 373 Chung E.Y. and Romano, J.P. (2015). Asymptotically valid and exact permutation tests  
374 based on two-sample U-statistics. *J. Statist. Plann. and Inf.* 168:97–105.
- 375 Dacks, J.B., Field, M.C., Buick, R., Eme, L., Gribaldo, S., Roger, A.J., Brochier-Armanet,  
376 C., and Devos, D.P. (2016). The changing view of eukaryogenesis - fossils, cells, lineages  
377 and how they all come together. *J. Cell Sci.* 129, 36953703.
- 378 Eme, L., Spang, A., Lombard, J., Stairs, C.W., and Ettema, T.J.G. (2017). Archaea and  
379 the origin of eukaryotes. *Nat. Rev. Microbiol.* 15, 711723.
- 380 Kasuya, E. (2001). Mann-Whitney U test when variances are unequal. *An. Behav.*  
381 *61*:1247–1249.
- 382 Lane N. and Martin W.F. (2010) The energetics of genome complexity *Nature* 467:929-934.
- 383 Martin, W.F., Roettger, M., Ku, C., Garg, S.G., Nelson-Sathi, S., and Landan, G. (2017).  
384 Late Mitochondrial Origin Is an Artifact. *Genome Biol. Evol.* 9:373-379.
- 385 Pittis, A.A. and Gabaldón, T. (2016a). Late acquisition of mitochondria by a host with  
386 chimaeric prokaryotic ancestry. *Nature.* 531:101-104.
- 387 Pittis, A.A. and Gabaldón, T. (2016b) On phylogenetic branch lengths distribution and  
388 the late acquisition of mitochondria. *bioRxiv doi: <https://doi.org/10.1101/064873>*
- 389 Porter, S.M. (2020). Insights into eukaryogenesis from the fossil record. *Interface Focus* 10,  
390 20190105.
- 391 Rochette, N.C., Brochier-Armanet, C., and Gouy, M. (2014). Phylogenomic test of the  
392 hypotheses for the evolutionary origin of eukaryotes. *Mol. Biol. Evol.* 31:832-845.
- 393 Roger, A.J., Muñoz-Gómez, S. and Kamikawa, R. (2017). The origin and diversification of  
394 mitochondria. *Curr. Biol.* 27:R1177-R1192.

- 395 Vosseberg, J., van Hooff, J.J.E, Marcet-Houben, M., van Vlimmeren, A., van Wijk, L.M.,  
396 Gabaldón, T. and Snel, B. (2020). Timing the origin of eukaryotic cellular complexity  
397 with ancient duplications. *Nat. Ecol. Evol.* 5:92-100.
- 398 Studer, R.A., Dessailly, B.H., and Orengo, C.A. (2013). Residue mutations and their  
399 impact on protein structure and function: detecting beneficial and pathogenic changes.  
400 *Biochem. J.* 449:581-594.