Journal of Phylogenetics & Evolutionary Biology

DeSalle, J Phylogen Evolution Biol 2015, 3:4
DOI: 10.4172/2329-9002.1000e116

Editorial Open Access

Can Single Protein and Protein Family Phylogenies be Resolved Better?

Rob DeSalle*

Sackler Institute for Comparative Genomic, American Museum of Natural History, New York, NY 10024, USA

The Problem

Single gene or protein sequences have been used in protein biology and molecular biology studies to make inferences about the dynamics of evolution of their structure and function. In addition, single gene or protein phylogenies are required input in statistical tests for natural selection or in the determination of sites in proteins that are coevolving. Gene phylogenies using members of a multigene family (the globin genes are a classic example) are important too, because they can decipher the relationships of the members of the overall gene family and can address potential neofunctionalization phenomena [1-3]. While the issue of how single gene and single proteins contribute to the overall phylogeny of a group of organisms is an important and copiously addressed topic [4-7], the focus here is on the use of gene and protein sequences in single gene studies.

The problem is that in many cases, the support at nodes in the phylogenies generated from single genes or proteins is extremely weak if not absent [8-11]. As an example ten randomly chosen proteins from the GenBank data base were chosen. The taxon representation focused on metazoans with complete genomes (Table-1). Bootstrap trees for Maximum likelihood (ML; RaxML) [12] and Maximum parsimony (MP; PAUP) [13,14] as well as Bayesian posteriors (BP) [15] were generated (see figure legend for models and other criteria imposed in these analyses) for all ten proteins. For this example amino acid sequences were used, however, DNA sequences would pose an even more extreme problem due to the smaller number of character states in DNA and also to the dynamics of change of the two kinds of data [16,17].

Figure 1A shows the results of this survey and indicates that support at a large number of nodes in phylogenies from single proteins of average length is lacking no matter what optimality criterion is used. Two levels of support were examined -whether a node had greater than 50% support, and whether a node had 85% supports for the bootstrap approach. For Bayesian analysis 0.95 and 0.85 posterior cutoffs were tallied. The figure shows that at best only 50% of the nodes in bootstrap trees and 60% of the nodes in the Bayesian trees are resolved at the higher cut-off. On average, for the ten proteins 22% (MP), 29% (ML) and 35% (MB) of the nodes at this upper cutoff are resolved. The ten proteins fare slightly better when the bootstrap/posterior cutoff is set at the lower cutoff with 39% (MP), 50% (ML) and 53% (MB) of the nodes in these trees being resolved.

A similar analysis of multigene families was performed on three multigene families–tetraspanins (TSPAN; 3107 terminals, 14413 characters, 4011 phylogenetically informative characters), spanning nexins (SNX; 2941 terminals, 3759 characters, 2169 phylogenetically informative characters) and the albumin (ALB; 310 terminals, 973 characters, 611 phylogenetically informative characters) gene family genes. Maximum parsimony was used to generate trees for this example because of the large number of terminals in this data set. All three of these gene families yield MP trees with strong support for the monophyly of individual members of the gene families (many subfamilies are recognized as monophyletic in all three gene families), but low levels of support for relationships within the gene families and between the various members of the overall larger gene family is the norm. Figure 1B shows that the results for multigene families using MP are similar (i.e., no more than half of the nodes in these trees) to the results for single genes. One gene family (ALB) fares better than the other two but this result is most assuredly a factor of the number of family members (terminals) in the three studies and hence nodes in the trees (see above).

The degree of resolution for these three gene families was also assessed across different depths in the trees (from the base of the tree to the tips). Figure 2A shows that for all three of the gene families deep nodes are very poorly resolved at 85% bootstrap values (below 40% of the nodes are resolved at this cut-off for the two larger families). As with the single gene analysis (Figure 1A) the resolution of the trees fare a bit better for the 50% cutoff at the base of the tree. In addition, as one might expect, the support level rises as the nodes appear further and further toward the tips of the tree, with the 50% bootstrap cutoffs rising to about 80% of the total tip nodes. However even at the tips of the tree for the 85% bootstrap cutoff, only 50% of the nodes are resolved.

Robustness

The lack of resolution of single gene and gene family phylogenies could be due to many factors. First and foremost, the relative paucity of characters that are available for reconstructing the evolutionary history of single proteins and genes relative to the number of terminals is a major factor. The examples in this communication range in size from 131 phylogenetically informative characters to 850 phylogenetically informative characters, but each of the analyses presented here have at least 85 terminals for the single gene analysis and up to 3100 terminals for the TSPAN analysis. A second reason for the lack of resolution might be incomplete search strategies that would produce suboptimal trees with many solutions. The multiple solutions produced by incomplete tree searches will cause the bootstraps and posteriors to be lowered and hence result in lower robustness. A third factor is sequence alignment. Slightly changing the alignment parameters (i.e., the gap insertion or extension cost) will change the phylogenetic outcome, often times drastically for single genes [18]. The alignment parameters will therefore contribute to phylogenetic robustness problems in single gene and gene family phylogenies.

	Taxa	max nodes	chars	phyl inf
CALR	119	116	1661	364
BID	95	92	245	225
GAPDH	77	74	482	263

Table 1: Ten randomly chosen nuclear genes from metazoan taxa with selected out-groups. Matrices are available upon request from the author. The taxa column gives the total number of taxa in the data set. The max nodes column indicates the maximum number of nodes in a fully resolved tree for the indicated data set. The chars column gives the size of the protein in amino acids. The phy inf column lists the number of phylogenetically informative sites in each data set.

***Corresponding author:** Rob DeSalle, Sackler Institute for Comparative Genomic, American Museum of Natural History, New York, NY 10024, USA, E-mail: desalle@amnh.org

Received September 04, 2015; **Accepted** September 07, 2015; **Published** September 14, 2015

Citation: DeSalle R (2015) Can Single Protein and Protein Family Phylogenies be Resolved Better? J Phylogen Evolution Biol 3: e116. doi:10.4172/2329-9002.1000e116

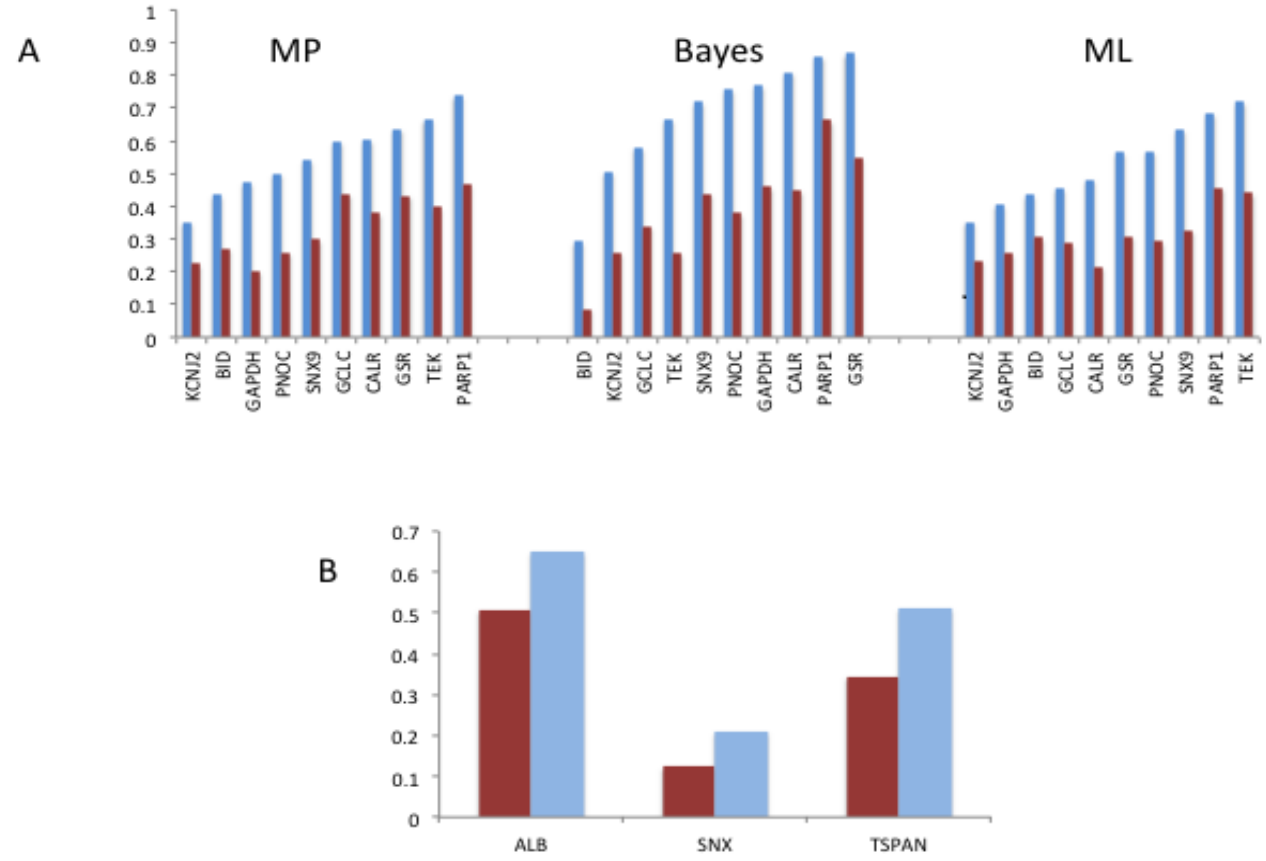

Figure 1: A) Bar charts showing the percentage of total possible nodes that show a bootstrap value >85% for MP and ML or a posterior >0.95 for Bayesian analysis (red bars) and bootstrap value >50% for MP and ML or a posterior >85% for Bayesian analysis (blue bars). MP trees were generated in PAUP ML trees were generated with RaxML BlackBox and Bayes trees were generated with Mr Bayes (Ronquist and Huelsenbeck, 2003). MP trees were generated without weighting, ML and Bayesian analysis used the GTR+G+I model. The Bayesian analysis used MCMC with four chains and two replicates with ngen=100000 and burnin=20000. Taxon number in all of the data sets were small enough (see Table 1) to do aggressive searches for all data sets for MP and ML (using TBR and 200 random taxon additions). The data sets are described in Table 1.

B) Bar chart showing the percentage of nodes in the three multigene families examined here that showed bootstrap percentages greater than 85% (blue bars) or greater than 50% (red bars). MP analysis was accomplished using TNT (Goloboff et al., 2008) with 100 re-samplings and the ratchet. Only MP was used because of the large number of terminals. Matrices are available from the author on request.

If all we were interested in using gene sequences in phylogenetics to elucidate the relationships of organisms at the terminals of trees, then our problem is obviously solved simply by adding more data from other genes. However with both single gene phylogenies used to make inferences about the evolution of specific proteins and gene or protein families, we do not have the luxury of adding more primary sequence information to increase character numbers. So what can be done when we have a single gene analysis and we have poorly resolved and non-robust inferences?

There are three major solutions to the problem, one relies on a more philosophical approach based on first principles [19,20], the second attempts to deal with robustness and the third ignores the phylogenetic inferences made from single genes or proteins and falls back on a references or "taxonomic" topology for further analyses.

The "first principles" solution is to simply pick and defend an optimality criterion and use the optimal solution for that optimality criterion. This choice would mean that the maximum parsimony tree for a data set or the maximum likelihood tree for a data set or the Bayesian consensus tree for a data set would be accepted as the solution to the problem. While this solution is defensible from first principles it is often times disregarded as too simplistic.

The second route is to somehow accommodate the problem of robustness. As stated before this is an easy route when obtaining a phylogeny of organisms is the goal by simply adding more sequence data. However when only a single gene or protein is the target of analysis, no new sequence characters can be added from other genes. There are other sources of information in single genes though. One mentioned earlier is the fact that different alignment parameters yield different alignments. Gatesy [21] argued that concatenating (eliding in their terminology) matrices constructed using different gap costs essentially created a matrix of characters that was effectively weighted by how stable alignment positions were. This procedure up-weights stable alignment positions and down weights those that are unstable. Agosti suggested that concatenating an amino acid matrix with its corresponding DNA matrix is another way to internally weight sequences. Both of these concatenation methods while they do not directly duplicate characters do however violate the assumption of non-independence of characters that most phylogeneticists impose in their analyses. Other weighting approaches using transition matrices for MP can be applied too. As well, different models in the ML framework might have some impact on robustness.

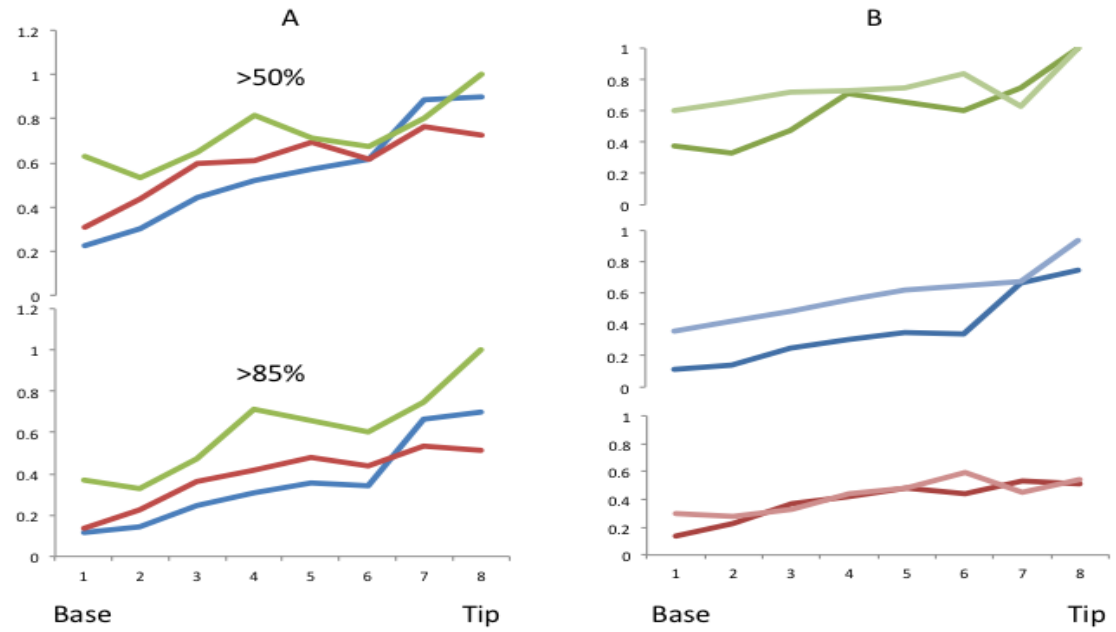

Figure 2: **A)** Graph showing the increase in percent of resolved nodes as a function of distance from the base of the tree. Red lines indicate SNX, blue lines indicate TSPAN and green lines indicate ALB. The nodes in the trees for these three large gene families were categorized into eight bins based on their depth in the tree. Each of the 8 bins was then analyzed for the percentage of nodes resolved at 50% bootstrap (top) and 85% bootstrap (bottom).

B) Graphs showing the increase in percentage of nodes resolved at 85% bootstrap (for MP) resulting from concatenating different alignments obtained from increasing gap opening costs (using MAFFT, go=1,2,3,4,5). Red lines indicate SNX, blue lines indicate TSPAN and green lines indicate ALB. In all cases the lighter line represents the concatenated analysis. MP trees were generated using TNT with 100 re-samplings and the ratchet. The nodes in the trees for these three large gene families were categorized into eight bins based on their depth in the tree. Each of the 8 bins was then analyzed for the percentage of nodes resolved at 50% bootstrap (top) and 85% bootstrap (bottom). Matrices are available from the author on request.

However, as Figure 1A suggests, imposing a model in ML fares about as equally well with robustness as the unweighted parsimony approach. The expectation would be that even the most parameter rich model that could be imposed in ML would not fare much better than what is shown in Figure 1A. A final way of adding information to a matrix would be to code some of the structural aspects of a gene (intron/exon junctions; secondary structure aspects; repeating motifs; function of protein; presence or absence of domains; etc.) as characters [22-24]. The number of characters that can be added using this approach is usually quite small compared to the sequence information, but often times these characters can add unexpected structure and robustness to single gene and protein trees.

The last approach mentioned above is to simply constrain the phylogeny of the organismal terminals based on a reference topology of known relationships. This is the "taxonomic" approach, and would be a valid approach when using gene and protein information to survey for natural selection (via dN/dS skew) or for coevolution of proteins or sites in proteins. In many of these approaches a tree is absolutely essential to the analysis and often times a Neighbor Joining tree is used. One caveat of this approach is that if there has been some lineage sorting for the gene under consideration and its sequences indicate a topology not broadly congruent with the "taxonomic" one, then dN/dS and co-evolutionary inferences might be compromised. It should be pointed out that this final approach is only appropriate if the pattern of terminals is not a primary goal of analysis. This stricture would leave out single gene and protein studies where gene and protein families are examined for the relationships of internal members to each other.

Concatenation: An Example

The bottom line is that for most single gene and protein phylogenetic analysis, we simply cannot expect a high degree of robustness with typical ways of analyzing sequence data. Concatenation of matrices with different gap costs and concatenation of DNA and amino acid sequences from the same protein can improve the robustness of trees. Adding structural or functional characters will often times surprisingly improve the robustness of nodes in trees too. As an example of using the concatenation approach, the ten nuclear gene data sets in Table 1 were used to generate concatenated matrices using five different gap opening costs with the alignment program MAFFT [25] In addition the same was accomplished for the three multigene families (ALB, SNX and TSPAN). Figure 3 shows the results of concatenation for the 85% cut-off for bootstraps for ML and MP and the 0.95 posterior probability cutoff for Bayesain analysis and clearly shows the increase in number of resolved nodes for all three methods for all ten proteins. Figure 2B shows the trend from base to tips for the multigene families (ALB, SNX and TSPAN). The pattern for this analysis is slightly more complex, but in general there is an increase in robustness using the concatenation method, but the strongest impact is seen at the base of all three trees. As one moves out toward the tips, the effect of concatenation is less conspicuous. Concatenation as used here might be a useful tool for exploring the robustness of nodes in single protein phylogenies.

Solutions/Recommendations

Any of the solutions to this problem of weak nodal support in single gene or protein trees should be articulated clearly in any publication. Perhaps the most logical solution would be to settle on an optimality criterion and report that tree. Bootstraps or Bayesian posteriors could then be placed on the MP or ML tree. Since the tree topology using an optimality criterion can be different from the overall bootstrap tree or Bayesain consensus tree, simply reporting the bootstrap or Bayesian

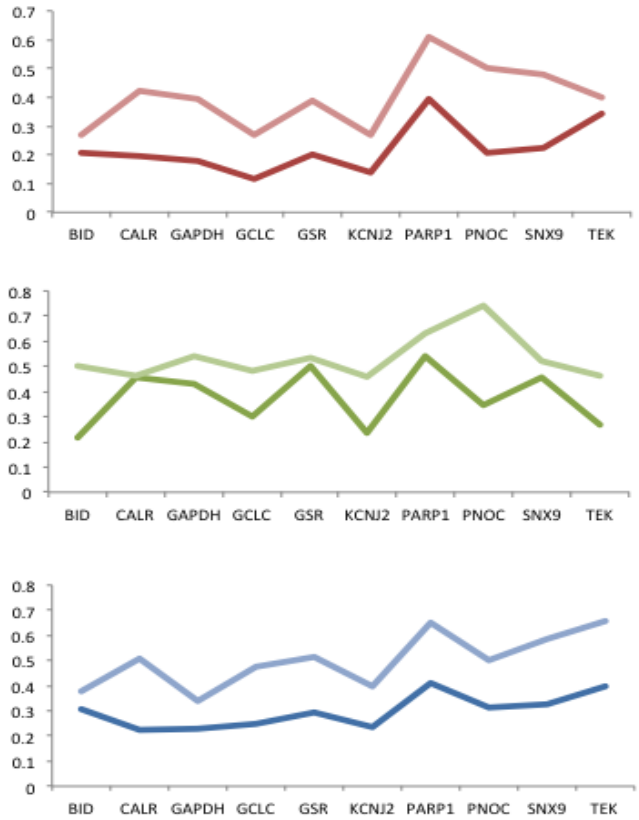

Figure 3: Graphs showing the increase in percentage of nodes resolved at 85% bootstrap (for ML and MP) and >0.95 posterior probability (for Bayesian analysis) resulting from concatenating different alignments obtained from increasing gap opening costs (using MAFFT, go=1,2,3,4,5) for MP (red), Bayesian analysis (green) and ML (blue). In all three graphs the lighter line represents the concatenated matrix. MP trees were generated in PAUP (Swofford, 2003), ML trees were generated with RaxML BlackBox and Bayes trees were generated with Mr Bayes (Ronquist and Huelsenbeck, 2003). MP trees were generated without weighting, ML and Bayesian analysis used the GTR+G+I model. The Bayesian analysis used MCMC with four chains and two replicates with ngen=100000 and burnin=20000. Taxon number in all of the data sets were small enough (see Table 1) to do aggressive searches for all data sets for MP and ML (using TBR and 200 random taxon additions). The data sets are described in Table 1. Matrices are available from the author on request.

consensus trees would obscure the MP or ML tree, so using the ML and MP trees is critical. The utility of accepted "taxonomic" trees in dN/dS analysis and in protein coevolution should also be considered as an alternative to weakly supported input trees in these kinds of analyses.

References

1. Lespinet O, Wolf YI, Koonin EV, Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome research 12: 1048-1059.
2. Rensing SA (2014) Gene duplication as a driver of plant morphogenetic evolution. Curr Opin Plant Biol 17: 43-48.
3. Escriva H, Bertrand S, Germain P, Robinson-Rechavi M, Umbhauer M, et al. (2006) Neofunctionalization in vertebrates: The example of retinoic acid receptors. PLoS Genet 2: e102.
4. Edwards SV, Liu L, Pearl DK (2007) High-resolution species trees without concatenation. Proc Natl Acad Sci U S A 104: 5936-5941.
5. Liu L, Yu L, Pearl DK, Edwards SV (2009) Estimating species phylogenies using coalescence times among sequences. Syst Biol 58: 468-477.
6. de Queiroz A, Gatesy J (2007) The super matrix approach to systematics. Trends Ecol Evol 22: 34-41.
7. Gatesy J, Springer MS (2013) Concatenation versus coalescence versus "concatalescence". Proc Natl Acad Sci U S A 110: E1179.
8. Nei M, Kumar S, Takahashi K (1998) The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proceedings of the National Academy of Sciences 95: 12390-12397.
9. Gatesy J, Springer MS (2014) Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Molecular phylogenetics and evolution 80: 231-266.
10. Sun M, Soltis DE, Soltis PS, Zhu X, Burleigh JG, et al. (2015) Deep phylogenetic incongruence in the angiosperm clade Rosidae. Mol Phylogenet Evol 83: 156-166.
11. Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, et al. (2011) Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol 9: e1000602.
12. Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol 57: 758-771.
13. Swofford DL (2003) {PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.}.
14. Goloboff PA, Farris JS, Nixon KC (2008) TNT, a free program for phylogenetic analysis. Cladistics 24: 774-786.
15. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.
16. Ward W (1999) Fixed character states and the optimization of molecular sequence data. Cladistics 15: 379-385.
17. Källersjö M, Farris JS, Kluge AG, Bult C (1992) Skewness and permutation. Cladistics 8: 275-287.
18. Gatesy J, DeSalle R, Wheeler W (1993) Alignment-ambiguous nucleotide sites and the exclusion of systematic data. Mol Phylogenet Evol 2: 152-157.
19. Kluge AG (1989) A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Systematic Zoology 38: 7-25.

20. Giribet G, DeSalle R, Wheeler WC (2002) 'Pluralism' and the aims of phylogenetic research. EXS: 141-146.

21. Wheeler WC, Gatsey J, DeSalle R (1995) Elision: A method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. Molecular phylogenetics and evolution 4: 1-9.

22. Rokas A, Kathirithamby J, Holland PW (1999) Intron insertion as a phylogenetic character: the engrailed homeobox of Strepsiptera does not indicate affinity with Diptera. Insect Mol Biol 8: 527-530.

23. Rokas A, Holland PW (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15: 454-459.

24. Ender A, Schierwater B (2003) Placozoa are not derived cnidarians: Evidence from molecular morphology. Mol Biol Evol 20: 130-134.

25. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772-780.