Journal of Biological Systems, Vol. 18, No. 2 (2010) 261–275

DOI: 10.1142/S0218339010003408

World Scientific
www.worldscientific.com

WHOLE GENOME COMPARISONS REVEALS A POSSIBLE CHIMERIC ORIGIN FOR A MAJOR METAZOAN ASSEMBLAGE

MICHAEL SYVANEN*,‡ and JONATHAN DUCORE†

*Department of Microbiology
†Department of Pediatrics
University of California at Davis School of Medicine
Davis, CA 95617, USA
‡syvanen@ucdavis.edu

Received 11 September 2009
Accepted 27 April 2010

The availability of whole genome sequences from multiple metazoan phyla is making it possible to determine their phylogeny. We have found that a sea urchin and human define a clade that excludes a tunicate, contradicting both classical and recent molecular studies that place the tunicate and vertebrate in the Chordate phylum. Intriguingly, by means of a novel four taxa analysis, we have partitioned the 2000 proteins responsible for this assignment into two groups. One group, containing about 40% of the proteins, supports the classical assemblage of the tunicate with vertebrates, while the remaining group places the tunicate outside of the chordate assemblage. The existence of these two phylogenetic groups is robustly maintained in five, six and nine taxa analyses. These results suggest that major horizontal gene transfer events occurred during the emergence of one of the metazoan phyla. The simplest explanation is that the modern tunicate (as represented by *Ciona intestinalis*) began as a hybrid between a primitive vertebrate and some other organism, perhaps from an extinct and unidentified protostome phylum, at a time close to but after the diversification of the chordates and echinoderms and before the lineages leading to *Drosophila melanogaster* and *Caenorhabditis elegans* diverged.

Keywords: Horizontal Gene Transfer; Four-Taxa Analysis; Tunicate; Chordate; Protostome; Deuterostome.

1. Introduction

Biologists have long been puzzled by the suddenness with which the metazoan phyla appeared in the fossil record during the early Cambrian period. Besides the sudden emergence of the metazoa, the widespread occurrence of parallel evolution among them was also considered puzzling. When one of us first started thinking about the evolutionary implications of horizontal gene transfer, it seemed quite natural to suggest that gene transfer between phyla could explain the phenomena of both the speed and observed parallelisms (homoplasy) of the Cambrian radiation.[1]

Initially it appeared that resolving the relationships among the metazoan phyla might not be possible since most emerged in the fossil record over a seemingly

brief time period. If so, this would make it difficult, if not impossible, to assess the likelihood that horizontal gene transfer played a role during the metazoan radiation. However, recent genomics studies[2,3] indicate that there are sufficient numbers of phylogenetic informative characters to reconstruct metazoan histories, though the signal is relatively weak and requires the information from many hundreds of genes. The current work began after the genomes representing multiple metazoan phyla became available, thereby making it possible to discern alternative evolutionary histories for different sets of genes. We encountered an unexpected result with regard to the tunicate, *Ciona intestinalis*. This organism, known as the sea squirt, has a completely unique metazoan body plan in its adult phase, but has a larval phase that quite clearly identifies it as a chordate. Given that embryonic and larval characters have, at least for the Chordates, been used in classification, the tunicates have been classified with them. The current results pose new questions about the chordate-tunicate affinity.

We will argue that conflicting phylogenies concerning the placement of the tunicates are a result of extensive horizontal gene transfer (HGT).

Most reports of HGT are based on finding a phylogeny of a single gene that conflicts with the known species phylogeny. There has been some work on finding techniques that evaluate reticulations from multiple genetic loci for which there is no known *a priori* taxa phylogeny.[4,5] However, these techniques have been tested primarily on reticulate events that have occurred in the recent past, such as plant introgressions, recombinant viruses, hybridizations and recombination patterns within interbreeding populations. There has been little application of these techniques in unraveling reticulation patterns found deep in phylogeny such as the emergence of metazoan phyla. This paper presents a series of steps for examining such a problem.

2. Results

Our initial observation showing that the tunicates may not be closely related to other members of the phylum Chordata is shown in Fig. 1. This tree is based on protein distances averaged from about 2000 genes. Parsimony trees supported the same assemblages. We can see that the sea urchin (*Strongylocentrotus purpuratus*) and human belong to a clade that excludes the tunicate (*Ciona intestinalis*), fruit fly (*Drosophila melanogaster*) and the round worm (*Caenorhabditis elegans)*. One point that is clear from this figure is that the internal branches that separate the vertebrate, sea urchin and tunicate are short compared to the terminal branches leading to the extant taxa. If the tree shown in Fig. 1 were based on one or a few genes, such an unexpected result could be attributed to the multiple replacements occurring in different terminal lineages, artifactually creating phylogenetic information. But with the very large number of genes and consequent large number of shared or phylogenetically informative characters, this unexpected result can be meaningfully analyzed.

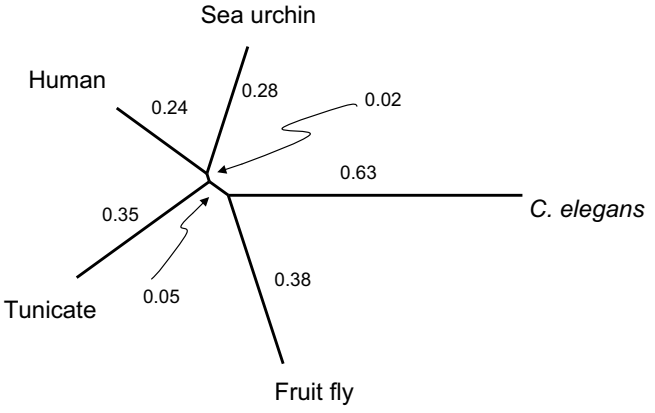


Fig. 1. Phylogeny of metazoan assemblages. The tree is based on the weighted average of protein distances for each gene (see methods) from the set of 1998 genes that these taxa had in common. Units are in average number of estimated changes per amino acid position.

2.1. *Four taxa analysis*

To assess the significance of contrasting phylogenetic trees, we reduced the problem to four taxa as was done in earlier work assessing contrasting phylogenies.[6] Four taxa analysis has the virtue of reducing the number of trees that need be compared. With four taxa there are only three competing unrooted trees and a single internal branch. The significance of the differences in lengths of these three trees can be tested using a simple chi-square test.

For purposes of this explanation we will use simple parsimony to define the different terms. Though the principle applies as well to weighted parsimony, maximum likelihood and protein distances though there are some small quantitative differences with large data sets. Figure 2 presents the topologies of the three unrooted

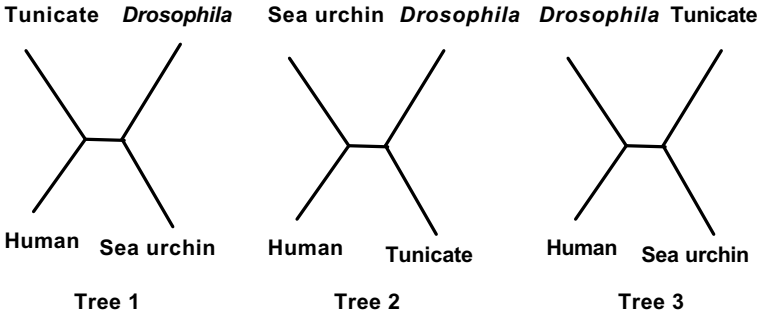


Fig. 2. Three possible unrooted four-taxa topologies. For the four taxa, human (hu), the tunicate (tu), the purple sea urchin (ur) and *Drosophila* (dr), the three possible topologies are written, according to convention as: tree 1=(hu, tu),(ur, dr), tree 2=(hu, ur),(ci, dr) and tree 3=(hu, dr),(ci, ur).

four-taxa trees. In simple parsimony, the best tree is the tree that has the most phylogenetically informative characters (defined as PIC) in its support. Let us assume that tree 1 represents the evolutionary history of the four taxa. If so, then tree 1 will be supported by single changes that occur on the internal branch (these include what are called synapomorphies, but because the tree is unrooted this is not a useful terminology). Tree 1 can also be supported by two or more changes that occur on the distal branches. These multiple changes that produce a PIC are called homoplastic changes (homoplasy in general is defined as convergent and parallel evolution and by reversion to an ancestral state, in four-taxa analysis of unrooted trees these are not distinctions that can be deduced from the data). The only character states that are phylogenetically informative are those where two of the taxa share one character and the other two share a different character. PICs that support tree 1 (there number $= N_1$) will be those where human and the tunicate share a character and the sea urchin and *Drosophila* another. If tree 1 is the correct tree, N_1 will be the sum of changes on the central branch and the homoplastic changes. There will also be PICs where the other two pairs of taxa share characters; these character states can only arise by means of homoplastic changes on the distal branches if tree 1 is the correct tree. If the distal branches are relatively equal in length, and the occurrence of homoplastic changes is randomly distributed, then we would expect to see the number of PICs where human and tunicate share a character (defined as N_2) and those where human and *Drosophila* share a character (defined as N_3) to be equal. Thus we would expect

$$N_1 > N_2 = N_3. \tag{2.1}$$

A priori we can consider N_1 as support for tree 1, N_2 as support for tree 2 and N_3 as support for tree 3. In traditional parsimony analysis the empirical finding of say $N_1 > N_2$ and N_3 is taken as evidence that tree 1 reflects the evolutionary history of the four taxa.

Table 1a shows the number of PICs that support each of the three trees. As can be seen, tree 2 containing the human-sea urchin clade wins over tree 1, as in Fig. 1. There are nearly 45,000 PIC states, but as is apparent from the numbers, many of these must arise through homoplastic replacements. For a rough estimation, if we assume that tree 3 is false, then that implies the nearly 13,000 characters supporting it are due to homoplastic replacements. There are likely comparable numbers supporting the other two trees as well. Thus we can attribute 39,000 PICs due to homoplastic replacements, leaving 6000 shared character states that support the true phylogeny.

2.2. *Curious distribution of phylogenetic informative characters*

There is additional information in Table 1a that sheds light on the evolution of the tunicates. The tunicate and human share more characters than would be expected by the inequality shown in Eq. (2.1) or the chance processes of homoplastic

Table 1. The number of phylogenetic informative characters (PIC) in support of each of the three trees in the four taxa analysis. Average is the number of PICs supporting each tree per gene while Observed is the total number of PICs. Expectations is based on a model where the tree supported by the largest value of N is the true tree (i) and the two false trees (j and k) will have equal numbers of homoplastic replacements thus will be distributed $N_i > N_j = N_k$ where the expected N_i and N_k equals the average of the observed values. The number of phylogenetic informative characters (N) that support tree i is $N_i = (P - 2T_i + T_j + T_k)/3$, where P is the total number of PICs and T is the length of the parsimony tree in units of unweighted amino acid differences. In a four taxa tree, the only PICs are those in which two taxa share one amino acid and the other two share another. (a) Tree 1 is (hu, tu)(ur, dr), tree 2 is (hu, ur)(tu, dr) and tree 3 is (hu, dr)(tu, ur). Based on 2537 proteins. (b) Tree 1 is (hu, xe)(tu, ur), tree 2 is (hu, tu)(xe, ur) and tree 3 is (hu, ur)(xe, tu). Based on 2469 proteins. (c) Tree 1 is (hu, ur)(dr, ce), Tree 2 is (hu, dr)(ur, ce) and tree 3 is (hu, ce)(ur, dr). Based on 1957 proteins.

Four texa analysis of:

	N_1	N_2	N_3
(a) Human, tunicate, sea urchin, *Drosophila.*			
Average	5.8	6.2	4.9
Observed	14,840	15,827	12,678
Expected	13,759	15,827	13,759
Chi-square 170, $P < 10–36$			
(b) Human, *xenopus*, tunicate, sea urchin.			
Average	15.9	2.4	2.3
Observed	40,868	6,188	5,977
Expected	40,868	6,082	6,082
Chi-square 3.6, $P = .2$			
(c) Human, sea urchin, *Drosophila, C. elegans.*			
Average	8.3	4.7	4.5
Observed	16,111	9,286	8,809
Expected	16,111	9,047	9,047
Chi-square 12.6, $P = 0.002$			

replacements if tree 1 were indeed a false tree. The distribution of characters is consistent with a reticulate event and in the Gauthier and Lapointe[5] HBS test would identify human, sea urchin and the sea squirt as a possible triad of a potential parent1, parent2, hybrid. This is also consistent with two recent studies using large number of protein sequences[2,3] that support a tunicate/chordate clade that excludes the sea urchin. To appreciate the support for the tunicate-chordate assemblage in the data in Tables 1a, some controls of the four taxa tests were performed on organisms for which there is no conflict between the classical and molecular phylogenies Tables 1b and 1c. In Table 1b we compare the two vertebrates, human and frog (*Xenopus laevis*), with the sea urchin and *drosophila*. As expected, tree 1 is very strongly supported over trees 2 and 3. Also, the number of phylogenetically informative characters that support the two alternative trees are comparable, as would be expected if homoplastic replacements are randomly distributed among the four distal branches. Table 1c shows a similar control analysis of the two deuterostomes, human and sea urchin, against the two protostomes, *Drosophila* and the *C. elegans*. Again, the classically defined clades are

strongly supported, and we can also see that the number of PICs supporting the two alternative trees is comparable if not equal. The expectations for the chi-square test in Table 1 is based on a model that for the true tree i, the distribution will be $N_i > N_k = N_j$.

Now returning to Table 1a, we can see the possible significance of the distribution of the PICs. Though tree 2 has the most support, the number of PICs supporting tree 1 is significantly larger than those supporting tree 3. As shown, the probability that $N_2 > N_1 = N_3$ represents the distribution is less than 10^{-36}. The simplest explanation for this is that some genes support tree 1 while others, a slight majority, support tree 2.

2.3. *Long-branch attraction?*

There is a potential artifact known as long-branch attractions[7] that needs to be addressed, especially in the current example with relatively long terminal branches and short internal ones. In the present case, the branch leading to the tunicate is slightly longer than the one leading to the sea urchin. Hence *Drosophila*'s even longer branch could have attracted the tunicate. This artifact is easy to understand when we realize that the probability of a homoplastic replacement increases with the total number of replacements in the two branches. To control for long-branch attraction, we used *C. elegans* instead of *Drosophila* as the fourth taxa; *C. elegans* has even a longer branch than does *Drosophila* (see Fig. 1). If Drosophila were attracting the tunicate artifactually, then we would expect *C. elegans* to do so even more. Table 2 shows the observed number of PICs. As in Table 1, tree 2 wins. There is some evidence for long-branch attraction; the high chi square value in Table 2 indicates that there is a significant redistribution of homoplastic characters. This is due mostly to the expected value of N_3 being somewhat larger than the observed value, which is what one would expect if the long-branch to *C. elegans* attracts the longer tunicate branch. However, the expected number of homoplastic replacements lost to tree 3 seems to move more to tree 1 than to tree 2. Long-branch attraction is insufficient to explain why human and sea urchin are more closely related to each other than they are to the tunicate, as is seen in Fig. 1, Table 1a and Table 2.

Table 2. As in Table 1a except that genes from *C. elegans* were used instead of *D. melanogaster*. The expectations are calculated with the assumption that the relative distribution of PICs seen in Table 1a is the same when *C. elegans* replaces *Drosophila*. Total of 2002 proteins.

Human, tunicate, sea urchin, *C. elegans*		
	obs	exp
N_1	10956	10,579
N_2	12,966	12,839
N_3	8831	9367
Chi-square = 45, $P < 10^{-10}$		

2.4. *Two classes of genes*

One mechanism that could account for the unusual distribution of PICs shown in Table 1a is that one group of genes support one phylogeny and another group of genes support a different one. If this is the case then it should be possible to partition the genes into one of the two groups. The procedure that follows is designed to test this hypothesis.

If the set of genes is a mix representing two different phylogenies, then we would expect the set of genes that support tree 1 should show a $N_1 > N_2 = N_3$ distribution, while those that support tree 2 should show a $N_2 > N_1 = N_3$ distribution. Anticipating that many of the individual genes will have too few PICs to distinguish between the competing trees we submitted each to bootstrap sampling.[5] Each of the aligned sequence sets represented by the 2208 genes were re-sampled 200 times. This consisted of determining the tree parameters derived from over 440,000 aligned sets. For each the number of phylogenetic informative characters supporting each of the three trees was recorded. At bootstrap values of greater than 70% we found prominent groups that supported tree 1 and those that supported tree 2. We can see the number of genes supporting tree 1 in Table 3a was significant at 518 and the number supporting tree 2 was somewhat larger at 632, which is consistent with our earlier results. More importantly, each of the two groups closely displayed a $N_i > N_j = N_k$ distribution. In Table 3a, where N_1 is the largest of the three, we see that the values of N_2 and N_3 are comparable if not equal ($P = 0.002$). A similar result is shown in Table 3b where N_2 is larger than N_1 and N_3 and the latter two values are comparable if not equal ($P = 0.03$). In both cases, deviation from equality favors either N_2 or N_1 by just a few hundred PICs, and this deviation is in the direction predicted by a long-branch attraction.

We are interpreting these results as supporting a model that the original set of genes was a mixture of at least two classes of genes, each having a different

Table 3. (a) Set of 518 genes with bootstrap values > 70% supporting tree 1. Average length = 312. Expectations are computed assuming an $N_1 > N_2 = N_3$ distribution. (b) Set of 632 genes with bootstrap values > 70% supporting tree 2. Average length = 338. Expectations are based on an $N_2 > N_1 = N_3$ distribution.

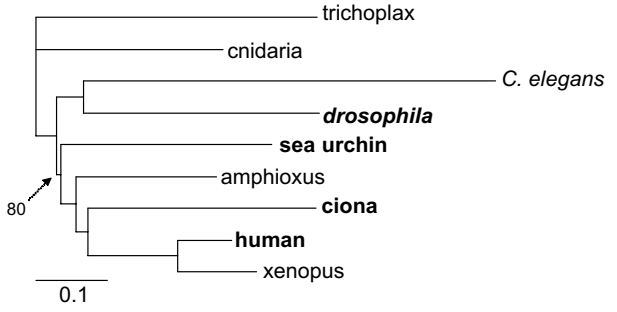
	N_1	N_2	N_3
(a)			
Average	8.49	4.81	4.34
Total observed	4372	2478	2235
Expected	4372	2356	2356
Chi-square = 12.6, $P = 0.002$			
(b)			
Average	4.81	9.05	4.49
Total observed	3037	5713	2832
Expected	2934	5713	2934
Chi-square = 7, $P = 0.03$			

phylogeny. The null hypothesis is that there is only a single class of genes, say those that support tree 1, but that the variances of N_1, N_2 and N_3 is very high. Thus, by this scenario, the bootstrap sampling selects incorrect trees by chance. If this was true we would expect two things that are not in the data, first, tree 2 supporting genes should show a distribution $N_2 > N_1 > N_3$. This is clearly not observed. Second, we would expect that for the tree 1 supporting genes the N_1, N_2, N_3 distribution would more robustly support its tree than would the tree 2 supporting genes. Again this is not observed in the numbers. Indeed the original set of 2208 genes statistically behave as if they are a mixture of genes with at least two different evolutionary histories.

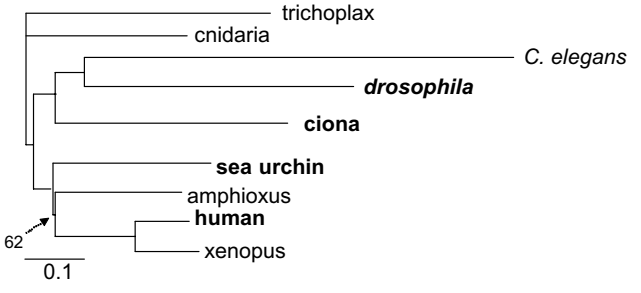
2.5. *The taxon sampling puzzle*

Other studies of metazoan phylogeny using sequences from multiple proteins have encountered a problem that is described as the 'poor' taxon sampling problem. That is, reconstructing clades using a given number of taxa often yields a different result when larger or lesser numbers of taxa are considered. This has been encountered even when large numbers of genes were analyzed.[2] Perhaps it could be argued that the current result is due to poor taxon sampling. If so, then we might expect that relationships reported above would change if larger numbers of taxa were included. To test this we constructed nine taxa trees for the group of proteins from Table 3a (those that support tree 1) and from Table 3b (those that support tree 2). To the four taxa shown in Table 3, we sequentially added proteins from other completed genomes. In addition to the sequences for *Xenopus* and *C. elegans,* we added *Trichoplax adherens* (a very primitive multi-cellular animal that is classified not as a metazoan but a placozoan), *Nematostella vectensis* (a Cnidarian, one of the most primitive metazoa) and *Branchistoma floridae* (amphioxus, another primitive chordate). Figure 3a shows the tree resulting from the set of proteins that originally supported tree 1. As can be seen, the relative relationship of human and *Ciona* with respect to the sea urchin and *Drosophila* is the same as in Table 3a. Figure 3b shows the tree resulting from the set of proteins that originally supported tree 2. Again, the relative relationship of human and sea urchin with respect to *Ciona and Drosophila* is the same as seen in Table 3b. In addition to examining nine taxa shown in Fig. 3, we also looked at five, six, seven and eight taxa trees for both tree 1- and tree 2-supporting genes (data not shown). In each case, the relative relationship of human, sea urchin, *Ciona* and *Drosophila* remains the same, as would be predicted by either tree 1 or tree 2. Thus we can conclude that the two sets of relationships documented in Table 3 are not dependent on the number of taxa examined.

The nine taxa trees shown in Fig. 3 were constructed using the Fitch-Margoliash distance method, but the same relative results were obtained using either parsimony or neighbor joining methods. All bootstrap values are 100% except for the 80% value shown in Fig. 3a and the 62% value shown in Fig. 3b. In Fig. 3a an alternative clade consisting of Sea Urchin, Drosophila and C. elegans is supported at 20% and in



(a) Tree 1 Supporting proteins.



(b) Tree 2 Supporting proteins.

Fig. 3. Proteins that supported tree 1 (Table 3a) and those that supported tree 2 (Table 3b) were used to construct the phylogenies for the same nine taxa in panels (a) and (b) respectively. (a) The 518 proteins from Table 3a were used as queries to probe the nine taxa database. This identified 369 proteins in common to all nine taxa. After multisequence alignment, editing and concatenation, this resulted in a single sequence of 97,737 amino acids. (b) The 632 proteins identified from Table 3b were used as queries to probe the same nine taxa and this identified 436 common proteins. This resulted in a concatenated sequence of 122,992 amino acids. All of the clades have 100% bootstrap support (50 replicates) except the two cases indicated by the arrows. Trees are based on distances as described in methods. Space bar is in JTT adjusted distances in units of amino acid replacements per site.

Fig. 3b a clade consisting of sea urchin and amphioxus is supported at 48%. These two alternative nine taxa topologies, identified by the bootstrap procedure, when evaluated by the maximum likelihood method resulted in log likelihood scores that were not significantly different from each other. It appears that even with the highly concatenated sequence files consisting of over 100,000 amino acids, the internal node involving the sea urchin is not resolved.

However, it should be stressed that the two basic topologies represented by Figs. 3a and 3b (including the ambiguity involving placement of the sea urchin) are highly significantly different using maximum likelihood. Table 4 presents these differences. In this table the log likelihood scores for both of the tree topologies were measured against sequences from the tree 1 and tree 2 input sequence sets.

Table 4. Tree 1 characters and tree 2 characters were defined by the results from Tables 3a and 3b respectively. The tree topologies from Figs. 3a and 3b were used as user defined trees and the log likelihood scores were computed using the maximum likelihood phylogenetics program in Phylip.

	Differences in log likelihood scores Nine taxa tree topology from:	
	Fig. 3a	Fig. 3b
Tree 1 characters	0	5500
Tree 2 characters	4300	0

The values shown are the difference between the best score and the highest. The chances, with log likelihood scores in the multiple thousands, for the two trees being the same are quite low (at least $P < 0.0001$ and probably lower than 10^{-6}) These P values are based on the Shimodaira-Hasegawa test(see methods).

3. Discussion

The simplest explanation for our results is that one of the metazoan genomes consists of at least two sets of genes with different and conflicting histories. That is, one of the four taxa descended from a chimera. In the four taxa analysis the incongruity of the two trees shown in Table 3 is unambiguous, though identification of the taxa that may be derived from the hybrid ancestor is not easily determined. Figure 3 with nine taxa has more information. If we remove *Ciona* from the two trees, we can see that the topology of the two eight taxa trees is the same (though it should be noted that internal branches close to the sea urchin/Amphioxus separation has low bootstrap support). Therefore the simplest explanation for these results is that the ancestor that gave rise to *Ciona* was a hybrid between an early protostome (i.e. related to an ancestor in the *Drosophila-C.elegans* clade) and a vertebrate ancestor (excluding Amphioxus). The reconstructed network from the two nine taxa trees in Fig. 3 is shown in Fig. 4.

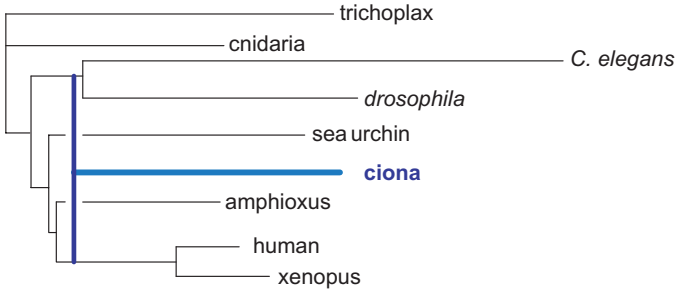


Fig. 4. Combined nine taxa trees into a single network. The two tree topologies in Fig. 3 are incongruent but the eight taxa trees excluding *Ciona* are the same. Shown is the nine taxa network based on the minimum number of reticulations.

The tunicate has traditionally been classified with the chordates because its larval form resembles the tadpole larvae of the chordates. However, the tunicate has an adult form that is completely unique among metazoan phyla. We are not the first to be intrigued with tunicate taxonomy. The tunicate anomaly has perplexed students of biology for more than a century and led Don Williamson[8] to suggest that the tunicate had hybrid origins with those genes controlling larval development coming from a chordate ancestor, and those genes controlling adult development coming from some other phylum.

3.1. *The sampling paradox*

Recent studies have examined large numbers of genes from many different phyla and have reached conclusions quite different from ours.[2,3] We believe that these differences are explicable and that, in fact, these other works have encountered a problem that is quite possibly consistent with our interpretation. This is the 'poor taxon sampling' or the 'taxon sampling artifact' problem mentioned above.[2,3,9–12] This seems to be a problem even when large numbers of genes are being analyzed. In some cases it is suggested that the number of taxa is too few and that the addition of more would lead to reliable phylogenies, while in other cases it looks like removal of taxa are needed to obtain more robust trees. To call this a 'sampling artifact' implies that there is an explanation based on knowable causes. However, though this problem has been encountered repeatedly for over 40 years[12] and recognized for the past 15 as being troublesome for resolving the metazoan phyla,[9] an adequate description of an underlying evolutionary mechanism has not been given, nor has any statistical problem been adequately described.

We would like to point out that the underappreciated evolutionary mechanism of horizontal gene transfer can provide insight into this taxon sampling problem. In practical terms, to say that a particular clade or branch is unstable means that during the bootstrap sampling of character states, some samples will support one tree while different samples will support another. That is, this is confirmation that there is homoplasy in the underlying character set. Homoplastic replacement is not the only mechanism that can cause homoplasy — horizontal gene transfer can as well. The sampling instability phenomena can be understood in this context. Consider a set of organisms that fall into two clades and further consider that all of the genes being sampled within any given genome shared the same history. We would expect the resulting organisms to map into their respective clades. Now what happens when a genome is included that is derived from a hybrid with parents from each of the two clades? There are two things that we would expect. First, the hybrid would have weak bootstrap support for either clade. But also, depending on the relationship of the two ancestors (that made up the hybrid) to other members of the two clades, addition of the hybrid to the data set could weaken already established relationships. Something like this could be influencing the data in the paper by Dunn *et al.*[2] where they found that by removing 12 organisms from their

data set that had weak bootstrap support for any clade, a number of the remaining organisms mapped into clades with strengthened bootstrap values. We suggest that those organisms causing tree instability are also evolved from hybrids and further, we suggest that one of their ancestors is related to those organisms whose bootstrap value to a clade is strengthened by removal of the hybrid.

We argue that it is differences in gene sampling and not taxon sampling that account for the differences in our results compared to the results of others. As an example from the current study, when we looked at a set of 60 ribosomal proteins, we found they supported the human-tunicate clade to the exclusion of the sea-urchin and *Drosophila* (data not shown). That is, for those taxa that have a major hybridization event in their past, one sampling of its genome may support one phylogeny while a different sampling may support another.

4. Methods

The nearly 30,000 protein sequences from the human genome sequence were used as query sequences in Blast[13] searches against a database consisting of the protein sequences obtained from the genome projects for the following metazoans: human,[14] Xenopus *laevis*,[15] the tunicate *Ciona intestinalis*,[16] the sea urchin, *Strongylocentrotus purpuratus*,[17] the fruit fly, *Drosophila melanogaster*[18] and the round worm *Caenorhabditis elegans*.[19] In another blast search the protein complement from the genomes of the amphioxus *Branchistoma floridae*,[20] *Trichoplax adherens*[21] and the Cnidarian *Nematostella vectensis*[22] were included. A Blast expectation score cutoff of less than 10^{-13} was used in all cases. This produced 30,000 blast output files. These were screened such that each contained at least one homologue for the tunicate, the sea urchin, and *Drosophila*. Large gene families were excluded by removing those files that had more than ten proteins from either the human, *Drosophila* or *C. elegans*. These various filters reduced the number of usable blast output files to about 3500. In the four taxa analysis, one sequence from each of the four taxa was recorded. For those proteins that had multiple listings for the same taxa, the protein with the smallest expectation value was used. The procedure outlined here does not guarantee that the data sets do not contain paralogues. In results not shown we repeated the analysis shown in Fig. 1 and Table 1 but excluded the blast output files that had more than either five or 20 entries for the same taxa. Other than changing the number of protein sets, changing gene family limit did not effect the pattern of the results, i.e. the topology of the tree in Fig. 1 or the $N_2 > N_1 > N_3$ distribution of PICs in Table 1. In addition we identified a group of about 30 protein sets (on the basis of being statistical outliers in the N_1, N_2, N_3 distribution) that clearly contained paralogs. Removing these protein sets did not affect the pattern of the results. These results indicate that the paralogs are distributed in a uniform manner among the taxa and are not biasing outcome.

Sequences from each blast output file were recovered, and multisequence alignments were performed using Clustal.[23] The aligned sequences were then edited

with the sequence editor Gblocks[24] to remove those regions that were within indels or that were hyper-variable. These sequences were then submitted to phylogenetic analysis using Phylip.[25] Three types of trees were obtained for each set of sequences — a parsimony tree using either the Pars subroutine or a distance tree using Protdist and the Fitch-Margoliash option. Protein distances were calculated for each aligned set and averaged over all sets or in some cases by concatenating the sequences and directly measuring distance. The Jones, Thornton, Taylor distance matrix[26] was used to determine distances. It was shown that distances up to 2.5 changes per residue were linear with time of divergence (data not shown), and those with distances in excess of 2.5 were removed from further consideration. The phylogenetic maximum likelihood program proml was used to calculate log likelihood scores that uses the Shimodaira-Hasegawa test.[27] Programs within Phylip were also used to perform the bootstrap procedure. TreeView[28] was used for tree visualization. All computations were performed on a standard pc with a Linux OS and data was processed using UNIX script files, Perl scripts and standard spread/sheets. Two procedures were used to transform the incongruent phylogenetic trees into a single network — these were triplet analysis[29] and a tree reduction approach[30] (described in the text). Reticulation was solved manually given that the program outputs[29,30] are difficult to translate into evolutionary events.

Acknowledgments

We thank Simone Linz and Balaji Venkatachalam for helpful discussions over some of the concepts in this paper. We thank an anonymous referee who made a number of useful suggestions that helped sharpen the text. Some of this material was presented at a Linnean Society meeting in July, 2008.

References

1. Syvanen M, Cross-species gene transfer: implications for a new theory of evolution, *J Theor Biol* **112**:333–343, 1985.
2. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sørensen MV, Haddock SH, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G, Broad phylogenomic sampling improves resolution of the animal tree of life, *Nature* **452**:745–749, 2008.
3. Delsuc F, Brinkmann H, Chourrout D, Philippe H, Tunicates and not cephalochordates are the closest living relatives of vertebrates, *Nature* **439**(7079):965–968, 2006.
4. Huson DH, Bryant D, Application of phylogenetic networks in evolutionary studies, *Mol Biol Evol* **23**:254–267, 2006.
5. Gauthier O, Lapointe, F, Hybrids and phylogenetics revisited: a statistical test of hybridization using quartets, *Syst Bot* **32**:8–15, 2007.
6. Syvanen M, On the Occurrence of horizontal gene transfer among an arbitrarily chosen group of 26 genes, *J Mol Evol* **54**:258–266, 2002.

7. Felsenstein J, Cases in which parsimony or compatibility methods will be positively misleading, *Syst Zool* **27**:401–410, 1978.
8. Williamson DI, *Larvae and Evolution*, Chapman and Hall, London, 1992.
9. Lecointre G, Philippe H, Lê HL, Le Guyader H, Species sampling has a major impact on phylogenetic inference, *Mol Phylogenet Evol* **2**:205–224, 1993.
10. Matus DQ, Copley RR, Dunn CW, Hejnol A, Eccleston H, Halanych KM, Martindale, MQ, Telford MJ, Broad taxon and gene sampling indicate that chaetognaths are protostomes, *Curr Biol* **8**:R575–R576, 2006.
11. Philip GK, Creevey CJ, McInerney JO, The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa, *Mol Biol Evol* **22**:1175–1184, 2005.
12. Blair JE, Ikeo K, Gojobori T, Hedges SB. The evolutionary position of nematodes, *BMC Evol Biol* **8**(2):7, 2002.
13. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res* **25**(17):3389–3402, 1997.
14. Human Genome Resources, National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/.
15. *Xentra tnpicalis* genome assembly, http://genome.jgi-psf.org/Xentr4/Xentr4.home.html.
16. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins, *Science* **298**:2157–2167, 2002.
17. Sea Urchin Genome Sequencing Consortium: Erica Sodergren *et al.* The genome of the sea urchin *Strongylocentrotus Purpuratus*, *Science* **314**:941–952, 2006.
18. Celniker SE *et al.* Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence, *Genome Biol* **3**:1–0079, 2002.
19. The *C. elegans* Sequencing Consortium Genome Sequence of the Nematode *C. elegans*: a platform for investigating biology, *Science* **11**:2012–2018, 1998.
20. Putnam NH, Butts T, Ferrier DEK *et al.* The amphioxus genome and the evolution of the chordate karyotype, *Nature* **453**:1064–1071, 2008.
21. Srivastava M, Begovic E, Chapman J *et al.* The Trichoplax genome and the nature of placozoans, *Nature* **454**:955–960, 2008.
22. Sullivan JC, Ryan JF, Watson JA *et al.* StellaBase: The *Nematostella Vectensis* Genomics Database, *Nucleic Acids Res* **1**:34, 2006.
23. Thompson JD, Higgins DG, Gibson TJ, CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res* **22**:4673–4680, 1994.
24. Castresana J, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, *Mol Biol Evol* **17**:540–552, 2000.
25. Felsenstein J, PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle, 2005.
26. Jones DT, Taylor WR, Thornton JM, The rapid generation of mutation data matrices from protein sequences, *Comput Appl Biosci* **8**:275–282, 1992.
27. Shimodaira H, Hasegawa M, Multiple comparisons of log-likelihoods with applications to phylogenetic inference, *Mol Biol Evol* **16**:1114–1116, 1999.

28. Page RDM, TREEVIEW: An application to display phylogenetic trees on personal computers, *Comput Appl Biosci* **12**:357–358, 1996.
29. van Iersel L, Kelk S, Constructing the simplest possible phylogenetic network from triplets, algorithmica, DOI: 10.1007/s00453-009-9333-0, 2009.
30. Linz S, Radtke A, von Haeseler A, A likelihood framework to measure horizontal gene transfer, *Mol Biol Evol* **24**:1312–1319, 2007.