LETTER

doi:10.1038/nature10526

# Resolving the evolutionary relationships of molluscs with phylogenomic tools

Stephen A. Smith[1,2], Nerida G. Wilson[3,4], Freya E. Goetz[1], Caitlin Feehery[1,4], Sónia C. S. Andrade[5], Greg W. Rouse[4], Gonzalo Giribet[5] & Casey W. Dunn[1]

**Molluscs (snails, octopuses, clams and their relatives) have a great disparity of body plans and, among the animals, only arthropods surpass them in species number. This diversity has made Mollusca one of the best-studied groups of animals, yet their evolutionary relationships remain poorly resolved[1]. Open questions have important implications for the origin of Mollusca and for morphological evolution within the group. These questions include whether the shell-less, vermiform aplacophoran molluscs diverged before the origin of the shelled molluscs (Conchifera)[2–4] or lost their shells secondarily. Monoplacophorans were not included in molecular studies until recently[5,6], when it was proposed that they constitute a clade named Serialia together with Polyplacophora (chitons), reflecting the serial repetition of body organs in both groups[5]. Attempts to understand the early evolution of molluscs become even more complex when considering the large diversity of Cambrian fossils. These can have multiple dorsal shell plates and sclerites[7–10] or can be shell-less but with a typical molluscan radula and serially repeated gills[11]. To better resolve the relationships among molluscs, we generated transcriptome data for 15 species that, in combination with existing data, represent for the first time all major molluscan groups. We analysed multiple data sets containing up to 216,402 sites and 1,185 gene regions using multiple models and methods. Our results support the clade Aculifera, containing the three molluscan groups with spicules but without true shells, and they support the monophyly of Conchifera. Monoplacophora is not the sister group to other Conchifera but to Cephalopoda. Strong support is found for a clade that comprises Scaphopoda (tusk shells), Gastropoda and Bivalvia, with most analyses placing Scaphopoda and Gastropoda as sister groups. This well-resolved tree will constitute a framework for further studies of mollusc evolution, development and anatomy.**

Since the first animal phylogenies based on molecular data, many researchers have struggled to resolve mollusc phylogenies even as taxon sampling improved[5,6,12] (see Fig. 1 for some hypotheses that have been proposed). Little support, if any, was found for the monophyly of Mollusca or most of its larger subclades. Better results were achieved for some internal relationships of these groups, including Polyplacophora, Bivalvia, Cephalopoda, Scaphopoda and Gastropoda, although often with difficulties recovering monophyly of the two largest clades, the gastropods and bivalves[5,13,14]. Unfortunately, fundamental questions in mollusc evolution remain largely unanswered by the molecular and morphological data. These questions include whether the aplacophoran molluscs are monophyletic[2] or paraphyletic[3,4]. There has also been conflicting evidence for the placement of Polyplacophora, which has been placed with the aplacophorans (forming the clade Aculifera), as the sister group to the shelled molluscs (forming the clade Testaria) or as the sister group to Monoplacophora (forming the clade Serialia). In addition, many hypotheses have been proposed for the interrelationships of the conchiferan groups. The extensive fossil record of Mollusca (which dates back to the Cambrian), combined with the numerous Palaeozoic forms that are considered stem-group molluscs and the lack of resolution in targeted-gene approaches to molluscan phylogenetics, pointed towards a possible rapid radiation with little phylogenetic signal left in the genomes of molluscs. However, the same has been argued for the radiation of Metazoa in the Cambrian or earlier[15], but large increases in gene representation using phylogenomic analyses have clearly ameliorated this problem and identified relationships that seemed impossible to resolve with target-gene

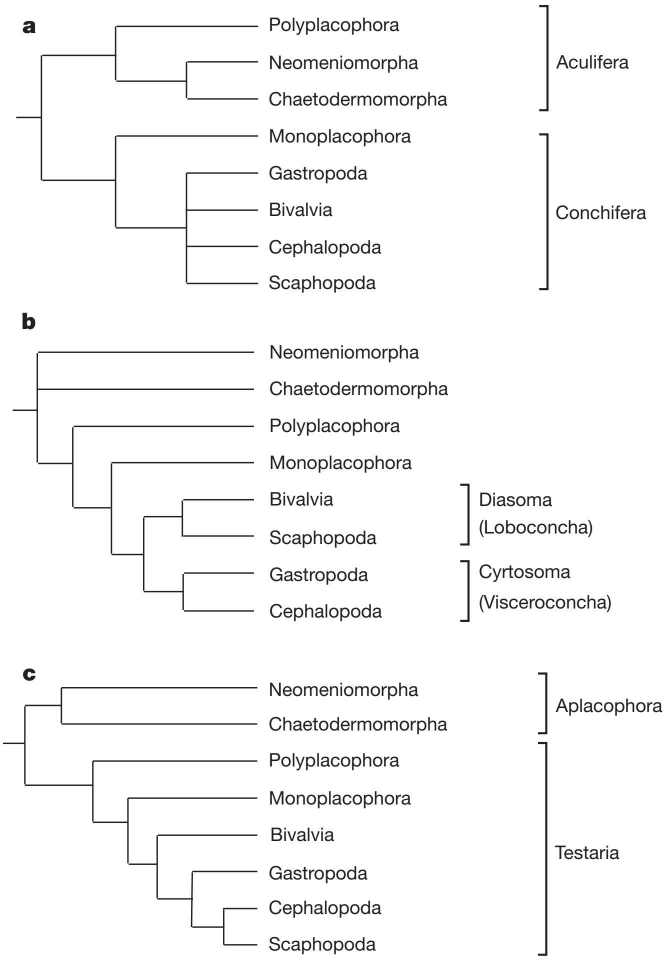

**Figure 1 | Selected hypotheses of extant molluscan relationships and relevant taxa.** Phylograms based on the hypotheses of Scheltema[3] (**a**), Salvini-Plawen and Steiner[2] (**b**) and Waller[28] (**c**). Most controversy centres on the monophyly of Aplacophora, the relationships within Conchifera and the placement of Polyplacophora (for example, in Aculifera versus in Testaria).

---

[1]Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island 02912, USA. [2]Heidelberg Institute for Theoretical Studies, Heidelberg D-69118, Germany. [3]The Australian Museum, Sydney, New South Wales 2010, Australia. [4]Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093, USA. [5]Museum of Comparative Zoology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

approaches[16,17]. We applied the same principles to Mollusca, one of the most challenging problems to solve in animal phylogenetics.

Only phylogenomic analyses have been able to recover molluscan monophyly with high support[17]; however, few molluscs were included in earlier analyses, and not all of the major subclades were represented. Therefore, little could be concluded about the interrelationships of the major molluscan groups. Morphology-based cladistic analyses have often relied on 'idealized' composite ground patterns to represent entire clades[2,4], a practice that has now been largely replaced with the use of exemplar species[18] and more detailed character descriptions. But an analysis of molluscan morphological features coding real exemplars has yet to be published, and the exemplar approach is much more amenable to molecular data.

Analyses of our broadly sampled, new phylogenomic data (see Methods, Supplementary Table 1 and Supplementary Fig. 1) result in a well-resolved and highly supported phylogeny of Mollusca (Fig. 2), in contrast to all previous molecular attempts[5,6,12]. These results are consistent across analytical methods, phylogenetic inference programs, matrices that vary in occupancy and the number of genes considered (Fig. 2 and Supplementary Figs 2–6 and 9), and the inclusion of different outgroup taxa (Supplementary Fig. 7).

Our results (Fig. 2) show a sister group relationship between the aculiferan molluscs and the conchiferan groups. Aculifera[3,19] includes Polyplacophora as the sister group to the two aplacophoran groups (Neomeniomorpha and Chaetodermomorpha). This topology lends support to the idea that the vermiform Aplacophora are not plesiomorphic but are derived from plated Palaeozoic molluscs such as *Acaenoplax*[10]. The aculiferans are characterized by spicules and dorsal shell plates. Chitons have eight dorsal shell plates, but their larva has an anlagen with seven rows of dorsal papillae, as observed in the serially arranged spiculoblasts of a chaetodermomorph larva[20], a character that may constitute a synapomorphy of the clade.

Conchifera is supported as a clade, suggesting that true shells may have originated only once, perhaps by the concentration of a diffuse shell gland into a single zone of the mantle (two zones in bivalves), at least as defined by the role of *engrailed* during organogenesis[21]. The support here for Conchifera rejects the recent Serialia hypothesis[5,6]. Comparing the site likelihoods in analyses in which Serialia is constrained with those in which it is not constrained reveals that there are many more characters that are incongruent with Serialia than support Serialia (Supplementary Fig. 8). Monoplacophora is not, however, the sister group of all other Conchifera, as has been suggested by most authors, and is instead the sister group to Cephalopoda, as has been proposed based on some palaeontological data[22]. Many palaeontologists have accepted the monoplacophoran 'ancestry' of Cephalopoda[23,24], although this relationship has been rejected by neontologists, who

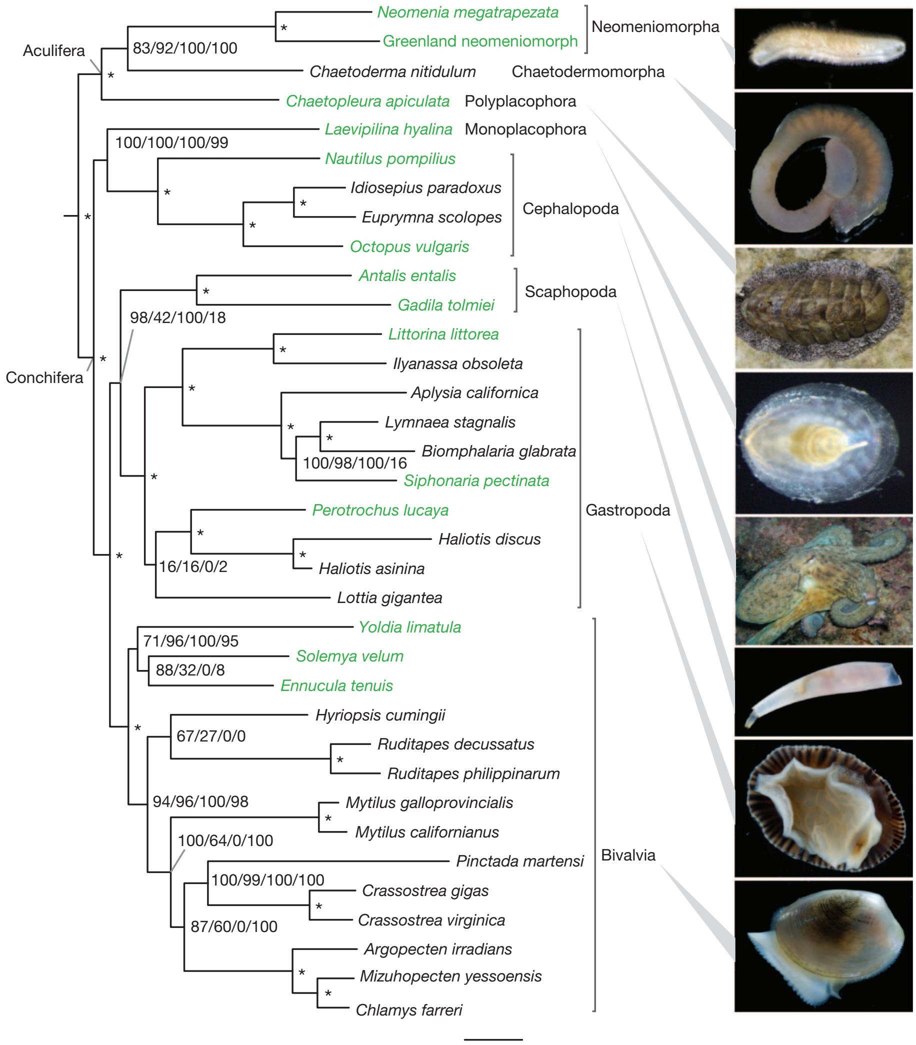

**Figure 2 | Phylogram of the RAxML maximum likelihood analysis of the big matrix (216,402 amino acids) under the WAG+$\Gamma$ model.** Support values for the topology obtained from four analyses are listed as percentages in the order A/B/C/D. A is the bootstrap support from RAxML analysis under the WAG model for the big matrix. B is the bootstrap from RAxML analysis under the WAG model for the small matrix. C is the posterior probability from MrBayes under the WAG model for the small matrix. D is the posterior probability from PhyloBayes under the CAT model for the small matrix. Asterisks indicate 100/100/100/100 support. Taxa with new data are shown in green. Scale bar, 0.08 expected changes per site.

consider that cephalopods and gastropods share important morphological features such as the presence of cephalic eyes, the isolation of the head from the visceral mass, the terminal position of the mantle cavity and the occurrence of muscle antagonistic systems[2,23]. The presence of multiseptate shells in fossil Hypseloconidae monoplacophorans, a character that is found in *Nautilus* and fossil cephalopods, has been interpreted as supporting this relationship between Cephalopoda and Monoplacophora[24]. The presence of two pairs of gills, kidneys and atria in the chambered *Nautilus* has been interpreted as an indication that secondary simplification took place during the early evolution of cephalopods from an ancestor with serially repeated structures[23]. This interpretation and the present trees suggest that the most recent common ancestor of Cephalopoda and Monoplacophora had some serially repeated structures.

The internal resolution of Cephalopoda is in agreement with all of the current hypotheses, with the chambered *Nautilus* forming the sister group of Coleoidea, and also identifies the monophyly of Decapoda[25,26]. Scaphopoda, Gastropoda and Bivalvia form a clade with thick multilayered shells, but this clade has received little attention in the literature[23]. Most morphological hypotheses place scaphopods as the sister group to bivalves in a clade named Diasoma[2,27] and, recently, molecular and developmental data have favoured a cephalopod–scaphopod relationship. Although there is strong support for the placement of Scaphopoda as the sister group to Gastropoda in maximum likelihood analyses of the big matrix (Fig. 2 and Supplementary Figs 2 and 4), maximum likelihood analyses of the small matrix recover this same relationship but with less support (Fig. 2 and Supplementary Figs 3 and 5). Bayesian analyses using the site-heterogeneous CAT model of protein evolution also place Scaphopoda as the sister group to Bivalvia, with a posterior probability of 81% (Supplementary Fig. 9).

Within Bivalvia, maximum likelihood analyses and Bayesian analyses under the Whelan and Goldman (WAG) model support the monophyly of Protobranchia, which includes bivalves with plesiomorphic ctenidia—gills comparable to those of many other molluscs. This contradicts some earlier bivalve phylogenies, based on fewer data, that proposed paraphyly of protobranchs[14] but supports the traditional morphological views[2,28]. Bayesian analyses with the CAT model are consistent with Protobranchia but do not provide strong support for it. The hypertrophied bivalve gill, which is responsible for filter feeding, had a single origin, and organisms with this type of gill constitute the well-recognized clade Autolamellibranchiata. Palaeoheterodonta (the group that includes freshwater pearl mussels) is the sister group to all other autolamellibranchiates, which can be divided into heterodonts and pteriomorphians. This hypothesis is similar to that proposed by some palaeontologists, although additional taxa, especially *Neotrigonia*, Anomalodesmata and Archiheterodonta, must be included before concluding more about the internal autolamellibranchiate relationships.

Likewise, the internal relationships of Gastropoda, although still limited in their taxonomic representation (the group includes nearly 100,000 living species), support some of the major divisions that are currently accepted[29]. The patellogastropod *Lottia* is either the sister group to Vetigastropoda (as in Thiele's Archaeogastropoda hypothesis) or the sister group to all other gastropods[29], depending on the data set that is analysed. The former alternative has been recovered in recent molecular analyses of gastropods[30]. The two representatives of Caenogastropoda form a sister clade to the representatives of Heterobranchia, including opisthobranchs and pulmonates, as suggested in all of the modern analyses of gastropod relationships[13,29].

For the first time, our data and analyses resolve the broad-scale relationships within Mollusca with strong support. This allows us to gain an understanding not only of the relationships of modern molluscs but also of the numerous Palaeozoic forms of molluscs. It also allows us to investigate several key characters that define the group. Molluscs are related to other animals with spiral development and a trochophore larva and have now been shown to share a close ancestor with annelids and brachiopods[16], both of which use chaetoblasts to produce chaetae. Spicules and chaetae may share a similar developmental mechanism[17]. Likewise, the appearance of dorsal plates or shells in addition to sclerites is now well documented in halwaxiids, *Acaenoplax* and Polyplacophora. These features are generated by multiple rows of secretory papillae in chiton and aplacophoran larvae. They may be plesiomorphic among molluscs, especially if halwaxiids are interpreted as stem-group molluscs, but they could also be apomorphic for Aculifera. The condensation of such papillae into a single shell gland[21] could be responsible for the origin of the conchiferan shell, arguably the single event that led to the extraordinary success of molluscs, first in the Cambrian oceans and later in many limnic and terrestrial environments. In addition to the presence of shell glands that can deposit calcium carbonate, the primitive mollusc may have had a rasping radula and serially repeated ctenidia along the mantle cavity, because both characters appear in the two lineages of extant molluscs, Aculifera and Conchifera, as well as in several extinct Palaeozoic stem molluscs. Like the arthropods, with their hardened exoskeletons, molluscs are true conquerors of our land and waters.

## METHODS SUMMARY

New transcriptome data were collected for 14 mollusc species that had been selected to optimize taxonomic representation (Supplementary Table 1). Collecting efforts included an oceanographic campaign to collect members of the key taxon Monoplacophora. Using several protocols, messenger RNA was extracted, and cDNA samples were sequenced on a 454 Genome Sequencer FLX Titanium (Roche) or a Genome Analyzer IIx (Illumina). After assembly and translation, the sequences from all taxa were compared to each other with BLASTP. These pairwise comparisons were used to cluster genes into homologues using the algorithm MCL. The phylogenetic analyses divided sets of homologues into orthologues, which were aligned, trimmed and concatenated into two supermatrices that differed in the number of genes and the average fraction of genes available for each species. The 'small' matrix consists of 301 genes that are present in at least 20 taxa. This matrix has 50% gene occupancy (that is, sequence data were available for an average of 50% of the genes across the taxa), 41% character occupancy (that is, 41% of the matrix consists of unambiguous amino acid data, with the remainder being missing data or alignment gaps) and is 50,930 sites in length. The 'big' matrix consists of 1,185 genes that are present in at least 15 taxa. This matrix has 40% gene occupancy, 32% character occupancy and is 216,402 sites long. Both matrices contain data for all of the 46 species that were included in the study. The matrices were analysed with the programs RAxML, MrBayes and PhyloBayes to infer relationships.

**Full Methods** and any associated references are available in the online version of the paper.

**Received 7 April; accepted 31 August 2011.**
**Published online 26 October 2011.**

1. Ponder, W. F. & Lindberg, D. R. (eds) *Phylogeny and Evolution of the Mollusca* (Univ. California Press, 2008).
2. Salvini-Plawen, L. V. & Steiner, G. in *Origin and Evolutionary Radiation of the Mollusca* (ed. Taylor, J. D.) 29–51 (Oxford Univ. Press, 1996).
3. Scheltema, A. H. in *Origin and Evolutionary Radiation of the Mollusca* (ed. Taylor, J. D.) 53–58 (Oxford Univ. Press, 1996).
4. Haszprunar, G. Is the Aplacophora monophyletic? A cladistic point of view. *Am. Malacol. Bull.* **15,** 115–130 (2000).
5. Giribet, G. *et al.* Evidence for a clade composed of molluscs with serially repeated structures: monoplacophorans are related to chitons. *Proc. Natl Acad. Sci. USA* **103,** 7723–7728 (2006).
6. Wilson, N. G., Rouse, G. W. & Giribet, G. Assessing the molluscan hypothesis Serialia (Monoplacophora + Polyplacophora) using novel molecular data. *Mol. Phylogenet. Evol.* **54,** 187–193 (2010).
7. Vinther, J. & Nielsen, C. The Early Cambrian *Halkieria* is a mollusc. *Zool. Scr.* **34,** 81–89 (2005).
8. Scheltema, A. H., Kerth, K. & Kuzirian, A. M. Original molluscan radula: comparisons among Aplacophora, Polyplacophora, Gastropoda, and the Cambrian fossil *Wiwaxia corrugata*. *J. Morphol.* **257,** 219–245 (2003).
9. Morris, S. C. & Caron, J. B. Halwaxiids and the early evolution of the lophotrochozoans. *Science* **315,** 1255–1258 (2007).
10. Sutton, M. D., Briggs, D. E. G., Siveter, D. J. & Siveter, D. J. Computer reconstruction and analysis of the vermiform mollusc *Acaenoplax hayae* from the Herefordshire Lagerstätte (Silurian, England), and implications for molluscan phylogeny. *Palaeontology* **47,** 293–318 (2004).
11. Caron, J.-B., Scheltema, A., Schander, C. & Rudkin, D. A soft-bodied mollusc with radula from the Middle Cambrian Burgess Shale. *Nature* **442,** 159–163 (2006).

12. Passamaneck, Y. J., Schander, C. & Halanych, K. M. Investigation of molluscan phylogeny using large-subunit and small-subunit nuclear rRNA sequences. *Mol. Phylogenet. Evol.* **32,** 25–38 (2004).
13. Aktipis, S. W., Giribet, G., Lindberg, D. R. & Ponder, W. F. in *Phylogeny and Evolution of the Mollusca* (eds Ponder, W. F. & Lindberg, D. R.) 201–237 (Univ. California Press, 2008).
14. Giribet, G. & Distel, D. L. in *Molecular Systematics and Phylogeography of Mollusks* (eds Lydeard, C. & Lindberg, D. R.) 45–90 (Smithsonian Books, 2003).
15. Rokas, A., Krüger, D. & Carroll, S. B. Animal evolution and the molecular signature of radiations compressed in time. *Science* **310,** 1933–1938 (2005).
16. Hejnol, A. *et al.* Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. R. Soc. B* **276,** 4261–4270 (2009).
17. Dunn, C. W. *et al.* Broad taxon sampling improves resolution of the animal tree of life. *Nature* **452,** 745–749 (2008).
18. Prendini, L. Species or supraspecific taxa as terminals in cladistic analysis? Groundplans versus exemplars revisited. *Syst. Biol.* **50,** 290–300 (2001).
19. Vendrasco, M. J., Wood, T. E. & Runnegar, B. N. Articulated Palaeozoic fossil with 17 plates greatly expands disparity of early chitons. *Nature* **429,** 288–291 (2004).
20. Nielsen, C., Haszprunar, G., Ruthensteiner, B. & Wanninger, A. Early development of the aplacophoran mollusc *Chaetoderma. Acta Zool.* **88,** 231–247 (2007).
21. Wanninger, A., Koop, D., Moshel-Lynch, S. & Degnan, B. M. in *Phylogeny and Evolution of the Mollusca* (eds Ponder, W. F. & Lindberg, D. R.) 427–445 (Univ. California Press, 2008).
22. Yochelson, E. L. An alternative approach to the interpretation of the phylogeny of ancient mollusks. *Malacologia* **17,** 165–191 (1978).
23. Runnegar, B. in *Origin and Evolutionary Radiation of the Mollusca* (ed. Taylor, J. D.) 77–87 (Oxford Univ. Press, 1996).
24. Yochelson, E. L., Flower, R. H. & Webers, G. F. The bearing of the new Late Cambrian monoplacophoran genus *Knightoconus* upon the origin of Cephalopoda. *Lethaia* **6,** 275–309 (1973).
25. Lindgren, A. R., Giribet, G. & Nishiguchi, M. K. A combined approach to the phylogeny of Cephalopoda (Mollusca). *Cladistics* **20,** 454–486 (2004).
26. Strugnell, J. & Nishiguchi, M. K. Molecular phylogeny of coleoid cephalopods (Mollusca: Cephalopoda) inferred from three mitochondrial and six nuclear loci: a comparison of alignment, implied alignment and analysis methods. *J. Molluscan Stud.* **73,** 399–410 (2007).
27. Runnegar, B. & Pojeta, J. Jr. Molluscan phylogeny: the paleontological viewpoint. *Science* **186,** 311–317 (1974).
28. Waller, T. R. in *Bivalves: An Eon of Evolution* (eds Johnston, P. A. & Haggart, J. W.) 1–45 (Univ. Calgary Press, 1998).
29. Ponder, W. F. & Lindberg, D. R. Towards a phylogeny of gastropod molluscs: an analysis using morphological characters. *Zool. J. Linn. Soc.* **119,** 83–265 (1997).
30. Aktipis, S. W. & Giribet, G. A phylogeny of Vetigastropoda and other 'archaeogastropods': re-organizing old gastropod clades. *Invertebr. Biol.* **129,** 220–240 (2010).

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

**Acknowledgements** This research was supported by the US National Science Foundation through the Systematics Program (awards 0844596, 0844881 and 0844652), the AToL Program (EF-0531757), EPSCoR (Infrastructure to Advance Life Sciences in the Ocean State, 1004057) and the iPlant Collaborative (0735191). Support was also provided by the Scripps Institution of Oceanography, the University of California Ship Funds and the Museum of Comparative Zoology. Collecting in Greenland was supported by the Carlsberg Foundation. A. Riesgo and J. Harasewych provided tissue samples of *Octopus* and *Perotrochus*, respectively. E. Röttinger assisted with *Nautilus*. C. Palacín allowed us to use an *Octopus vulgaris* photograph. At Brown University, Illumina sequencing was enabled by the Genomics Core Facility, and computational analyses were facilitated by L. Dong and the Center for Computing and Visualization. At Harvard University, Illumina sequencing was enabled by the Bauer Core in the Faculty of Arts and Sciences (FAS) Center for Systems Biology, and analyses were supported by the staff of the Research Computing cluster Odyssey facility in the FAS.

**Author Contributions** C.W.D., G.G. and N.G.W. conceived of and oversaw the study. S.A.S. and C.W.D. designed and implemented the data analyses. N.G.W., G.G. and G.W.R. collected the specimens. F.E.G., S.C.S.A. and C.F. prepared the specimens for sequencing. S.A.S., C.W.D., G.G., G.W.R. and N.G.W. wrote the manuscript. All authors read and provided input into the manuscript and approved the final version.

**Author Information** Illumina and 454 reads have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive under accession number SRA044948. Sanger reads for *Laevipilina hyalina* have been deposited in the NCBI Trace Archive under the sequencing centre name BUDL with TI range 2317135955-2317139410. The assembled data, matrices and trees have been deposited in Dryad (http://dx.doi.org/10.5061/dryad.24cb8). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.G.W. (nerida.wilson@austmus.gov.au), G.G. (ggiribet@oeb.harvard.edu) or C.W.D. (casey_dunn@brown.edu).

## METHODS

**Taxon sampling and RNA isolation.** The taxa were selected to optimize taxonomic representation within Mollusca. Collecting efforts included an oceanographic campaign to collect members of the key taxon Monoplacophora[31]. New transcriptome data were collected for one outgroup taxon, *Lingula anatina*, and for 14 other taxa that were broadly sampled across Mollusca (Supplementary Table 1). All tissues were collected fresh and were prepared immediately or preserved for subsequent RNA work. Stored tissue was frozen (at −80 °C) or added to RNA*later* (and frozen at −80 °C or −20 °C). Total RNA was isolated with TRI Reagent (Invitrogen) and further cleaned up with an RNeasy kit (QIAGEN), including a DNase I digestion step.

**Sequencing.** Samples were sequenced on a 454 Genome Sequencer FLX Titanium (Roche) or a Genome Analyzer IIx (GA IIx, Illumina). The sample preparation protocol and sequencing technology used for each sample is listed in Supplementary Table 1.

All 454 samples were sequenced by 454 Life Sciences on one-eighth of a Titanium flow cell. For five of the 454 samples, RNA was sent to the sequencing facility for library preparation and sequencing according to the standard 454 cDNA protocols (these samples are marked Roche in the Library Protocol column of Supplementary Table 1). *Nautilus pompilius* mRNA was enriched by one round of binding to Dynabeads (Invitrogen); for the other specimens, total RNA was sent to the sequencing centre, where mRNA enrichment was performed. For four of the 454 samples, full-length cDNA was prepared according to a template-switching protocol[32] (these samples are marked TS in the Library Protocol column of Supplementary Table 1). Adaptors were modified to include restriction sites and were removed by cleavage before sequencing. An MmeI site was incorporated into the 3′ adaptor (5′-ATT CTA GAG CGC ACC TTG GCC TCC GAC TTT TCT TTT CTT TTT TTT TCT TTT TTT TTT VN-3′, where V and N are ambiguous nucleotides), and a SfiI site (5′-AAG CAG TGG TAT CAA CGC AGA GTG GCC ACG AAG GCC GGG-3′) or an AsiSI site (5′-AAG CAG TGG TAT CAA CGC AGA GTG CGA TCG CGG G-3′) was included in the 5′ adaptor. Titanium sequencing reagents were used for all samples. Additional expressed sequence tags for *L. hyalina* were sequenced with Sanger technology according to previously described methods[17].

Most Illumina samples were prepared with the NEBNext mRNA Sample Prep kit (New England BioLabs), with size selection for 400 base pair (bp) products. These samples were sequenced (paired-end, 104 bp), with one per lane on an Illumina GA IIx at the Genomics Core Facility at Brown University. One sample (marked Fragmentase in the Library Protocol column of Supplementary Table 1) was prepared with a modified NEBNext mRNA protocol, in which the full-length cDNA was fragmented with NEBNext dsDNA Fragmentase (New England BioLabs) instead of the mRNA being fragmented. This sample was sequenced (paired-end, 150 bp) in a single lane on an Illumina GA IIx at the FAS Center for Systems Biology at Harvard University.

**Assembly.** Publicly available data from the NCBI dbEST database were processed with a version of the PartiGene pipeline (version 3.0.5)[33] that had been modified to run without user intervention. Trace Archive data were processed as described previously[16].

Roche 454 data were assembled with the Newbler GS *De novo* Assembler (version 2.3, Roche) with the flags '-cdna -nrm -nosplit'. In cases in which multiple splice variants (isotigs in Newbler terminology) were produced for a gene (an isogroup in Newbler terminology), a single exemplar splice variant was selected. The selected isotig was the one with the highest geometric mean of reads spanning each splice site between contigs. This roughly corresponds to the most abundant splice variant for the gene. Singletons that were not assembled by Newbler were assembled with CAP3 (version 10/15/07, with the options '-z 1' and '-y 100'). The sequences that were assembled by Newbler, the sequences that were assembled by CAP3 and all singletons that were not assembled by either were used in subsequent analyses.

Illumina data were assembled with Velvet[34] (version 1.0.12) and Oases (version 0.1.15). Insert lengths for Oases were estimated with a 2100 Bioanalyzer (Agilent). Reads that did not have an average quality score of at least 35 were removed. We examined the assemblies over a range of *k* values (21–61, in increments of 10). We selected a *k* value of 61 for all samples, except for *Octopus* (for which we used 31). As for the 454 assemblies, we selected a single splice variant (transcript in Oases terminology) for each gene (locus in Oases terminology). To accomplish this, we developed a procedure whereby we chose transcripts that were at least 150 nucleotides, had a length of at least 85% of the longest transcript for the gene and had the highest read coverage. We ignored loci that had more than 50 transcripts, as these often appeared to be the result of misassembly.

Assembled data were compared to NCBI's nr protein database with BLASTX, with an *e* cutoff of 0.00001. Large data sets were compared to a reduced nr database by masking nr sequences from taxa that do not belong to the clade designated by NCBI Taxon ID 33154 (Fungi/Metazoa group). Nucleotide sequences were translated with a version of the prot4EST (version 2.3)[35] pipeline that had been modified to run without user intervention, using these BLASTX results.

**Orthology assignment.** The orthology assessment for data set assemblies followed one described previously[16]. All-by-all comparisons were conducted with BLASTP as in ref. 17. Clustering analyses were conducted on these results by using MCL[36]. At the suggestion of recent analyses[37], we excluded edges with $-\log_{10}$ BLASTP *e* values lower than 20, to reduce spurious cluster connections. We examined cluster composition with inflation parameters between 1.1 and 6 and found that the final cluster composition was not particularly sensitive to different inflation values in this range. We selected an inflation value of 2.1. Clusters with at least four taxa and at least one ingroup taxon were aligned by using MAFFT[38] and trimmed with Gblocks[39], and maximum likelihood analyses were conducted with RAxML[40]. The assessment of these phylogenies was conducted as in ref. 16. Monophyly masking was conducted to reduce the number of monophyletic sequences from the same taxon to one sequence. The resultant phylogenies were then analysed by an iterative paralogy pruning procedure, by which maximally inclusive subtrees with no more than one sequence per taxon were pruned and retained. FASTA-formatted files were generated from subtrees that were produced by the paralogy pruning procedure. These files were then aligned with MAFFT, trimmed with Gblocks, filtered (alignments with fewer than 150 sites were excluded) and concatenated into the final matrices.

**Phylogenetic analyses.** We constructed two phylogenetic matrices from the translated sequences. The 'small' matrix consists of 301 genes that are present in at least 20 taxa. It has 50% gene occupancy and is 50,930 sites long. The 'big matrix' consists of 1,185 genes that are present in at least 15 taxa. It has 40% occupancy and 216,402 sites. Both matrices contain data for all of the 46 species included in the study

Maximum likelihood analyses were performed for both matrices by using RAxML (version 7.2.6)[40] with both the Le and Gascuel (LG)[41] and WAG[42] models with each gene region partitioned. Likelihood analyses consisted of first conducting a bootstrap analysis with 200 replicates, which was followed by a thorough maximum likelihood search.

Bayesian analyses of the small matrix were conducted with MrBayes (version 3.1.2)[43] and PhyloBayes (version 3.3b)[44,45]. The big matrix was too large to analyse with these tools. With MrBayes, we conducted two searches each with two runs (four runs and 16 chains total). We allowed MrBayes to estimate the fixed rate model of evolution. Each chain was run for 1,000,000 generations, and convergence was determined with time-series plots and an estimated sample size of tree likelihoods of at least 100. Samples recorded before burn-in were removed, and post-burn-in samples of the runs were combined. We summarized the posterior probabilities of the clades with majority rule consensus trees.

We conducted analyses of the reduced-outgroup small matrix with PhyloBayes (version 3.3b) using the CAT model of evolution[45]. Five PhyloBayes runs under the fully parameterized CAT model were run, and each converged by 2,000 cycles based on time-series plots of the likelihood scores and number of partitions. The runs were allowed to run, each for more than 3,500 cycles, and estimated about 300 categories for the model. We removed pre-burn-in samples and constructed a majority rule consensus tree using all five runs (Supplementary Fig. 9)

31. Wilson, N. G. *et al.* Field collection of *Laevipilina hyalina* McLean, 1979 from southern California, the most accessible living monoplacophoran. *J. Molluscan Stud.* **75,** 195–197 (2009).
32. Ewen-Campen, B. *et al.* The maternal and early embryonic transcriptome of the milkweed bug *Oncopeltus fasciatus*. *BMC Genomics* **12,** 61 (2011).
33. Parkinson, J. *et al.* PartiGene: constructing partial genomes. *Bioinformatics* **20,** 1398–1404 (2004).
34. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
35. Wasmuth, J. D. & Blaxter, M. L. prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics* **5,** 187 (2004).
36. van Dongen, S. *A Cluster Algorithm for Graphs* Technical Report No. INS-R0010 (National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, 2000).
37. Apeltsin, L., Morris, J. H., Babbitt, P. C. & Ferrin, T. E. Improving the quality of protein similarity network clustering algorithms using the network edge weight distribution. *Bioinformatics* **27,** 326–333 (2011).
38. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinformatics* **9,** 286–298 (2008).
39. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17,** 540–552 (2000).
40. Stamatakis, A. P. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
41. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25,** 1307–1320 (2008).

42. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18,** 691–699 (2001).
43. Huelsenbeck, J., Larget, B., van der Mark, P., Ronquist, F. & Simon, D. *MrBayes: Bayesian Analysis of Phylogeny* 〈http://mrbayes.csit.fsu.edu/index.php〉 (2005).
44. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25,** 2286–2288 (2009).
45. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21,** 1095–1109 (2004).

## CORRIGENDUM

doi:10.1038/nature11736

# Corrigendum: Resolving the evolutionary relationships of molluscs with phylogenomic tools

Stephen A. Smith, Nerida G. Wilson, Freya E. Goetz, Caitlin Feehery, Sónia C. S. Andrade, Greg W. Rouse, Gonzalo Giribet & Casey W. Dunn

*Nature* **480,** 364–367 (2011); doi:10.1038/nature10526

In this Letter, we investigated the evolutionary relationships of molluscs with multigene data sets assembled from new transcriptome data and published genomes and transcriptomes. Since publishing these results, examination of our gene sequence matrix by others revealed that all instances of six amino acids (E, F, I, L, P and Q) were replaced by ambiguous characters in our super matrix. This led to the exclusion of data that should have been in the final analyses. The data exclusion was caused by incorrect handling of protein data at the final stage of matrix concatenation by the published program Phyutility (http://code.google.com/p/phyutility/). We have fixed the program, regenerated the final matrices, and re-run our analyses. There was minimal impact on our results, with no changes in the topology of the tree at deep nodes that had consistent strong support in our published analyses. There are minor variations in support values in the corrected analyses. The corrected matrices have been deposited at Dryad under the existing accession number (http://dx.doi.org/10.5061/dryad.24cb8). Figure 1 shows the corrected Fig. 2 of the original Letter, with corrected support values. In the text of the original Letter, the sentence "Bayesian analyses using the site-heterogeneous CAT model of protein evolution also place Scaphopoda as the sister group to Gastropoda, with a posterior probability of 89%" should read "Bayesian analyses using the site-heterogeneous CAT model of protein evolution also place Scaphopoda as the sister group to Bivalvia, with a posterior probability of 81%". In the Methods Summary, both instances of "27%" in the following phrase should be "41%": "27% character occupancy (that is, 27% of the matrix consists of unambiguous amino acid data, with the remainder being missing data or alignment gaps)" and "21%" should be "32%" in the sentence "This matrix has 40% gene occupancy, 21% character occupancy and is 216,402 sites long.". In the Methods, the following two sentences should be removed: "PhyloBayes misidentified the data type of our matrix as DNA, resulting in model misspecification and lack of convergence. We conducted the analyses presented here with a modified version that was forced to read all matrices as protein sequences." The next three sentences from the final paragraph of the Methods should now read "Five PhyloBayes runs under the fully parameterized CAT model were run, and each converged by 2,000 cycles based on time series plots of the likelihood scores and number of partitions. The runs were allowed to run, each for more than 3,500 cycles, and estimated about 300 categories for the model." instead of "Five PhyloBayes runs under the fully parameterized CAT model each converged at around 1,500 cycles (at least 86,000 generations) based on time-series plots of the likelihood scores and number of partitions. The runs were allowed to run for 5,000 cycles for two runs and 2,500 cycles for three runs. The runs estimated 140 (±10) categories for the model." The original Letter's Supplementary Figures 2–9 have also been updated with the results of analyses based on the corrected matrices. Differences in results for these figures include increased support for some relationships highlighted in the original manuscript, and changes in some relationships within Bivalvia. These changes do not alter any of the conclusions of our manuscript. We thank Hervé Philippe, Raphael Poujol and Béatrice Roure for bringing this error to our attention.

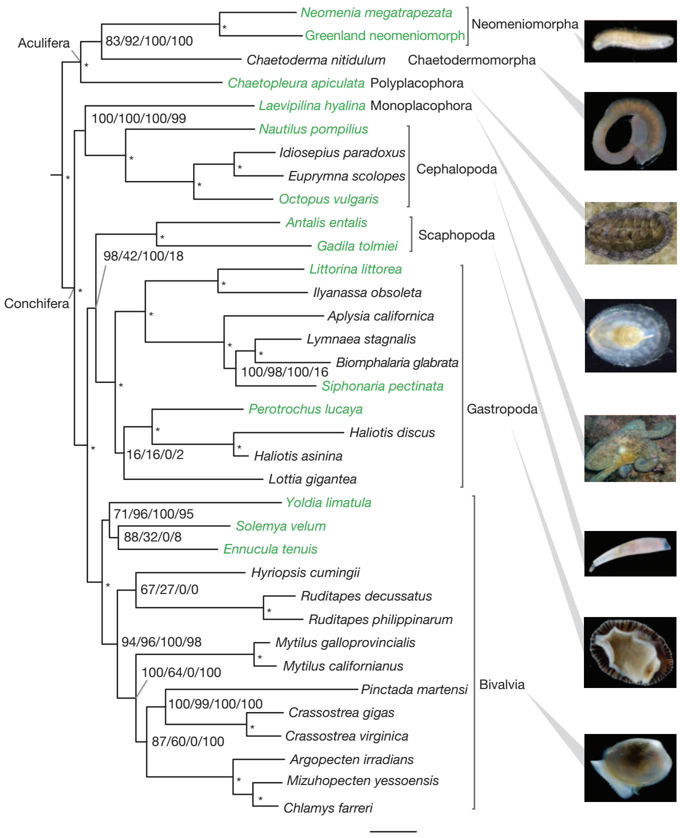

**Figure 1 | This is the corrected Fig. 2 of the original Letter.**