Review

Improving animal phylogenies with genomic data

Maximilian J. Telford[1] and Richard R. Copley[2]

[1] Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK
[2] Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK

Since the first animal genomes were completely sequenced ten years ago, evolutionary biologists have attempted to use the encoded information to reconstruct different aspects of the earliest stages of animal evolution. One of the most important uses of genome sequences is to understand relationships between animal phyla. Despite the wealth of data available, ranging from primary sequence data to gene and genome structures, our lack of understanding of the modes of evolution of genomic characters means that using these data is fraught with potential difficulties, leading to errors in phylogeny reconstruction. Improved understanding of how different character types evolve, the use of this knowledge to develop more accurate models of evolution, and denser taxonomic sampling, are now minimizing the sources of error. The wealth of genomic data now being produced promises that a well-resolved tree of the animal phyla will be available in the near future.

Genomes as a repository of evolutionary information

The major groups of animals – the bilaterian phyla – appear suddenly in the fossil record ~530 million years ago in what is known as the Cambrian explosion. Although the suddenness of their appearance could be explained in part by the limitations of the fossil record, the cladogenesis and morphological innovation that produced these diverse body plans is likely, nevertheless, to have taken place during a relatively short period of time, very long ago [1]. The apparent brevity and great antiquity of the Cambrian explosion has led some to argue [2,3] that there is little hope of unpicking the genomic signal that might remain from this explosion of evolutionary innovation – the evolutionary equivalent of the 'cosmic background radiation' left over from the Big Bang.

But is it naïve to suppose that complete genome sequences might provide a complete and final picture of animal relationships, a solution to a problem that arose long before Darwin? Logically, genomes must be replete with evolutionary information, representing a more or less interpretable historical record of genotypic and phenotypic change.

Crucially, gaining an understanding how genomes have evolved is not possible via the study of individual species; understanding depends, instead, on comparisons of genomes (and the phenotypes they encode) which represent the endpoints of different branches of evolution. To make meaningful evolutionary comparisons of genomes they must be interpreted in the light of a phylogeny relating the organisms in which they reside [4].

Considering this requirement for phylogenetic context it is perhaps not surprising that one of the greatest successes of comparative genomics to date has been the use of genomic data to aid the reconstruction of evolutionary relationships. But, as we will see, although the idea of using genome-sized datasets to reconstruct animal evolution is immensely attractive, it has also proved to be far from straightforward. The problem is that the ancient evolutionary information encoded in the genome is overwritten

Glossary

Bilateria: bilaterally symmetrical animals; synonymous with triploblasts because they possess three tissue layers – ectoderm, mesoderm and endoderm. Bilateria contain deuterostomes and protostomes and exclude poriferans (sponges), cnidarians (jellyfish and corals), placozoans (*Trichoplax*) and ctenophores (sea gooseberries).
Cambrian explosion: the sudden appearance of many of the extant bilaterian animal phyla in the fossil record 530–540 million years ago.
Cladogenesis: a speciation event resulting in two monophyletic groups or clades of organisms.
Deuterostomia: a major division of Bilateria; contains the phyla Chordata, Echinodermata, Hemichordata and Xenacoelomorpha.
Ecdysozoa: a major division of Protostomia; ecdysozoan phyla include Arthropoda, Nematoda and Priapulida.
Homoplasy: convergent evolution of a given character state in unrelated lineages. Homoplasy leads to the incorrect grouping of unrelated species on phylogenetic trees.
Long-branch attraction (LBA): LBA is an error of tree reconstruction resulting from undetected convergent evolution (homoplasy) in unrelated branches of phylogenetic trees. These homoplasies are more likely to occur along long branches, leading to the artefactual clustering of long-branch taxa. LBA could arise from rapid evolution in a subset of species, or because one or more species are particularly evolutionarily distant from others, or from a combination (e.g. rapidly evolving nematodes are attracted to the phylogenetically distant fungi). LBA affects all types of tree-reconstruction methods to some degree, but probabilistic models are generally better able to infer the existence of convergent changes.
Lophotrochozoa: a major division of Protostomia; lophotrochozoan phyla include Mollusca, Annelida and Platyhelminthes.
Phylogeny: an evolutionary tree of relationships, also called a phylogenetic tree.
Plesiomorphies: primitive or ancestral character states. These can be shared (symplesiomorphies) by a subset of species of interest but, being primitive, will also be present in more distantly related taxa. Symplesiomorphies cannot be used to infer close relationships between the species that share them.
Protostomia: a major division of Bilateria; contains Lophotrochozoa and Ecdysozoa.
Synapomorphies: derived or novel character states shared by two or more taxa – these then constitute a monophyletic group (clade) whose common ancestor also possessed the character state. The character will not be present in any outgroups to the clade and therefore is unique to and defines the clade.

Corresponding author: Telford, M.J. (m.telford@ucl.ac.uk).

0168-9525/$ – see front matter doi:10.1016/j.tig.2011.02.003

by the messy course of subsequent change and, just as phylogenies based on morphology can be dogged by convergence and secondary absence of characters, so too can those derived from genomes.

The dawn of animal phylogenomics: the Ecdysozoa versus Coelomata dispute

For a phylogenetic tree to contain non-trivial information it must connect a minimum of three species plus an outgroup to root the tree. For animal groups this minimum condition was fulfilled 10 years ago with the complete sequencing of genomes of the model organisms *Drosophila melanogaster*, *Caenorhabditis elegans* and of humans [5–7]. Although a phylogenetic problem concerning just three species (and with only three possible solutions) could appear trivial, this particular sample of genomes was a phylogeneticist's dream because, for the first time, it allowed the testing with genome-scale data of one of the most controversial claims of the 'new animal phylogeny' – the existence of a clade termed the Ecdysozoa. The Ecdysozoa are named after the common occurrence of ecdysis – moulting of the cuticle for growth. The clade includes the insects and nematodes and excludes non-moulting animals including annelids and molluscs as well as vertebrates such as humans [8,9]. The Ecdysozoa hypothesis (*C elegans*, *D. melanogaster*) contradicts the traditional view that unites apparently more complex animals, in possession of an epithelium-lined body cavity or coelom (Human, *D. melanogaster*), to the exclusion of more simple pseudocoelomate animals including nematodes.

The apparent simplicity of this question, and the abundant means for its resolution in the form of genome-scale data, set the scene for a debate lasting a decade. The inability to agree on a solution seems surprising, but the basis for the disagreement is illustrative of the root of the most recalcitrant of phylogenetic problems – the diverse methods used to solve this problem indicate future directions for the use of genome-scale data in phylogenetics.

Long-branch errors and their solution

The first evidence for the Ecdysozoa hypothesis came from authors who recognized the unusually rapid evolution of the available nematode small subunit (SSU) ribosomal RNA sequences meant there was a strong likelihood of a

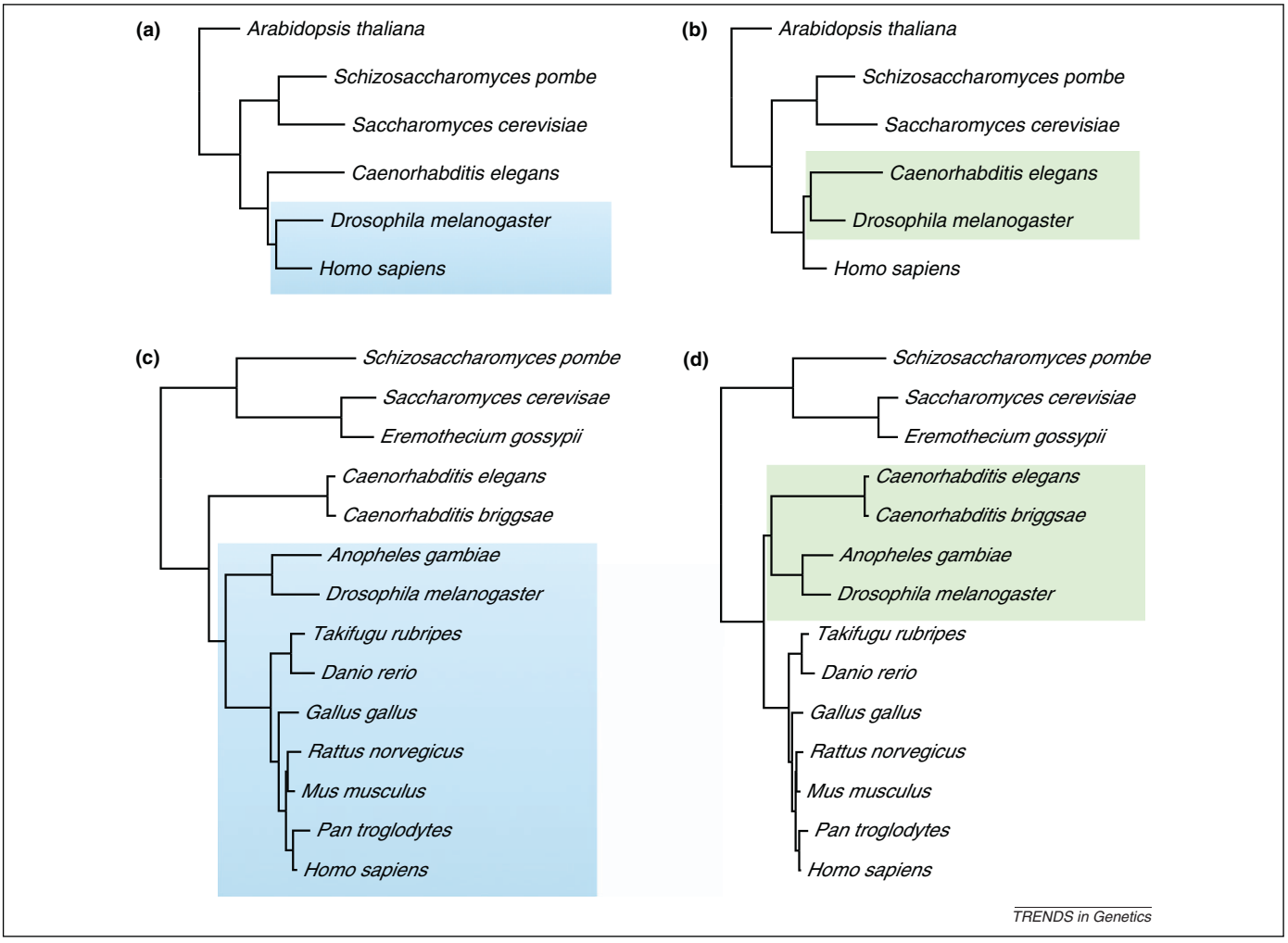

Figure 1. More realistic models of molecular evolution have a major impact on tree reconstruction. **(a)** Analysis of phylogenomic data from [13] (29 625 characters) using the original site homogenous model (JTT matrix, gamma parameter with five discrete categories) shows support for monophyletic Coelomata (blue box). **(b)** Analysis of the same data using a site-heterogeneous model (CAT–GTR + Γ) shows support for Ecdysozoa (green box). **(c)** Analysis of data from [22] (8089 characters) using site homogenous model (as above) shows support for Coelomata (blue box). **(d)** Analysis using the CAT model shows support for Ecdysozoa (green box). All analyses were performed using Phylobayes; all nodes have a Bayesian posterior probability of 1.0. CAT, categories model; GTR, general time-reversible; JTT, Jones, Taylor and Thornton model.

systematic error in tree reconstruction known as long-branch attraction (LBA) [9]. LBA is a well-understood and pervasive source of systematic error in phylogeny reconstruction that causes unrelated long branches to cluster together [10]. Long branches could occur through elevated rates of evolution in specific taxa or might simply be because no close relatives have been sampled for particular species on the tree. In general, systematic errors are generated when the genes/genomes of different taxa have experienced divergent trends in their evolution (in the case of LBA, different rates of substitution) that are not adequately accounted for by the models of evolution employed [11]. The recognition of the rapid evolution of the available nematodes led to a search for nematodes with more typical rates of evolution [9]. This effort resulted in a change from a tree in which the long-branched nematodes diverged close to the distant root (leaving the coelomate animals as a clade – Coelomata) to a tree in which the more moderately evolving nematodes branched with the arthropods forming the Ecdysozoa, thus disbanding the coelomate clade [9].

The first phylogenomic analyses of the Ecdysozoa/Coelomata question were for the most part restricted to a single and notably long-branched nematode – *C. elegans* [12,13] (Figure 1). Although measures were taken in these studies to address the problems of LBA, they nevertheless overwhelmingly supported Coelomata, with flies and humans sharing a more recent common ancestor with each other than with nematodes [12,13]. Whereas the taxonomic coverage was much lower than for the work based on SSU, these studies were given serious credence in view of their overwhelming superiority in terms of overall alignment length. The much smaller SSU studies supporting Ecdysozoa were widely dismissed as resulting from stochastic error. Hindsight tells us that, although stochastic (small sample) error had certainly been removed in these phylogenomic studies, systematic error (whose effects are felt with greater certainty with larger datasets) had been emphasized – in particular by the use of the idiosyncratic *Caenorhabditis* data [11,14–16].

Solving LBA errors in phylogenomic datasets

In the original SSU-based study supporting Ecdysozoa, LBA had been tackled by sampling additional nematodes and, in particular, slowly evolving exemplars. Equivalent methods employed with phylogenomic datasets – using less phylogenetically distant (and thus shorter-branched) outgroup taxa to root the tree [16,17], increasing the density of taxon sampling within the nematodes [11,16,18], and using the more slowly evolving priapulids in place of nematodes [19] – have similarly led to strong support for Ecdysozoa. In addition to adding new taxa, the influence of parameter-rich probabilistic models of evolution, designed explicitly to model the realities of sequence evolution, should not be underestimated. Including additional biologically relevant parameters in probabilistic models inevitably results in a closer fit to the data, and estimating additional parameters is generally well-tolerated with the very large datasets available under phylogenomic studies. The effects of using one of the most realistic models currently available (CAT–GTR + Γ mixtures model from the Phylobayes software [20,21]) can be seen in Figure 1. Here two venerable datasets [13,22] that strongly support Coelomata under their original methods of analysis, now switch their allegiance to strongly supporting Ecdysozoa simply by the use of a better-fitting model [20,21].

Presence/absence data and the violation of Dollo's law

As a complement to these whole-genome-based studies that use typical analyses of aligned gene sequences, several authors have used an alternative approach involving consideration of data that can be broadly classed as presence/absence characters [23,24]. The principle is simply to look for common, complex heritable features of the genomes under consideration (genes, specific combinations of

Box 1. Examples of rare genomic changes from gene and genome structure

Introns

Although intron locations in isolated orthologous genes have previously been used for phylogenetic inference (e.g. [66]), the arrival of complete genome sequences provides the means for this method to be approached in a more rigorous manner. Although conceptually straightforward, early analyses focused on resolving the Coelomata versus Ecdysozoa controversy produced conflicting results depending on outgroup choice and the methods for deciding which introns were informative. The sequencing of the *Nematostella vectensis* genome helped to clarify which introns are likely to be plesiomorphies of the bilaterians, and strongly supports large-scale correlated loss of introns within the ecdysozoans [59]. Intron dynamics also support the Olfactores hypothesis, despite the fact that tunicates have undergone substantial intron loss [67].

In general, on a genome-wide scale, intron presence/absence is not necessarily a well-conserved trait. Genomes can undergo substantial turnover of introns – in *Oikopleura dioica,* for instance, 76% of intron locations are not shared with other animals [68] (in comparison, human and Amphioxus share 85% of introns in alignable regions). Such data need to be interpreted with caution, and care must be taken to distinguish primitive from secondary absence.

Unique gene structures

Related to intron locations, some genes have unusual structures that can be treated as synapomorphies. One such useful contribution arises from consideration of vertebrate immunoglobulin loci, which have tandem cassettes of V, D and J gene segments. Hagfish and lampreys, in contrast, share an unusual immune receptor that uses tandem cassettes of leucine-rich repeats to generate diversity – this has been interpreted as providing support for grouping these jawless vertebrates (Cyclostomata) [36].

Protein domain structure

Kawashima *et al.* identified 1000 new domain pairs found uniquely within the vertebrate lineage [69], although their study tested no phylogenetic hypotheses. The existence of a vertebrate clade is obviously uncontested, but many unique domain combinations can be shared by non-monophyletic clades – for instance by Lophotrochozoa plus Ambulacraria (unpublished observations). Under the assumption that domain-fusion events are rare, and multiple independent occurrences are therefore unlikely, this suggests extensive secondary loss of ancestral domain combinations. Such losses are unlikely to occur randomly, but instead are concentrated in particular taxa, thus confounding the use of shared presence as a synapomorphy. The alternative, that domain-fusion events are common, has been considered unlikely [70].

Box 2. Lineage-specific genes and gene families as markers of relationships

Gene families can serve as clade synapomorphies. Examples include the ANTP class homeobox genes found within the Metazoa, but in no other eukaryotic phyla, and the entire class of four-helix cytokine-like proteins (including interferons and various hormones) within the vertebrates, but not other Metazoa. This class of synapomorphy is seductive because it appeals to the intuition that specific genes define particular types of clade-specific biology: in the case of the Hox genes, patterning of the metazoan body plan; in the case of the cytokines, the vertebrate adaptive immune system. As with other traits, however, the phylogenetic utility of the presence/absence of such gene classes is affected by poor taxon sampling and by secondary absence. Nematodes, for instance, have lost the entire class of NFκB-like transcription factors – before the identification of these genes in *Nematostella vectensis* and other outgroups this fact would have appeared to support the Coelomata hypothesis.

The developmental 'toolkit' of animal genes, absent from the first sequenced non-animal outgroup genomes, including yeasts and the choanoflagellate *Monosiga brevicollis* [71], were thus apparent metazoan synapomorphies. Recently, however, examples such as the T-box and RUNT domain have been found in the single-celled *Capsaspora owczarzaki* [72]. This case is instructive because it highlights the fact that gene families are often found to have arisen before there was any necessity for the functions with which they are typically associated, and that the chosen outgroup for a particular clade (in this case *Monosiga brevicollis*) is not necessarily a good proxy for the genome of the last common ancestor of a group, having itself undergone secondary loss.

In preference to gene families, one would ideally identify precisely the individual genes that constitute clade-specific synapomorphies. Although the principles are clear, namely reconstructing a phylogeny and identifying clade-specific gene-duplication events, the challenges of performing these analyses robustly on a genome-wide scale are formidable. Gene duplications are likely to lead to functional shifts (and hence accelerated evolution) that are unevenly distributed over the paralogous copies, leading to difficulties in inferring when the duplication event occurred. The problem of secondary absence of duplicated genes is particularly acute because duplicated genes are likely to have some level of functional redundancy and are therefore at elevated risk of being lost.

protein domains, introns, indels and rarely changing amino acids [13,14,25–28]) and to score each homologous character as being present or absent in each genome (Boxes 1,2). The characters are considered sufficiently complex to be unlikely to have evolved convergently, and common occurrence is used to indicate close relationship. For the Ecdysozoa/Coelomata problem the question is simply whether fruit flies share more of these derived characters (those absent in the outgroup) with humans or with nematodes. These presence/absence data have been analyzed using Dollo parsimony, which allows a single instance of evolution of a character but permits parallel losses in multiple lineages (loss of a character is easier than gain) [29,30]. Dollo parsimony should not be confused with 'Dollo's law of irreversibility' which states that evolution cannot proceed in reverse such that any character returns exactly to its ancestral character state [31].

Just as with the initial sequence-based analyses, these studies were restricted to the few taxa with a complete genome (a character can only reliably be scored as absent if the complete genome is known), and again the studies overwhelmingly supported Coelomata [13,25–28]. Dollo parsimony is a sensible approach for this class of characters, but nevertheless relies on the assumption that losses of characters will be randomly distributed among taxa. The inevitable suspicion, considering the (with hindsight) inappropriate support for Coelomata, is that this model of randomly distributed character loss is violated.

A greatly enhanced tendency for loss of characters in *Caenorhabditis* compared to humans or flies has indeed been demonstrated [14], and this systematic bias in *Caenorhabditis* results in an LBA effect exactly as encountered with the standard sequence-based approach. Methods to address LBA (more closely related outgroup and more densely sampled ingroup, especially additional nematodes) have been applied to several of these datasets and, in each case, these efforts to address LBA switch support from Coelomata to Ecdysozoa [32–35].

Problems with presence/absence data arise because their evolution violates Dollo's law of irreversibility. In instances of loss of a character, and *contra* Dollo, it is not possible to differentiate between (i) the case in which a taxon lacks a character because the taxon diverged before the character evolved (primitive absence) and (ii) the case in which a taxon lacks a character due to loss in a subset of a clade whose ancestor possessed the character (derived or secondary absence). In essence, although the derived state might be complex and resistant to convergent evolution, the primitive state (absence) is a very simple character and is easily re-evolved. This inequality becomes a problem in cases where differences in propensity for loss result in systematic error [14].

MicroRNAs as phylogenetic indicators

One possible solution is to identify characters that, once gained, are unlikely to be lost. This mode of evolution has been expected to hold for microRNAs (miRNAS) which, because they are generally thought to interact with multiple target genes, are expected to be resistant to loss due to the deleterious pleiotropic effects that would result. miRNAs have indeed had many notable successes and have given invaluable support to a number controversial animal groups including the Cyclostomata (lampreys and hagfish) [36], the Mandibulata (insects, crustaceans and myriapods) [37], the Olfactores (urochordates and vertebrates) [38], and the Deuterostomia [39]. Although they have an impressive record, the immunity of miRNAs to loss turns out not to be universally true. It was recently demonstrated, for example, that numerous miRNAs have been lost from (or at least are undetectable in) the genome of the supposedly primitive acoel flatworm *Symsagittifera roscoffensis* [39]. This type of systematic bias in propensity for loss is exactly that which can lead to wrong answers with presence/absence characters. In the case of the acoels, the paucity of miRNAs led to incorrect support for a position close to the outgroup and outside of the main clade of Bilateria [39].

Characters that conform to Dollo: the example of the NAD5 gene

A consequence of these considerations is the realization that complex characters that do not violate Dollo's law of

Box 3. The *NAD5* gene as an ideal synapomorphy of protostomes

The usefulness of *NAD5* was first pointed out in a study considering the phylogenetic position of chaetognaths using whole mitochondrial genome sequences [41]. These authors noted clustered series of amino acids in the NAD5 protein sequence that were highly conserved across many eukaryotic groups including sponges, cnidarians, ctenophores and deuterostomes. They showed that these same amino acids were different, but equally conserved, in all protostomes they looked at. In addition to these 'conserved-but-different' amino acids, there is both a deletion of a single (otherwise conserved) amino acid in all protostomes and a greatly shortened N-terminus in protostomes when compared to all other eukaryotes. In essence, the *NAD5* gene conforms perfectly to our ideal character with two complex character states, one primitive and one derived. It is highly unlikely that the diverse animals that possess the derived condition evolved it convergently, but it is equally unlikely that those with the 'primitive' character state evolved it by reverse evolution from the 'derived' state.

The protostome *NAD5* character defines just a single clade on the metazoan tree, but is nevertheless particularly interesting because the existence of this clade serves to rule out a series of long-disputed phylogenetic relationships:

(i) Arthropods and nematodes are both protostomes, and are therefore more closely related to each other than either is to the chordates (thus ruling out the Coelomata hypothesis) [13,25–28,32–35].

(ii) Chaetognaths are protostomes (and therefore not deuterostomes) [41,73–75].

(iii) The lophophorate phyla (brachiopods, bryozoans and phoronids) are protostomes (and therefore not deuterostomes) [76–79].

(iv) The acoel flatworms are not protostomes (and hence not the sistergroup of the platyhelminthes) [39,80,81].

irreversibility would be robust indicators of phylogeny. The straightforward way to avoid this problem is to use characters in which the primitive state is as complex, and hence as resistant to convergent (re) evolution, as the derived state [40]. In essence, the character must be one that is always present (instead of being absent or present) but which exists in two equally complex character states – primitive and derived. One such character, the mitochondrial *NADH dehydrogenase subunit 5* (*NAD5*), clearly illustrates this approach. *NAD5* exists in two distinct but conserved states that differ by multiple substitutions and indels: one state is found in almost all eukaryotes studied, including cnidarians and deuterostomes, and represents the primitive character state for the animals; the other state is found uniquely in the protostomes and thus represents a shared derived character (synapomorphy) that unites the protostomes [41] (Box 3).

The successes of phylogenomics

The prolonged debate regarding Ecdysozoa versus Coelomata has certainly highlighted many of the pitfalls associated with the use of such large datasets and has contributed in great part to the first and arguably most important contribution of phylogenomics studies – the support they have given to the general view of animal phylum level relationships derived from studies of SSU rRNA: the so-called 'new animal phylogeny' [8]. Recently, the more influential of such phylogenomic studies have gone to great lengths to sample the broadest possible range of taxa with a very large selection of orthologous genes [42–44]. Despite the controversy we have seen over the position of the nematodes, this subsequent work has confirmed the monophyly of the Ecdysozoa, Lophotrochozoa, Protostomia and Deuterostomia and identified their constituent taxa.

When molecules and morphology clash

Most higher-level groupings of animals have been fairly intensively sampled; why then have we not already reached the end of metazoan high-level phylogenetics (notable remaining controversies are illustrated in Figures 2,3)? We suggest that part of the explanation lies in the inability of large-scale holistic studies to address adequately the possibility of systematic artefacts affecting specific parts of the tree. The second important contribution made by recent genomic studies, therefore, has emerged from more in-depth research into specific sets of relationships. These questions have been targeted either because existing trees make little sense in terms of the characteristics of the groups (they conflict with the known distribution of presumed homologous morphological characters) or, more generally, because different datasets have suggested differing conclusions. These more targeted studies highlight the importance of stress-testing contentious aspects of these large phylogenies.

A classic example of a clash between molecules and morphology is found when considering the relationships of the myriapods (millipedes and centipedes). The head segments and in particular the mouthparts of myriapods strongly resemble those of insects, and to a lesser extent those of crustaceans; all three groups have therefore been traditionally linked in a clade named Mandibulata in recognition of the biting mandible that all three share as their third head appendage. Molecular analyses, including large phylogenomic studies, however, indicated that the closest relatives of the myriapods were the chelicerates (arachnids and horseshoe crabs) [45–48]. This separation of myriapods from insects and crustaceans implied a surprising (albeit not unprecedented [49]) degree of convergent evolution between Myriapods and the crustacean/insect clade [50–52]. Only work using increased sampling, careful outgroup selection, assessment of the best-fitting models of sequence evolution, and novel miRNA characters was able to overcome the apparent effects of an LBA artefact made worse by the short branch leading to the Mandibulata [37,50,53,54]. With this somewhat labor-intensive approach to the problem it was possible to show the Mandibulata is a credible clade, reconciling morphology with the results from molecular analyses [37].

Systematic errors from both molecules and morphology

The Mandibulata clade provides a clear example where discordance between molecular and morphological characters has tested the molecular phylogeny. Occasionally however, both molecules and morphology have been found to mislead in the same way. One important example is the discovery that the closest relatives of the vertebrates are not the fish-like cephalochordates but the unusual and fast evolving urochordates [55]. The urochordates (e.g. sea squirts) have a typical chordate tadpole larva but an unusual sac-like, sessile and filter-feeding adult stage. Molecules and morphology conspired to mislead because

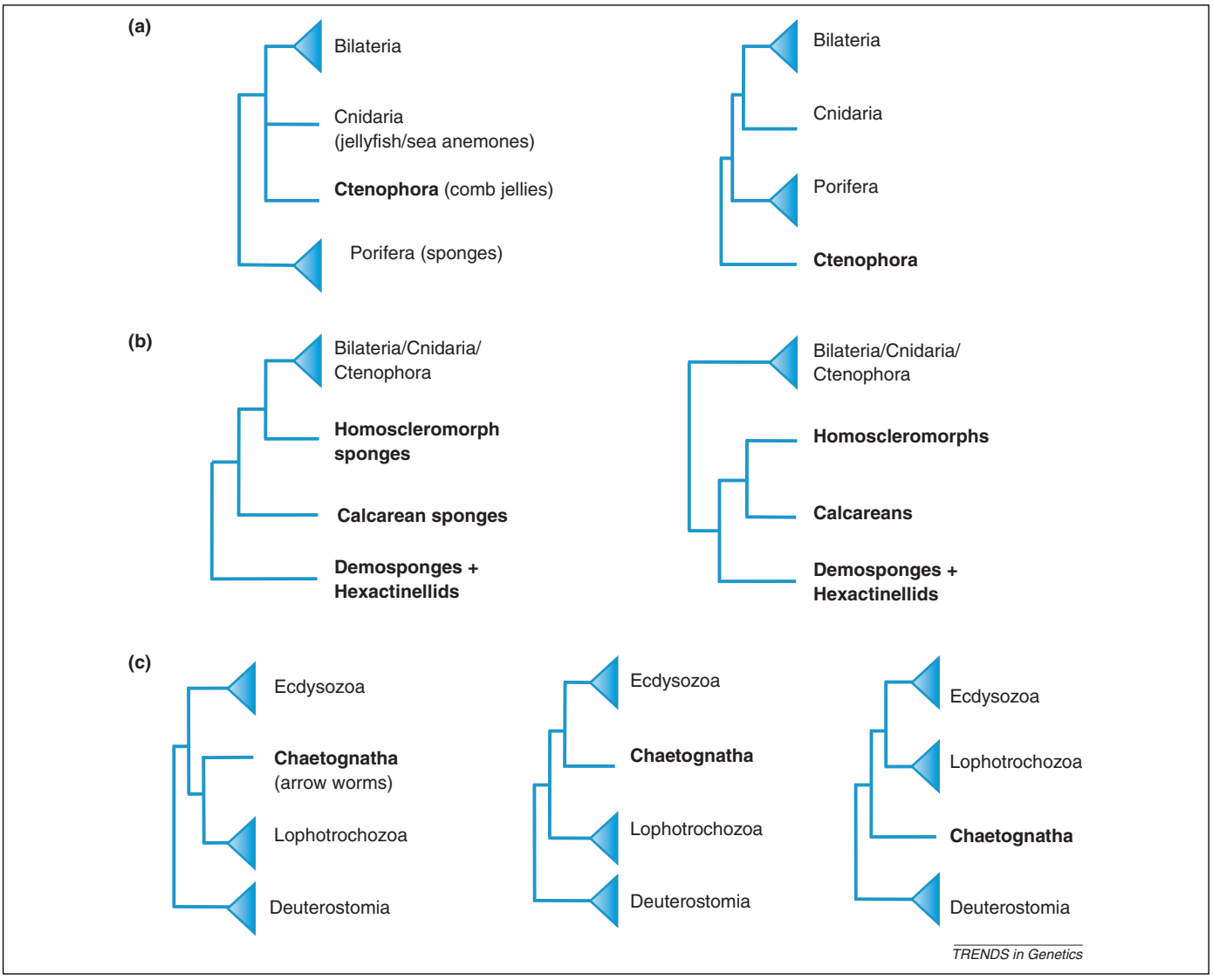

Figure 2. Notable open questions in the higher-level relationships of the Metazoa. **(a)** Position of the Ctenophora (comb jellies) relative to the Cnidaria, Porifera and Bilateria. Some analyses support the traditional view in which the Porifera are basal to other groups [15], others place ctenophores as most basal branch of Metazoa [43,61,62]. **(b)** Origins of Porifera (sponges). Different studies have supported multiple lineages (paraphyletic) [82], whereas others favor a single origin (monophyletic) [15]. **(c)** Chaetognaths (arrow worms) are protostomes [41] but it is not clear whether they belong with Lophotrochozoa, or with Ecdysozoa, or are outside both of these clades [73].

both the high substitution rate (and, it turns out, secondary absence of various genes and violent reorganization of the genome as a whole) and the loss of typical chordate features in the adult pointed to the urochordates as being an early branch in chordate evolution. Dense species sampling was the key to revealing the unexpected truth of a close relationship between urochordates and vertebrates (Olfactores) that has since been corroborated by additional evidence [56–59].

The benefits of widespread genome and transcriptome sequencing

Much progress has been made in the past 20 years in resolving the relationships between all known metazoan phyla and the influence of genome-scale approaches is increasingly evident. Until recently, transcriptome sequencing was the principal route to providing massive alignments of orthologous genes for phylogenomics. The increasing taxonomically-broad reach of genome sequencing is improving matters further: by minimizing missing data [17], by making available a sample of genes unbiased by expression levels, and by providing large collections of non-gene-sequence based characters. There are now many parts of the metazoan phylogeny with broadly accepted relationships in which different clades have diverse sources of support; as can be seen from the unresolved bushy areas of the consensus animal tree (Figure 3), however, a number of outstanding areas of disagreement or uncertainty remain – most notably at the base of the tree and within the Lophotrochozoa.

When looking to the solution of the remaining questions at least one aspect appears clear: a mass of new data from completed genomes and transcriptomes will emerge in the near future in view of the decreasing cost and increased availability of next-generation sequencing. This deluge of data indicates, for sequence-alignment-based phylogenomic analyses, that we have essentially overcome the problem of stochastic error.

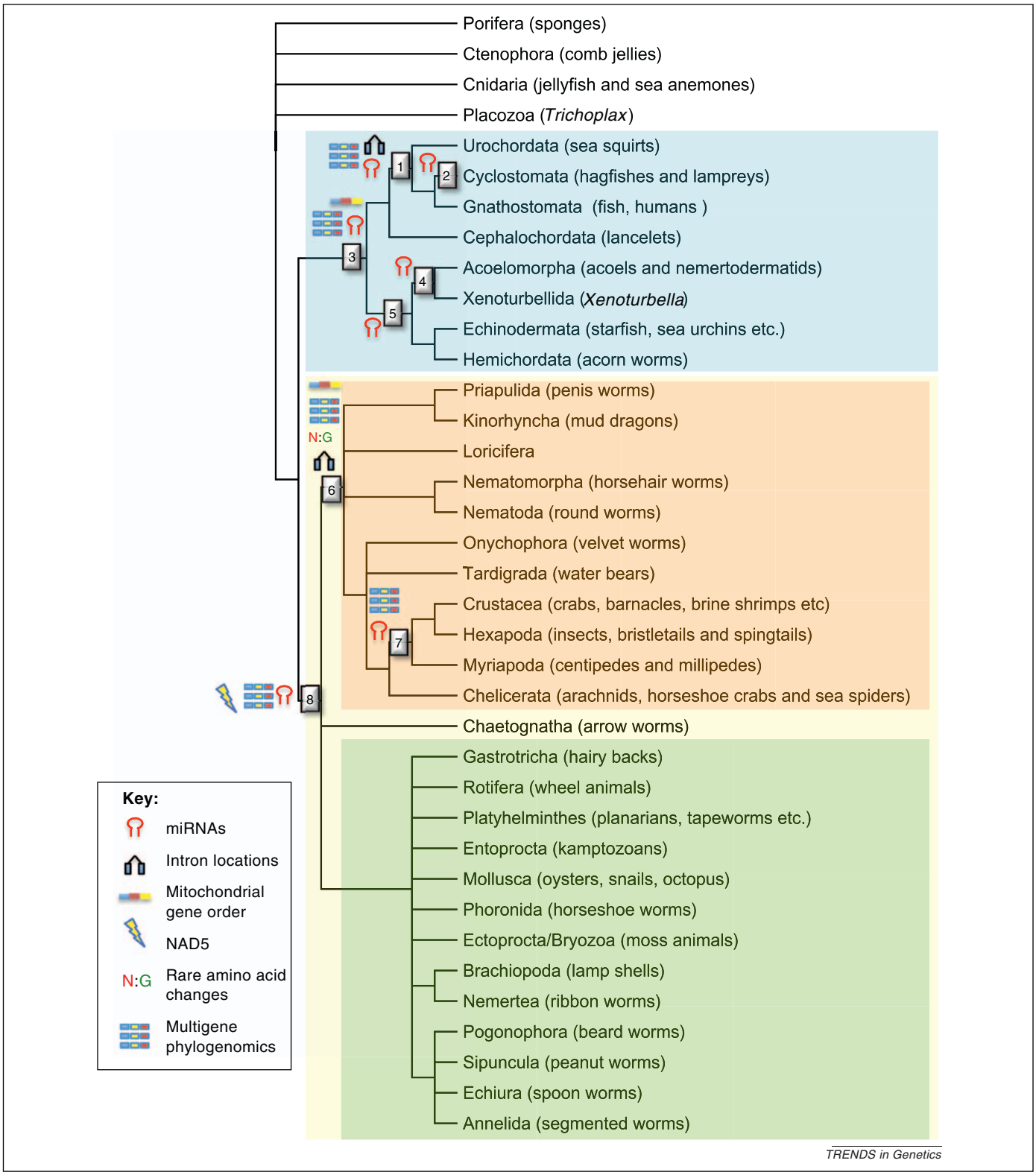

Figure 3. Consensus view of metazoan higher-level taxonomy highlighting the evidence for specific nodes discussed in the text. (1) Olfactores not Euchordata [38,55,59]. (2) Monophyletic Cyclostomata [36]. (3) Deuterostomia including Xenacoelomorpha [39,57]. (4) Monophyletic Xenacoelomorph [39] (5) Xenambulacraria [39,57]. (6) The Ecdysozoa [16,19,32,34,35]. (7) Mandibulata not Myriochelata [37,64]. (8) Protostomes including chaetognaths [41,73,75,83]. Note the lack of resolution within lophotrochozoan and ecdysozoan phyla. The Deuterostomia are boxed in blue; Protosomia in yellow. Within Protostomia, Ecdysozoa are boxed in orange and Lophotrochozoa in green. Symbols indicate sources of support for numbered clades as discussed in the text.

As some of the examples we have given above show, however, large datasets are not necessarily sufficient in themselves to minimize or eliminate errors in tree reconstruction. Systematic errors can in fact be accentuated by large datasets because the likelihood that they are cancelled out by stochastic error begins to disappear – they are said to be inconsistent. Moroever, it is axiomatic that the remaining areas of uncertainty in the metaozan tree are

precisely those difficult nodes (typically closely spaced) where systematic errors can most easily exert a major effect. In consequence, simply building big datasets is not sufficient to tackle these questions; further progress in using genome- and transcriptome-grade data for metazoan phylum-level phylogenetics will depend to a great extent on adopting approaches to minimize systematic error.

Broader taxonomic sampling

The first and most obvious improvement is to widen taxonomic sampling. Although it is difficult to be prescriptive about where this sampling should be done (because this depends on the interests of the community), the falling cost of genome sequencing gives grounds for optimism that even minor phyla will soon have fully sequenced representatives. However, to date a number of phyla have been barely sampled despite notable efforts to generate sequences from the more obscure, minute or difficult-to-collect groups of taxa [42,43].

The primary and obvious benefit of including new phyla in the analysis is simply that this is the only way in which they can be placed. More indirect but equally important benefits are achieved by reducing the effects of homoplasy. This bonus arises because insertion of additional taxa onto existing branches provides information about intermediate character states along those branches. If there is a convergent change in the branches leading to two taxa, tracking of intermediate states along the branches can reveal the independent evolution of the convergent changes, and homoplasy will not result in false support for grouping the two taxa. When dealing with LBA, adding intermediate species can be more simply thought of as a means to divide long branches. For some problematic groups, however, more intensive sampling seems unlikely to help: all extant species of some problematic phyla (chaetognaths [60] and ctenophores [61–63] for example) appear to derive from relatively recent radiations – meaning that essentially no intermediate (long-branch breaking) taxa are available for sampling.

Additional complete genomes simplify inference of loss

The advantages accruing from deeper sampling are also applicable to the seductive, but occasionally problematic, presence/absence data. As discussed, the major problem with such data is that secondary absence is generally indistinguishable from the ancestral (plesiomorphic) state: if secondary absences are sufficiently frequent and biased towards only a subset of the taxa under investigation (as is the case with introns in tunicates, compared to other chordates for instance [59]), inferred phylogenies could be incorrect. Sampling of sister taxa is able to reveal the previous existence of secondarily absent characters in the branch leading to the problematic species [14]. The usefulness of additional taxa has been clearly demonstrated in studies [32–35] of the presence/absence data that had initially supported Coelomata [13,25–27].

Can less sometimes be more?

Alongside deeper taxon sampling, we have seen the profound effects that a more accurate evolutionary model can have on the correct reconstruction of phylogenetic relationships. In a perfect world the ideal combination would be the largest possible alignment of orthologous genes analyzed with the best-fitting model possible, thus simultaneously minimizing stochastic and systematic error. In reality, however, the use of very large datasets in conjunction with the models most able to reflect biological reality imposes a toll on tractability in terms of computer capacity.

The most comprehensive analysis of metazoan phylum-level relationships to date involved a dataset of 98 taxa (including outgroups) and 270 580 aligned amino acids [43]. This vast dataset was analyzed on the IBM Blue Gene/L supercomputer using the rapid RaxML software rewritten for this distributed computing – but even with this relatively simple approach the analysis took 2.25 million CPU hours to complete. Although a *tour de force*, this analysis nevertheless seems likely to have suffered from a degree of systematic error due to the suboptimal model available to analyze such a large dataset [39]. More complex models taking into account variation in amino acid composition across sites as well as rates across sites (CAT) have been shown to have a closer fit to real datasets and to be less prone to LBA [20]. Reanalyses of this complete dataset using CAT-type models are practically impossible, however, as they are considerably more CPU-intensive.

To achieve an accurate tree using current software and hardware, a tradeoff must be made between the size of the dataset and the sophistication of the method used to analyze it. One attractive approach is simply to discard positions (and also taxa) within the concatenated alignment that are missing most data. This approach serves to shorten the alignment while improving at least one measure of quality: the proportion of missing data. One possible way to make a virtue out of the necessity to cut the size of such large datasets is to use a jackknife approach whereby a randomly selected (large) proportion of the alignment is omitted from multiple repetitions of the analysis. Each random subsample can be analyzed relatively rapidly and the repeated small samples used as nonparametric bootstrap replicates providing an additional measure of node support. An alternative is to address individual phylogenetic problems with bespoke datasets containing many genes but fewer taxa (e.g. restricting an analysis to the Lophotrochozoa), their selection depending on the problem in hand. Finally, some workers are successfully using datasets based on a large and predefined set of genes (approximately 70) gathered by the more traditional approach of using degenerate PCR primers to amplify the same gene regions from many taxa [64].

Systematic searches for 'constant but different' NAD5-like genes

For future phylogenomic research the *NAD5* example could be an indicator of another way to proceed (Box 1). The characteristic of *NAD5* that makes it so useful can be summarized as 'constant but different' because it is, in effect, a binary switch with either a conserved primitive or a conserved derived state. A systematic search for such characters would need to identify genes with typical rates of evolution within taxa but with an atypically high rate

along an internode leading to one specific (and predefined) clade – the branch along which the derived character state has evolved. Some work in this direction has already been carried out, albeit with an eye to discovering genes with a step-change in function rather than for their use in phylogenetics [65].

Concluding remarks

To the outsider, phylogenetic debates can seem particularly fractious. If different investigators can reach strongly supported but entirely different conclusions by analyzing the same data, how does a consensus emerge and should it be trusted? A central guide here must be concordant results from distinct sources of data, be they molecular or morphological, and the biological credibility of the alternative views.

In this context, the value of rare genomic changes and genomic presence/absence characters is not as traits that will definitively resolve all phylogenetic questions, but as additional and relatively independent tests of hypotheses. If the evidence of such rare genomic changes is entirely inconsistent with all other hypotheses, it is as much our assumptions about these changes, and the manner in which genomes can evolve, that should be subject to scrutiny.

The falling cost of DNA sequencing means that, in the near future, phylogenetic questions will be approached with greatly expanded molecular datasets, both in terms of sampled taxa and quantity of data, as transcript sampling becomes less attractive than whole-genome sequencing. Dense taxon sampling of genomes will lead to a better understanding of the evolutionary dynamics of processes such as changing intron locations and the secondary absence of characters such as miRNAs, and will also provide greatly expanded gene sets. Concordant phylogenetic results from these multiple data sources will ultimately prove hard to argue with. Such a situation returns us to the question of whether Cambrian cladogenesis has left a sufficient genomic signal to resolve inter-phylum relationships. Based on the results from the disparate datasets reviewed here, we see reasons for optimism.

References

1 Budd, G. and Jensen, S. (2000) A critical reappraisal of the fossil record of the bilaterian phyla. *Biol. Rev. Camb. Philos. Soc.* 75, 253–295
2 Philippe, H. *et al.* (1994) Can the Cambrian explosion be inferred through molecular phylogeny? *Development* (Suppl.), 15–25
3 Rokas, A. *et al.* (2005) Animal evolution and the molecular signature of radiations compressed in time. *Science* 310, 1933–1938
4 Telford, M.J. (2002) Cladistic analyses of molecular characters: the good, the bad and the ugly. *Contrib. Zool.* 71, 93–100
5 Consortium, C.e.S. (1998) Genome sequencing of the nematode *C.elegans*: a platform for investigating biology. *Science* 282, 2012–2018
6 Adams, M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195
7 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
8 Adoutte, A. *et al.* (2000) The new animal phylogeny: reliability and implications. *Proc. Natl. Acad. Sci. U.S.A.* 97, 4453–4456
9 Aguinaldo, A.M. *et al.* (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387, 489–493
10 Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410
11 Philippe, H. *et al.* (2005) Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol.* 5, 50
12 Blair, J.E. *et al.* (2002) The evolutionary position of nematodes. *BMC Evol. Biol.* 2, 7
13 Wolf, Y.I. *et al.* (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res.* 14, 29–36
14 Copley, R.R. *et al.* (2004) Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of *Caenorhabditis elegans*. *Evol. Dev.* 6, 164–169
15 Philippe, H. *et al.* (2009) Phylogenomics restores traditional views on deep animal relationships. *Curr. Biol.* 19, 706–712
16 Philippe, H. *et al.* (2005) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol. Biol. Evol.* 22, 1246–1253
17 Holton, T.A. and Pisani, D. (2010) Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm. *Genome Biol. Evol.* 2, 310–324
18 Delsuc, F. *et al.* (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361–375
19 Webster, B.L. *et al.* (2006) Mitogenomics and phylogenomics reveal priapulid worms as extant models of the ancestral Ecdysozoan. *Evol. Dev.* 8, 502–510
20 Lartillot, N. *et al.* (2007) Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* 7 (Suppl. 1), S4
21 Lartillot, N. and Philippe, H. (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* 21, 1095–1109
22 Ciccarelli, F.D. *et al.* (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311, 1283–1287
23 Rokas, A. and Holland, P.W.H. (2000) Rare genomic changes as a tool for phylogenetics. *Trends Evol. Ecol.* 15, 454–459
24 Telford, M.J. and Budd, G.E. (2003) The place of phylogeny and cladistics in Evo-Devo research. *Int. J. Dev. Biol.* 47, 479–490
25 Rogozin, I.B. *et al.* (2007) Analysis of rare amino acid replacements supports the Coelomata clade. *Mol. Biol. Evol.* 24, 2594–2597
26 Rogozin, I.B. *et al.* (2008) Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. *Biol. Direct* 3, 7
27 Rogozin, I.B. *et al.* (2007) Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. *Mol. Biol. Evol.* 24, 1080–1090
28 Zheng, J. *et al.* (2007) Support for the Coelomata clade of animals from a rigorous analysis of the pattern of intron conservation. *Mol. Biol. Evol.* 24, 2583–2592
29 Swofford, D.L. *et al.* (1996) Phylogenetic inference, In *Molecular Systematics* (2nd edn) (Hillis, D.M. *et al.*, eds), pp. 407–514, Sinauer
30 Felsenstein, J. (2005) *PHYLIP (Phylogeny Inference Package)*, . Department of Genome Sciences, University of Washington
31 Dollo, L. (1893) Les lois de l'évolution. *Bull. de la Soc. Belge de Geologie Paléontologie et d'Hydrologie* 7, 164–166
32 Irimia, M. *et al.* (2007) Rare coding sequence changes are consistent with Ecdysozoa, not Coelomata. *Mol. Biol. Evol.* 24, 1604–1607
33 Roy, S.W. and Irimia, M. (2008) Rare genomic characters do not support Coelomata: intron loss/gain. *Mol. Biol. Evol.* 25, 620–623
34 Roy, S.W. and Irimia, M. (2008) Rare genomic characters do not support Coelomata: RGC_CAMs. *J. Mol. Evol.* 66, 308–315
35 Roy, S.W. and Gilbert, W. (2005) Resolution of a deep animal divergence by the pattern of intron conservation. *Proc. Natl. Acad. Sci. U.S.A.* 102, 4403–4408
36 Heimberg, A.M. *et al.* (2010) microRNAs reveal the interrelationships of hagfish, lampreys and gnathostomes and the nature of the ancestral vertebrate. *Proc. Natl. Acad. Sci. U.S.A.* 107, 19379–19383
37 Rota-Stabelli, O. *et al.* (2011) A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc. R. Soc. Lond. B.* 278, 298–306
38 Heimberg, A.M. *et al.* (2008) microRNAs and the advent of vertebrate morphological complexity. *Proc. Natl. Acad. Sci. U.S.A.* 105, 2946–2950
39 Philippe, H. *et al.* (2011) Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature* 470, 255–258
40 Hillis, D.M. (1999) SINEs of the perfect character. *Proc. Natl. Acad. Sci. U.S.A.* 96, 9979–9981

41 Papillon, D. *et al.* (2004) Identification of chaetognaths as protostomes is supported by the analysis of their mitochondrial genome. *Mol. Biol. Evol.* 21, 2122–2129
42 Dunn, C.W. *et al.* (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452, 745–749
43 Hejnol, A. *et al.* (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. R. Soc. Lond. B.* 276, 4261–4270
44 Philippe, H. and Telford, M.J. (2006) Large-scale sequencing and the new animal phylogeny. *Trends Ecol. Evol. (Amst.)* 21, 614–620
45 Cook, C.E. *et al.* (2001) Hox genes and the phylogeny of the arthropods. *Curr. Biol.* 11, 759–763
46 Friedrich, M. and Tautz, D. (1995) rDNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* 376, 165–167
47 Pisani, D. *et al.* (2004) The colonization of land by animals: molecular phylogeny and divergence times among arthropods. *BMC Biol.* 2, 1
48 Janssen, R. and Budd, G.E. (2010) Gene expression suggests conserved aspects of Hox gene regulation in arthropods and provides additional support for monophyletic Myriapoda. *Evodevo* 1, 4
49 Telford, M.J. and Thomas, R.H. (1995) Demise of the Atelocerata? *Nature* 376, 123–124
50 Telford, M.J. *et al.* (2008) The evolution of the Ecdysozoa. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 1529–1537
51 Stollewerk, A. and Simpson, P. (2005) Evolution of early development of the nervous system: a comparison between arthropods. *Bioessays* 27, 874–883
52 Chipman, A.D. and Stollewerk, A. (2006) Specification of neural precursor identity in the geophilomorph centipede *Strigamia maritima*. *Dev. Biol.* 290, 337–350
53 Rota-Stabelli, O. and Telford, M.J. (2008) A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. *Mol. Phylogenet. Evol.* 48, 103–111
54 Budd, G.E. and Telford, M.J. (2009) The origin and evolution of the arthropods. *Nature* 457, 812–817
55 Delsuc, F. *et al.* (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* 439, 965–968
56 Ruppert, E.E. (2005) Key characters uniting hemichordates and chordates: homologies or homoplasies? *Can. J. Zool.* 83, 8–23
57 Bourlat, S.J. *et al.* (2006) Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* 444, 85–88
58 Delsuc, F. *et al.* (2008) Additional molecular support for the new chordate phylogeny. *Genesis* 46, 592–604
59 Putnam, N.H. *et al.* (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317, 86–94
60 Telford, M.J. and Holland, P.W.H. (1997) Evolution of 28S ribosomal DNA in chaetognaths: duplicate genes and molecular phylogeny. *J. Mol. Evol.* 44, 135–144
61 Ryan, J.F. *et al.* (2010) The homeodomain complement of the ctenophore *Mnemiopsis leidyi* suggests that Ctenophora and Porifera diverged prior to the Parahoxozoa. *EvoDevo* 1, 9
62 Pang, K. *et al.* (2010) Gegnomic insights into Wnt signalling in an early diverging metazoan, the ctenophore *Mnemiopsis leidyi*. *EvoDevo* 1, 10
63 Podar, M. *et al.* (2001) A molecular phylogentic framework for the phylum Ctenophora using 18S genes. *Mol. Phyl. Evol.* 21, 218–2230
64 Regier, J.C. *et al.* (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* 463, 1079–1083
65 Studer, R.A. and Robinson-Rechavi, M. (2010) Large-scale analysis of orthologs and paralogs under covarion-like and constant but different models of amino acids evolution. *Mol. Biol. Evol.* 27, 2618–2627
66 Rokas, A. *et al.* (1999) Intron insertion as a phylogenetic character: the *engrailed* homeobox of Strepsiptera does not indicate affinity with Diptera. *Insect Mol. Biol.* 8, 527–530
67 Srivastava, M. *et al.* (2010) The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* 466, 720–726
68 Denoeud, F. *et al.* (2010) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* 330, 1381–1385
69 Kawashima, T. *et al.* (2009) Domain shuffling and the evolution of vertebrates. *Genome Res.* 19, 1393–1403
70 Gough, J. (2005) Convergent evolution of domain architectures (is rare). *Bioinformatics* 15, 1464–1471
71 King, N. *et al.* (2008) The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451, 783–788
72 Sebé-Pedrós, A. *et al.* (2011) Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsaspora owczarzaki*. *Mol. Biol. Evol.* 28, 1241–1254
73 Matus, D. *et al.* (2006) Broad taxon and gene sampling indicate that chaetognaths are protostomes. *Curr. Biol.* 16, R575–576
74 Telford, M.J. and Holland, P.W. (1993) The phylogenetic affinities of the chaetognaths: a molecular analysis. *Mol. Biol. Evol.* 10, 660–676
75 Marletaz, F. *et al.* (2006) Chaetognath phylogenomics: a protostome with deuterostome-like development. *Curr. Biol.* 16, R577–R578
76 de Rosa, R. *et al.* (1999) Hox genes in brachiopods and priapulids and protostome evolution. *Nature* 399, 772–776
77 Nielsen, C. (2001) *Animal Evolution. Interrelationships of the Living Phyla*, Oxford University Press
78 Halanych, K.M. *et al.* (1995) Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science* 267, 1641–1643
79 Halanych, K.M. (1996) Convergence in the feeding apparatuses of Lophophorates and pterobranch Hemichordates revealed by 18S rDNA: an interpretation. *Biol. Bull.* 190, 1–5
80 Ruiz Trillo, I. *et al.* (1999) Acoel flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. *Science* 283, 1919–1923
81 Egger, B. *et al.* (2009) To be or not to be a flatworm: the acoel controversy. *PLoS ONE* 4, e5502
82 Sperling, E.A. *et al.* (2009) Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. *Mol. Biol. Evol.* 26, 2261–2274
83 Peterson, K.J. *et al.* (2009) MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. *BioEssays* 31, 736–747