Phylogenomics and the reconstruction of the tree of life.

Frédéric Delsuc, Henner Brinkmann, Hervé Philippe

PHYLOGENOMICS AND THE RECONSTRUCTION OF THE TREE OF LIFE

Frédéric Delsuc, Henner Brinkmann and Hervé Philippe

Canadian Institute for Advanced Research, Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C3J7, Canada.

Correspondence to H.P. email: herve.philippe@umontreal.ca

Preface

As more complete genomes are sequenced, phylogenetic analysis is entering a new era — that of phylogenomics. One branch of this expanding field aims to reconstruct the evolutionary history of organisms based on the analysis of their genomes. Recent studies have demonstrated the power of this approach, which has the potential to provide answers to a number of fundamental evolutionary questions. However, challenges for the future have also been revealed. The very nature of the evolutionary history of organisms and the limitations of current phylogenetic reconstruction methods mean that part of the tree of life may prove difficult, if not impossible, to resolve with confidence.

Introductory paragraph

Understanding phylogenetic relationships between organisms is a prerequisite of almost any evolutionary study, as contemporary species all share a common history through their ancestry. The notion of phylogeny follows directly from the theory of evolution presented by Charles Darwin in *"The Origin of Species"*[1]: the only illustration in his famous book is the first representation of evolutionary relationships among species, in the form of a phylogenetic tree. The subsequent enthusiasm of biologists for the phylogenetic concept is illustrated by the publication of Ernst Haeckel's famous "trees" as early as 1866[2].

Today, phylogenetics — the reconstruction of evolutionary history — relies on using mathematical methods to infer the past from features of contemporary species, with only the fossil record to provide a window on the evolutionary past of life on our planet. This reconstruction involves the identification of **HOMOLOGOUS CHARACTERS** that are shared among different organisms, and the inference of phylogenetic trees from the comparison of these characters using reconstruction methods (BOX 1). The accuracy of

the inference is therefore heavily dependent upon the quality of models for the evolution of such characters. Because the underlying mechanisms are not yet well understood, reconstructing the evolutionary history of life on Earth based solely on the information provided by living organisms has turned out to be difficult.

Until the 1970s, which brought the dawn of molecular techniques for sequencing proteins and DNA, phylogenetic reconstruction was essentially based on the analysis of morphological or ultrastructural characters. The comparative anatomy of fossils and extant species has proved powerful in some respects; for example, the main groups of animals and plants have been delineated fairly easily using these methods. However, this approach is hampered by the limited number of reliable homologous characters available; these are almost non-existent in micro-organisms[3] and are rare even in complex organisms.

The introduction of the use of molecular data in phylogenetics[4] led to a revolution. In the late 1980s, access to DNA sequences increased the number of homologous characters that could be compared from less than 100 to more than 1,000, greatly improving the resolving power of phylogenetic inference. A few genes became reference markers. In particular, owing to its considerable degree of conservation across all organisms, the gene encoding small subunit ribosomal RNA (SSU rRNA) was extensively used for the classification of micro-organisms and allowed the recognition of the Archaea as a third distinct domain of the tree of life[5]. However, as more genes were analysed, topological conflicts between phylogenies based on individual genes were revealed. Moreover, information from a single gene is often insufficient to obtain firm statistical support for particular **NODES** of a phylogeny. As a consequence, numerous parts of the tree of life remained poorly resolved simply because of sampling effects due to the limited availability of data.

However, in the time it has taken you to read these lines, thousands more base pairs of sequence information will have been generated by large-scale genomic projects. This wealth of data, hardly imaginable only a decade ago, is giving birth to a new field of research, termed phylogenomics, which uses phylogenetic principles to make sense of genomic data[6]. One branch of phylogenomics involves the use of these data to reconstruct the evolutionary history of organisms. Indeed, access to genomic data could potentially alleviate previous problems of phylogenetics due to sampling effects by expanding the number of characters that can be used in phylogenetic analysis from a few thousand to tens of thousands. With this increase, the emphasis of phylogenetic inference is shifting from the search for informative characters to the development of better reconstruction methods for using genomic data. Indeed, existing models used in tree-building algorithms only partially take into account molecular evolutionary processes, and phylogenomic inference will benefit from an increased understanding of these mechanisms. Interestingly, phylogenomics is also providing the opportunity to use new "morphological-like" characters based on genome structure, such as rare genomic changes (RGCs[7,8]).

In this review, we describe current methods for phylogenomic inference and discuss their merits and pitfalls in light of their recent application to diverse phylogenetic problems. The recent improvements in the resolution of the tree of life due to large-scale studies in each of the three domains — Archaea, Bacteria and Eukaryota — are discussed. Despite holding considerable promise, the phylogenomic approach also has potential problems, stemming from the limitations of current phylogenetic reconstruction methods. We present examples of method **INCONSISTENCY**, leading to tree reconstruction artefacts, and tentatively propose solutions to these issues. Finally, we discuss the future potential of phylogenomics, and specifically address the issue of how to corroborate results from phylogenomic analyses.

Current methods for phylogenomics – an overview

The two crucial steps of classical phylogenetic inference — the identification of homologous characters and tree reconstruction — are generally preserved in phylogenomics. Therefore, as for phylogenies based on morphological data or single genes, the reliability of a phylogenomic tree depends on the quality of the characters and the accuracy of the reconstruction methods. Theoretically, reliable characters can be considered as those that have undergone only a few changes over time (ideally, a single change). Multiple changes create **HOMOPLASY** (noise) in the form of **CONVERGENCE** and **REVERSAL**, masking genuine phylogenetic signal[9].

The three main types of standard phylogenetic reconstruction method (distance, parsimony and likelihood methods; BOX 1)[10] have been adapted for use in phylogenomics. Phylogenomic reconstruction methods can be divided roughly into sequence-based methods and methods based on whole-genome features. The latter are becoming increasingly popular, but their relatively recent introduction limits their critical evaluation at this stage. As a consequence, methods based on multiple sequence alignment, for which an extensive methodological background exists, currently remain the methods of choice. Both of these types of method, as well as the study of rare genomic changes, are discussed below.

Sequence-based methods

Number of characters versus number of species. Sequence-based phylogenomic methods are based on the comparison of primary sequences, and phylogenetic trees are

inferred from multiple sequence alignments. Around the year 2000, the move from single-gene to multiple-gene analyses, using sets of fewer than 20 genes (e.g. REF.[11-15]), slightly preceded the phylogenomic era. This later progressed to the use of datasets containing more than 100 genes[16-21], but at the cost of considering fewer species than in single gene studies because of data availability and computational time constraints (for recent exceptions see REF. [22,23]).

A longstanding debate in phylogenetics is the question of whether the greatest improvement in accuracy results from an increased number of characters (in this case, genes) or species[24-28]. Evidence from computer simulations has been equivocal[27,28], whereas empirical studies tend to support the importance of extensive species sampling[24,29,30]. In practice, increasing the number of genes is straightforward for species for which complete genomes have been sequenced (e.g. REF. [16,18,19,21,31]). However, the largest complete datasets (in which all genes are represented for all species) that can be mined from sequence databases are asymmetrical — that is, they include many species and few genes, or few species and many genes[22,32]. As it is likely to produce more accurate results[29,30], assembling phylogenomic datasets rich in both species and genes is necessary. However, this is invariantly associated with missing data, as some genes are not represented for all species (e.g. REF. [17,20,22]), and the effects of these missing data are discussed below.

Supermatrices and supertrees. Once multiple gene alignments have been assembled from the chosen dataset as described above, two alternative approaches can be used for phylogenomic reconstruction. Following from the total evidence principle of using all the relevant available data[33], the most popular strategy is to analyse the "supermatrix" constituted by the concatenation of individual genes (BOX 2) using standard sequence-

based methods[34]. In this approach, the sequences of genes that are not represented for some species are coded as question marks. Several studies have implicitly made the assumption that a certain amount of missing data can be tolerated by tree reconstruction methods (e.g. 20% missing data in REF. [11], 12.5% in REF. [15], 25% in REF. [17]). Recent empirical[20,22,35] and simulation[20,36] studies have shown that this proportion can be surprisingly high without losing too much accuracy. The impact of missing data is limited because species for which sequence information is incomplete are still represented by a large number of informative characters in phylogenomic studies[20]. In fact, having a species represented by only 10 genes out of 100 that are analysed in the whole study might generally be less problematic than not considering that species at all (see BOX 3).

The robustness of the supermatrix approach to missing data makes it powerful for phylogenetic reconstruction, as phylogenomic datasets can be assembled at low cost by mining existing databases[22,32] or by the sequencing of multiple PCR-targeted loci[11,14,15], as well as cDNAs and ESTs[17,20,23]. This allows the incorporation of a large number of species, instead of being restricted to model organisms for which complete genome sequences are available.

The second sequence-based phylogenomic approach consists of analysing each data partition (such as genes) individually, and combining the resulting trees — which contain information from partially overlapping species — into a “supertree” (BOX 2). Different methods for constructing supertrees have been proposed[37], but because of its intrinsic simplicity the matrix representation using parsimony (MRP) method[38,39] is the most popular[40]. Supertree methods have mainly been used for combining trees obtained from disparate sources of data (for example, morphological and molecular data) in order to provide an overview of the phylogeny of a given group. For example, this approach has been used in studies of placental mammals[41]. In phylogenomics, this approach has so far

been restricted to a few studies that have attempted to reconstruct the phylogenies of Bacteria[42] and of model eukaryotic species for which complete genomes are available[31]. However, supertree reconstruction is currently an active area of research[40], and its use is likely to expand in the near future.

The relative merits of the two sequence-based approaches have not been thoroughly explored. Empirical comparisons suggest the superiority of the supermatrix approach over MRP in a study of crocodylians (crocodiles, caimans, alligators and gavials)[43], whereas the two approaches were found to be roughly equivalent in terms of the results produced when reconstructing the phylogeny of grasses[44]. However, the comparison between the two approaches is made difficult by the different types of characters used in each method[45], and more work is needed to address these issues. Nevertheless, in phylogenomic studies of Bacteria, supermatrix[46,47] and supertree[42] analyses have produced fairly similar trees based on different datasets.

Methods based on whole-genome features

Gene content and gene order. Methods of phylogenetic reconstruction based on the comparison of whole-genome features beyond the sequence level — such as gene order and gene content (i.e. the specific genes found in a genome) — have recently been developed (BOX 2). Unlike classical sequence-based approaches, methods based on gene content and gene order do not rely on a multiple-sequence alignment step. However, they do still depend on **HOMOLOGY** and **ORTHOLOGY** assessment (see below). Changes in gene content and gene order within genomes result in characters with billions of possible states, as compared to only four states for nucleotide sequences. As a result, they are less prone to homoplasy by convergence or reversal, and may therefore potentially represent good

phylogenetic markers[9], as long as they contain enough phylogenetic information[48]. Although they use different character types to those employed in sequence-based approaches, these methods nevertheless use standard tree-reconstruction algorithms (see REF. [49] for a recent review).

Phylogenetic trees reconstructed from gene-content information have generally been reconstructed using distance[50-56] or parsimony[55-58] methods (BOX 2). One concern with gene-content analyses is the erroneous grouping of organisms with a similar number of genes[53,56,58,59]. This phenomenon, known as the 'big genome attraction' artefact[60], is thought to result from substantial convergent gene losses occurring in certain genomes, for example, those of intracellular parasites[58,59]. Progress in explicitly modelling the molecular details of genome evolution has recently been made with the development of probabilistic approaches, and this should ultimately lead to more accurate inferences based on gene content[60-62].

Gene order was recognized early on as a valuable phylogenetic character[63]. However, its use involves estimating evolutionary distance from the number of rearrangements necessary to transform one genome into another, which is a complex mathematical problem[64]. Even with the **HEURISTIC** approach of **BREAKPOINT** minimisation, computational burden has significantly restricted the application of phylogenetic reconstruction based on gene order[65]. Only inversions are currently considered, but explicit models of gene order evolution are being implemented in order to handle duplications, insertions and deletions[66]. Progress in devising efficient algorithms still needs to be achieved before realising the full potential of this approach. Whole-genome prokaryotic trees have nevertheless been constructed through this approach using parsimony and distance methods, under the drastically simplifying assumption that gene order can be

described by the presence or absence of pairs of orthologous genes[53,56]. However, this approach suffers from the current lack of evaluation of its accuracy.

The issue of orthology assessment. Dependence on the assessment of orthology is an important issue in phylogenomic studies. This assessment is primarily based on sequence-similarity searches that can be misleading[67] because of differences in evolutionary rates and/or base composition between species, and the occurrence of **HORIZONTAL GENE TRANSFER** (HGT; see REF. [68]). Orthology assessment ideally requires rigorous and time-consuming phylogenetic analyses of individual genes[47,69,70]. Although automation procedures have been proposed[71], this step is often overlooked in the reconstruction of phylogenies based on gene content, rendering their critical evaluation difficult. However, analyses of homologous[58] or orthologous[56] genes using the gene-content approach have yielded fairly congruent trees. Furthermore, recent gene-content analyses attempting to filter the noise introduced by HGT and **PARALLEL GENE LOSSES** showed only a limited effect of these factors on the results[52,54]. This suggests that gene-content methods might be more robust to the potential problems of orthology assessment than was first thought. The accurate modelling of HGT events and parallel gene losses will nevertheless be necessary to avoid the "big genome attraction" artefact[60].

The 'DNA string' approach. Finally, another approach derived from whole-genome features, which is not dependent upon homology/orthology assessment, is based on the distribution of oligonucleotides ("DNA strings") in genomes[55,72-75] (BOX 2). This approach is based on the observation that each genome has a characteristic "signature" with regard to these strings; these are defined, for example, as the ratio between observed dinucleotide frequencies and those expected if neighbouring nucleotides were chosen at random[72]. The

few methods currently used for this approach show that it is possible to extract phylogenetic signal using this oligonucleotide 'word usage'[55,73-75]. However, a comparison with SSU rRNA sequences in the context of the prokaryotic phylogeny shows that this usage seems to evolve much faster than SSU rRNA[74]. Therefore, **SATURATION** of the phylogenetic signal contained in oligonucleotides may limit the use of such approaches for inferring ancient divergence events.

The study of rare genomic changes

Genomes can also be studied using the traditional methodology of comparative morphologists by looking for shared complex characters — known as rare genomic changes (RGCs) — that have a very low probability of being the result of convergence (BOX 2). As well as gene order, such RGCs include intron positions, insertions and deletions (indels), retroposon (SINE and LINE) integrations and gene fusion and fission events[8,9].

Until now, only a few characters of this kind have been used to address specific phylogenetic questions, such as the phylogeny of placental mammals[76,77] and jawed vertebrates[78], or the position of the **ROOT** in the eukaryotic tree[79,80]. Although rare, homoplasy can also affect RGCs[81-84], so inferences should not rely on only a few characters. With the sequencing of complete genomes, the statistical study of large numbers of RGCs certainly represents a promising avenue.

Recent achievements of phylogenomics

At the time of writing, 260 complete genomes have been sequenced (33 eukaryotes, 206 bacteria and 21 archaea), and more than 1,000 genome projects are ongoing. These figures illustrate the large datasets becoming available for phylogenomic studies and the extraordinary potential of the phylogenomic approach to shed light on longstanding phylogenetic questions, spanning all levels of the tree of life (FIGURE 1). In this section, we present recent advances for each of the three domains of life that have been enabled by phylogenomics.

Eukaryotes. The reconstructions of phylogenies of placental mammals and land plants represent the most spectacular examples of recent advances enabled by phylogenomics. The evolutionary history of placental mammals was traditionally considered to be irresolvable, due to the explosive radiation of species that occurred in a short space of time. However, this has now been elucidated by the analysis of supermatrices containing about 20 nuclear genes[14,15,85,86], and the resulting phylogeny has been confirmed by recent analyses of complete mitochondrial genomes[30,87], with only few nodes left unresolved. All but one of the 18 morphologically defined extant mammalian orders have been confirmed by molecular studies. The notable exception to this is the insectivores, which have been split into two distinct orders by the recognition of an unexpected group of African origins named Afrotheria[14,15,77]. Four major groups of placental orders have been identified, and their origins might be explained by geographic isolation, resulting from plate tectonics, in the early stages of diversification among placental mammals[85]. Molecular studies have also revealed the prevalence of morphological convergence during the evolution of placental mammals with the occurrence of parallel adaptive radiations[14]. This might partly explain why reconstructing the phylogeny of placental orders on the basis of morphological characters has been difficult.

Similarly, studies of multiple genes from all three compartments of the plant cell (the mitochondrion, chloroplast and nucleus) have helped to overcome longstanding uncertainties in land plant phylogeny[11,88-92]. These advances have brought several changes to the traditional botanical classification of flowering plants, which were previously based on morphological features, such as floral characteristics and leaf shape[93]. In this case too, molecular evidence revealed the plasticity of phenotypic characters, highlighting several examples of previously unrecognised convergent evolution. This is perhaps best illustrated by the case of the sacred lotus (*Nelumbo nucifera*), which was previously thought to be related to water lilies (Nymphaeaceae), whereas molecular studies unveiled a close phylogenetic affinity with plane trees (genus *Platanus*)[88].

At a larger scale, sequence-based phylogenomic studies of eukaryotic phylogeny confirmed the **MONOPHYLY** of most phyla, which were originally defined based on ultrastructural or rRNA analyses. However, they also demonstrated the common origin of a group of morphologically diverse amoebae, which were previously thought to have evolved independently[17]. Phylogenomic analyses of nuclear[20,94] and mitochondrial genes[95] have also corroborated the long suspected sister-group relationship between the unicellular choanoflagellates and multicellular animals. However, the lack of representatives from several major lineages (i.e. Rhizaria, Cryptophytes, Haptophytes and Jakobids) in phylogenomic studies currently prevents the study of most of the working hypotheses derived from decades of ultrastuctural and molecular studies, which have proposed the division of the eukaryotic world into six major groups[80,96]: the Opisthokonta, Amoebozoa, Plantae, Chromalveolata, Rhizaria, and Excavata (FIGURE 1). The validity of these proposed “kingdoms” represents one of the most important outstanding questions that phylogenomics has the potential to answer.

Finally, the question of the origin of the eukaryotes has been recently addressed[97] using a gene-content method that allows the modelling of genome-fusion events[60]. The authors proposed that the eukaryotes originated from a fusion event between a bacterial species and an archaeal species, leading to a ring-like structure at the root of the tree of life[97]. This scenario would account for the chimeric nature of the eukaryotes, which has been inferred from the observation that many eukaryotic metabolic genes have a greater number of similar counterparts in Bacteria, whereas most of those involved in information processing, such as transcription and translation, have more similar counterparts in Archaea. However, the mitochondrial endosymbiosis, a well-characterised fusion event, was followed by massive, lateral gene transfers to the eukaryotic nucleus and constitutes a major source of "bacterial-like" genes in Eukaryotes[98]. Distinguishing between this endosymbiosis and an additional earlier genome fusion event is difficult, and the accuracy of the new gene-content method described above[60] has not yet been sufficiently evaluated to make a definitive statement on this fundamental evolutionary question.

Prokaryotes (Bacteria and Archaea). Despite the large number of complete prokaryotic genomes available, the picture of bacterial and archaeal evolution provided by SSU rRNA in the 1980s[99] remains surprisingly unchanged. The development of phylogenomic studies in Prokaryotes have been largely held back by the supposedly predominant role of HGTs in shaping the evolutionary history of micro-organisms, which is thought to have been so widespread that it may have blurred the phylogenetic signal for a prokaryotic phylogeny[100]. HGTs are undeniably an important source of genome evolution and innovation in prokaryotes[101]. Nevertheless, phylogenomic methods based on whole-genome features such as DNA strings[55,74,75], gene content[50,53,54], conservation of gene pairs[53,56] and protein domain structure[55,102], have all yielded phylogenetic trees that are similar to the

corresponding SSU rRNA tree in the sense that they recovered the three domains of life and the main groups within both prokaryotic domains. Moreover, both supertree[42] and supermatrix[46,47,103] analyses identified a core of genes that rarely undergo HGT, from which it is possible to infer the phylogeny of prokaryotes. This suggests that HGTs do not prevent the recovery of a phylogenetic signal in Prokaryotes, although they do constitute an additional source of noise[68]. For example, in Archaea, a major division between the Euryarchaeota and Crenarchaeota (FIGURE 1) is supported by evidence from rRNA[99] and sequence-based phylogenomic studies[42,103]. However, there has been difficulty in recovering this division using gene-content methods[49], and this has been interpreted as being a consequence of HGT[53].

In the Bacteria, methods based on whole-genome features[50,53-56,58,74] and sequence-based phylogenomic inferences[42,46,47,103] have recovered the respective monophyly of all major groups that were suggested by SSU rRNA (e.g. Cyanobacteria, Spirochaetes, Chlamydiales and Proteobacteria; FIGURE 1). However, the relationships among these groups, which are unresolved in the SSU rRNA tree[99], remain weakly supported. The only tentative groupings that might be proposed at this stage are Chlamydiales with Spirochaetes, Aquificales with Thermotogales, and High-GC Bacteria with Deinococcales and Cyanobacteria, as they were recovered in independent analyses[42,46,47,56]. However, biases in amino-acid composition and differences in evolutionary rates, instead of genuine phylogenetic signal (discussed below), might also explain these results. These potential new groupings can thus only be considered as working hypotheses for future phylogenomic studies, which will be based on more species and will use methods specifically designed to tackle the issues raised above. The resolution of the bacterial radiation is perhaps the biggest challenge to phylogenomics at present.

Future challenges of phylogenomics

Because it uses many characters, phylogenomics leads to a drastic reduction in **STOCHASTIC** or **SAMPLING ERRORS** associated with the finite length of single genes in traditional phylogenetic analyses. It is not, however, immune to **SYSTEMATIC ERRORS**, which are dependent upon data quality and inference methods. The emergence of phylogenomics therefore brings the field full-circle to the roots of molecular phylogenetic analysis, with potential pitfalls in the form of tree-reconstruction artefacts, which were among the earliest issues faced by phylogenetics[104]. Here, we will discuss systematic errors in the case of classical sequence-based methods (those based on supermatrices), as they are the best characterised. However, systematic error can also affect all other approaches, as exemplified by the occurrence of the "big genome attraction artefact" in gene-content methods. This artefact is analogous to the problem of compositional bias in sequence data (see below), and its impact can be reduced by the use of models of genome evolution that have been inferred from sequence-based models[60].

***Misleading effects of inconsistency*.** The use of large datasets generally results in a global increase in the resolution of phylogenetic trees, as measured by standard statistical indices such as bootstrap percentages obtained from performing a **BOOTSTRAP ANALYSIS**[105] and **BAYESIAN POSTERIOR PROBABILITIES**[106]. However, obtaining a strongly supported tree does not necessarily mean that it is correct. Indeed, these statistical indices only assess sampling effects, and give an indication of tree reliability that is conditional on the data and the method. So, if the method does not correctly handle properties of the data, an incorrect tree can receive strong statistical support (BOX 3).

A phylogenetic reconstruction method is statistically consistent if it converges towards the true tree as more data are analysed. All phylogenetic reconstruction methods make assumptions about the process of sequence evolution either implicitly (in the case of parsimony methods) or explicitly (in the case of distance and probabilistic methods). In theory, these tree-building methods are statistically consistent as long as their assumptions are met; however, every method is known to be inconsistent under some conditions[34]. When their assumptions are violated, current methods are prone to converge towards an incorrect solution, as shown by simulation studies[107-109]. In practice, method assumptions are always violated to some extent, as current models fail to capture the full complexity of sequence evolution[110]. These model violations generate an erroneous signal (noise) that will compete with the genuine phylogenetic (historical) signal. In general, noise is randomly distributed in sequences, and tree reconstruction methods are able to extract the more structured phylogenetic signal. However, when the historical signal is weak, such as for ancient phylogenetic relationships, and/or the noise is predominant, because the same biases are shared by phylogenetically unrelated organisms (see below), the phylogenetic inference can be misled.

Sources of inconsistency. There are several causes of model inadequacy, as several simplifying assumptions are generally made. These include the independence of evolutionary changes at different sites and the homogeneity of the nucleotide-substitution process. For example, compositional biases can result in the artefactual grouping of species with similar nucleotide or amino-acid compositions[111], because most methods assume the homogeneity of the substitution process and the constancy of sequence composition (stationarity) through time (BOX 3a). Moreover, variations in the evolutionary rate among species can cause the well known and widespread long-branch attraction

(LBA) artefact[104,112]. Here, high evolutionary rates increase the chance of convergence and reversal, leading to the artefactual grouping together of fast-evolving species[104] (BOX 3b). These biases are the best characterised sources of inconsistency in sequence data. In addition, the confounding effects on phylogenetic inference of heterotachy[113] (variation of evolutionary rate through time) are only now beginning to be better understood (BOX 3).

Examples of inconsistency. Two recent examples illustrate the problem of inconsistency in phylogenomic studies. The first concerns the controversy surrounding the phylogenetic position of *Amborella trichopoda* within the flowering plants (angiosperms). In contrast with the classical two-way division of flowering plants into monocots and dicots, this dicotyledonous plant has been removed from dicots (now called eudicots) and is now considered to have been one of the earliest angiosperms to evolve, based on several lines of molecular evidence[11,114]. However, recent phylogenomic analyses of complete chloroplast genomes argued for a return to the traditional phylogeny separating monocots and dicots, with *Amborella* being part of dicots[115,116]. The limited taxon sampling used in these studies, combined with high levels of heterogeneity in evolutionary rates among species (the grasses, which were used as representatives of monocots, are particularly fast evolving), have been pointed out as possible sources of this discrepancy[117,118]. More specifically, maximum likelihood (ML) phylogenetic analyses have been carried out using models that take into account among-site rate variation and use a dataset that includes a slowly evolving monocot (*Acorus*)[118]. These analyses found strong support for the early emergence of *Amborella*, and this suggests that an LBA artefact was responsible for the apparent early emergence of fast evolving grasses included in previous analyses. However, as *Amborella* is the only extant representative of its family, its basal position within the angiosperms might prove difficult to resolve conclusively[118].

The second controversial example relates to the phylogeny of bilaterian animals. The classical gradualist view that divides bilaterians into Acoelomata, Pseudocoelomata and Coelomata on the basis of the nature of their body cavity (coelom) has been overturned by accumulating molecular evidence[119,120]. The starting point of this revolution was the recognition based on analyses of SSU rRNA of a monophyletic group of moulting animals — named the Ecdysozoa — that includes arthropods and nematodes[121]. The new animal phylogeny consists of three major bilaterian groups: the Deuterostomia (including vertebrates), the Lophotrochozoa (including brachiopods, annelids and molluscs) and the Ecdysozoa. However, recent phylogenomic studies of model animals have resurrected the ancient view in supporting the grouping of arthropods and vertebrates (Coelomata), to the exclusion of the nematodes[16,21,31,122]. Although these studies considered a very large number of genes, they nevertheless suffer from poor species sampling as they included a maximum of 10 species, a configuration potentially prone to phylogenetic artefacts (see BOX 3b). In fact, the analysis of a much larger species sample has revealed the extent of effect of the LBA artefact on animal phylogenomic analyses, leading to the artificial grouping of fast-evolving nematodes and platyhelminthes[23]. When different methods were used to avoid this artefact, strong support was obtained in favour of the new animal phylogeny including Ecdysozoa[23].

Reducing the perils of inconsistency. In the pre-genomic era, the most straightforward way of detecting an erroneous phylogenetic result was the observation that incongruent trees are obtained from different genes. For example, the misplacement of microsporidia at the base of the eukaryotic tree, instead of as a sister-group of fungi, was revealed by the comparison of several single-gene phylogenies[123]. However, when whole genome

information is used, an erroneous result due to method inconsistency is difficult to ascribe as only one tree is produced.

As increasingly large datasets are analysed, the probability increases that strongly supported but erroneous groupings remain undetected. Therefore, using the most accurate tree-reconstruction methods available is of the utmost importance. Arguably, probabilistic methods of phylogenetic inference, such as ML and Bayesian methods should be preferred, as they explicitly incorporate the processes of sequence evolution in the models that they use[110]. The use of the most complex models will reduce the probability of becoming inconsistent, as they will fit the data better. However, despite the fact that simulation studies suggest that probabilistic methods are relatively robust to violations of the model's primary assumptions[124-126], this may not hold in extreme cases[109].

More realistic models of sequence evolution are therefore needed. Research in this area is ongoing, with the most recently developed models relaxing the assumption of independence among sites by taking into account the occurrence of context-dependent[127], multiple-nucleotide[128] and structurally constrained[129,130] substitutions. Likelihood models that relax the assumptions of homogeneity and stationarity have also been designed to handle sequences with heterogeneous composition[131,132]. Methods of tree reconstruction based on the **COVARION MODEL OF MOLECULAR EVOLUTION** first proposed by Fitch[133] have been proposed[134,135] and implemented within a maximum likelihood[136] and Bayesian[137] framework to handle heterotachy. These efforts have been recently followed by the development of mixture models, allowing distinct models for different classes of sites, in order to accommodate among-site heterogeneities in evolutionary dynamics[138,139]. In general, mixture models seem to represent a promising avenue to correctly handle sequences that evolved through heterogeneous processes[109]. Ultimately, evolutionary models should integrate all these improvements simultaneously. However, ML methods

could misbehave because of the increased variance associated with the estimates of large number of parameters. The development of complex models should therefore avoid falling into the “infinitely many parameter trap”[34].

However, improved probabilistic methods, as described above, might not hold all the answers to the inconsistency problem. Indeed, fast-evolving characters are particularly challenging for phylogenetic inference because they are likely to have experienced multiple changes, eroding the phylogenetic signal[9]. At these sites, using the correct model of sequence evolution is of particular importance for inferring hidden changes in order to separate signal from noise. However, since the perfect model does not exist, various strategies have been developed in order to reduce the impact of systematic errors. One efficient approach is to improve species sampling (BOX3b), as multiple changes are most easily detected when many species are analysed. Moreover, different species violate model assumptions to variable degrees, and inconsistency may only occur for particular combinations of species. For instance, datasets constituted of species with heterogeneous evolutionary rates are more likely to exhibit inconsistency than those with more homogeneous rates. With single-gene datasets, focusing on the most slowly evolving taxa has been shown to counteract the LBA artefact[121]. In phylogenomics, although increasing the number of species is important (BOX 3b), it also significantly raises the computational burden, which can become a serious issue (BOX 4).

Focusing on the rarest substitution events is another way of improving phylogenetic inference. Because nucleotide substitutions between bases belonging to the same family (transitions) occur more frequently than between bases from different families (transversions), reducing the data to purines (A,G = R) and pyrimidines (C,T = Y) efficiently reduces saturation (FIGURE 2a). Compositional bias is also decreased by this so-called RY-coding strategy, as base-composition differences are often most pronounced

between bases of the same family[140]. RY-coding has proved helpful in studies that have used mitochondrial genomes[141-143], and has enabled the inconsistency of distance methods in yeast phylogenomics to be avoided[144].

Other approaches have been proposed for identifying and removing the fastest evolving sites [145-149]. For example, the slow/fast (SF) method has been developed primarily for studying ancient divergence events[147] and has been helpful in tackling the LBA artefact when reconstructing the phylogeny of Eukaryotes[79]. For example, FIGURE 2b illustrates the utility of this approach in the phylogenomics of bilaterian animals. When using a complete phylogenomic dataset of almost 150 genes, the fast-evolving nematodes tend to be artefactually "attracted" by the distant fungal outgroup[23] (see BOX3b for an explanation). This illustrates that even a sophisticated phylogenetic method such as ML can be misled by the bias introduced by differences in evolutionary rates among species. The objective exclusion of the noisiest characters using the SF method led to an alternative topology, in which the long branch of nematodes is no longer grouped with the long branch of fungi.

Phylogenomic datasets offer the luxury of focusing solely on the more reliable characters using the methods described above. Indeed, when 100,000 characters are available, removing the 20-30% fastest evolving is unlikely to alter the statistical significance of the results, as would be the case using single-gene datasets (e.g. REF. [79]). At present, these types of approaches[52,54,142,145,147,149] might be the only way to handle cases where the presence of several confounding factors misleads even the most accurate methods of phylogenetic inference.

Phylogenomics and corroboration. The congruence of results obtained from various datasets and/or various methods is the key validation of evolutionary inferences[150]. To corroborate results, single-gene phylogenies have been compared to classical

morphological and ultrastructural studies, and subsequent multigene phylogenies were generally contrasted with previously obtained molecular trees. However, whole genomes represent the ultimate source of characters from which the evolutionary history of organisms can be reconstructed; therefore, how can we corroborate phylogenomic results?

The similarity between phylogenomic results and the SSU rRNA tree of Prokaryotes has been viewed as a first validation of the new large-scale approaches. Eventually, corroborating the results of different large-scale approaches should become a standard method of validation (e.g. REF. [56,151]). It is therefore desirable that methods based on whole-genome features should become as sophisticated and accurate as sequence-based methods. This necessitates a better understanding of the processes driving genome evolution, which could be achieved by comparing genomes from closely related taxa. With better methods and the use of the more reliable characters, rendering inconsistency less likely, the definitive proof of corroboration of phylogenomic results will certainly be their robustness to varying the species sampling. This will enable verification that the same results are obtained with different subsets of species. Corroboration in phylogenomics is a necessary prerequisite for tackling the large-scale resolution of the tree of life.

Perspectives — towards a fully resolved Tree of Life?

Recently, concerted efforts have been made towards realising Darwin's dream of having "fairly true genealogical trees for each great kingdom of Nature" in the form of collaborative network initiatives for assembling the tree of life. Several phylogenomic research programs (see ONLINE LINKS) targeting various groups — such as eukaryotes, fungi, arthropods,

nematodes, dipterans, or birds — will soon lead to important improvements through the consideration of a dense taxon sampling for these groups. Moreover, continuous progress from **METAGENOMICS** will continue to reveal the extent of microbial diversity[152].

At first sight, these efforts should ultimately lead to a fully resolved tree of life. However, as shown in FIGURE 3, not all nodes of a phylogenetic tree are equal with respect to the increase in resolving power provided by the phylogenomic approach. This is expected, as the resolution of phylogenetic trees ultimately depends on the evolutionary pattern by which organisms diversified. If time intervals between speciations were particularly short, it is likely that even complete genome data might not provide enough characters to accurately resolve certain nodes, for which hardly any phylogenetic signal will be recovered[153]. Furthermore, given the importance of taxon sampling for phylogenetic inference, it is possible that isolated species that are the sole living member of a particular group (e.g. the coelacanth, the tuatara and *Amborella*), or groups for which sampling is naturally scarce with only few representative extant species (e.g. monotremes), will prove difficult to position with confidence. Therefore, there will be nodes that are likely to be left unresolved because of the very nature of the evolutionary process, and this will in itself tell us a lot about the evolution of organisms. However, despite the power of phylogenomics to reveal evolutionary relationships, we might have to accept the idea of a partially resolved tree of life.

Finally, the reconstruction of the topology of the organismal phylogeny is not in itself the ultimate goal. The challenge is to understand the evolutionary history of organisms and their genomes, the functions of their genes, and how this relates to their interactions with the environment. Assembling the tree of life represents the first step towards achieving the big picture of phylogenomics where, to paraphrase Theodosius Dobzhansky[154], “nothing in genomics makes sense except in the light of evolution”.

Display Items

Box 1 | **Basic principles and methods of phylogenetic inference**

Phylogenetic inference involves two crucial steps: first, homologous characters (those that are descended from a common ancestor) are identified among species; second, the evolutionary history of species is reconstructed from the comparison of these characters using tree-building methods for phylogenetic inference. Almost any kind of character (for example, morphological structures, ultrastructural characteristics of cells, biochemical pathways, genes, amino-acids or nucleotides) can be used for inferring phylogenies, provided that they are homologous. In sequence data, homology is determined by similarity searching. Once homologous characters are identified, a character matrix is constructed, which scores the different character states (columns on the matrix) observed in each species (rows on the matrix).

Three main kinds of reconstruction method can then be used to infer phylogenetic trees from this character matrix as follows (see REF. [10] for an overview and REF. [34] for details):

Distance methods. These methods first convert the character matrix into a distance matrix that represents the evolutionary distances between all pairs of species. The phylogenetic tree is then inferred from this distance matrix using algorithms such as Neighbour-Joining (NJ)[155] or Minimum Evolution (ME)[156].

Maximum Parsimony. This method selects the tree that requires the minimum number of changes to explain the observed data.

Likelihood methods. These methods are based on a function that calculates the probability that a given tree could have produced the observed data (i.e. the likelihood). This function allows the explicit incorporation of the processes of character evolution into probabilistic models. Maximum Likelihood[157] (ML) selects the tree that maximises the probability of observing the data under a given model. Bayesian methods[106] derive the distribution of trees according to their posterior probability, using Bayes' mathematical formula to combine the likelihood function with prior probabilities on trees. Unlike ML, which optimises model parameters by finding the highest peak in the parameter space, Bayesian approaches integrate out model parameters (see REF. [10]).

Box 2 | **Methods of phylogenomic inference**

The flowchart shows steps in the inference of evolutionary trees from genomic data. Genomic information is obtained by large scale sequencing of DNA. In general, sets of orthologous genes are then assembled from specific sets of species for phylogenetic analysis. This homology or orthology assessment is a crucial step that is almost always based on simple similarity comparisons (e.g. BLAST[158] searches). Most methods used for the subsequent reconstruction of phylogenetic trees are either sequence-based or are based on whole-genome features.

Sequence-based methods. These methods necessitate orthologous genes to be aligned using tools for multiple-sequence alignment (e.g. the CLUSTAL W program[159]) and the determination of unambiguously aligned positions (e.g. using GBLOCKS[160]). Once this critical step is achieved, two alternative approaches can be used to infer phylogenetic trees from the different gene alignments, which are usually of unequal lengths and contain different sets of species. The supermatrix approach involves analysing the concatenation

of individual genes, and non-overlapping taxa are coded as missing data. Likelihood-based reconstruction methods (Box 1) are particularly suited for the analysis of supermatrices. These methods take into account across-gene heterogeneity in evolutionary rates by using partitioned-likelihood models, which allow each gene to evolve under a different model[161]. Despite the increased number of additional parameters introduced by using an independent model for each gene, these partitioned models usually fit the data better than concatenated models[17,20,162]. Alternatively, the supertree approach[37] combines the optimal trees obtained from the analysis of individual genes, each of which contain data from only partially overlapping sets of taxa.

Methods based on whole-genome features. These methods infer phylogenetic trees from the comparison of gene content (also known as gene repertoire), gene order, and "DNA strings". Gene-content methods reconstruct phylogenetic trees from 'distances', which represent the proportion of shared orthologous genes between genomes using classical distance algorithms[50,51,55], or from matrices, which score the presence or absence of homologues or orthologues in genomes using maximum parsimony[56,57]. Gene-order methods construct phylogenetic trees by minimizing the number of **BREAKPOINTS** between genomes[65], or simply by scoring the presence or absence of pairs of orthologous genes[53,56]. Methods based on the distribution of "DNA strings", which do not rely on homology assessment, can also be used[55,73-75]. These are based on oligonucleotide 'word usage' (the frequency of short-oligonucleotide combinations), which provides a characteristic signature of genome structure[72]. The few approaches that are currently implemented for this method calculate evolutionary distances among species from the difference in their oligonucleotide word usage, and reconstruct phylogenetic trees using standard distance-based algorithms[55,73-75].

Rare genomic changes. Rare genomic changes[7,8] — such as insertions and deletions (indels), intron positions, retroposon (SINE and LINE) integrations, and gene fusion and fission events — can be used as signatures supporting particular nodes, and to reconstruct phylogenetic trees based on their presence or absence.

Box 3 | **Inconsistency and its causes**

A phylogenetic reconstruction method is statistically inconsistent if it converges towards supporting an incorrect solution as more data are analysed. This happens when the assumptions made by the methods about the sequence evolutionary process are violated by the data properties. The three main kinds of bias that are not efficiently handled by most current reconstruction methods are known to be responsible for inconsistency.

Compositional bias. Similar nucleotide composition can lead phylogenetic methods to artefactually group unrelated species together. As an illustration (**a**), a phylogenomic dataset of 127,026 nucleotide sites (106 genes) for 8 yeast species[19] was analysed using variable-length bootstrap analysis[163]. This method allows visualisation of the change in statistical support (expressed as bootstrap percentages, BP) for a particular phylogenetic hypothesis as the number of sites increases. In the left panel, maximum likelihood (ML) using a parameter-rich model (GTR+Γ+I) that accounts for substitution-rate heterogeneity among sites converges towards supporting a tree (1) that groups *Saccharomyces kudriavzevii* with S*. mikatae*, *S. cerevisiae* and *S. paradoxus* to the exclusion of *S. bayanus*. The statistical support for this tree reaches 100% when more than 10,000 sites are analysed (blue diamonds). By contrast, on the right panel, the distance-based Minimum Evolution (ME) method using the same model converges towards supporting an alternative tree (2) where *S. kudriavzevii* and *S. bayanus* are grouped together also with 100% BP support for more than 10,000 sites (orange squares). Note that these two

phylogenetic hypotheses are mutually exclusive since on both panels the support for the alternative solution converges towards 0. As the two trees cannot be both correct, one of the methods must be inconsistent. In this case, ME has been shown to be misled by the fact that *S. kudriavzevii* and *S. bayanus* share similar base compositions[144]. This illustrates that the increase in statistical support provided by phylogenomics does not always guarantee convergence on the correct tree. All calculations were done with PAUP*[164].

Long-branch attraction. Unrelated species sharing high evolutionary rates can be artefactually grouped together because most phylogenetic methods become inconsistent under these conditions[104]. As an example (**b**), a phylogenomic dataset encompassing 146 nuclear proteins (35,346 amino acid positions) was assembled to study the relationships between *Homo sapiens*, *Drosophila melanogaster*, the *Caenorhabditis elegans* and the *Saccharomyces cerevisiae*[23]. The ML tree obtained using a JTT+Γ model (which accounts for substitution-rate heterogeneity among sites for these four species) strongly groups *H. sapiens* and *D. melanogaster* together (BP = 100). This arrangement corresponds to the classical Coelomata hypothesis (in which arthropods are grouped with vertebrates). The same analysis including six additional outgroups (three fungi, two choanoflagellates and a cnidarian) results in a highly supported tree where *D. melanogaster* and *C. elegans* are grouped together (BP = 96). This corresponds to the Ecdysozoa hypothesis (in which arthropods are grouped with nematodes). In this case, ML is probably inconsistent for the 4-taxa dataset, as the 'long branch' of the *C. elegans* is attracted by the 'long branch' of *S. cerevisiae*, which is broken by additional outgroup species in the 10-taxa dataset[23,112]. All calculations were done with PHYML[165]. Numbers indicated above nodes correspond to bootstrap percentages and the scale bars represent the number of estimated substitutions per site.

Heterotachy. Heterotachy[166] refers to the variation in the evolutionary rate of a given position of a gene or protein through time. This phenomenon has been recently confirmed as an important process of sequence evolution[113], and can lead to phylogenetic reconstruction artefacts[109] in cases where unrelated taxa have converged in their proportions of invariable sites[167-169]. Unlike other types of bias, heterotachy does not leave any evident footprints in sequences[109], and therefore leads to insidious artefacts that are particularly difficult to detect[168,169].

Box 4 | **Phylogenomics and computational burden**

The problem of 'tree space'. Phylogenetic inference is a computationally demanding task, given the large number of trees that are possible — the 'tree space' — even when a relatively small number of species is considered. Indeed, the number of rooted trees (i.e. those for which the common ancestor of all the species included is known**)** for 10 species is 34,459,425, and for 50 species this number increases to more than 10^{75}. It is therefore impossible to look at all of these trees and assess which best explains the data. As a consequence, phylogenetic inference relies on various heuristic optimisation algorithms that explore only a subset of the possible trees[34], but this should not be done at the expense of the accuracy of tree reconstruction. Likelihood-based methods are computationally very demanding because they incorporate numerous parameters in their underlying models. However, by using numerical procedures, such as **MARKOV CHAIN MONTE CARLO** (MCMC) methods, the Bayesian approach allows the implementation of complex models while remaining computationally tractable[106].

Assessing confidence. Assessing the statistical confidence of phylogenetic trees adds another dimension to the computational burden. Indeed, using resampling procedures such as the non-parametric bootstrap[105] is very time-consuming because they involve

repeating the initial phylogenetic analysis multiple times. The computational burden is particularly high with maximum likelihood methods, but it can be considerably reduced by resampling the sitewise log-likelihoods (RELL bootstrap) instead of the original characters[170]. With Bayesian methods, the measure of confidence is less computationally demanding, since it is directly computed from the original data in the form of Bayesian posterior probabilities (PPs)[106]. Originally thought to be roughly equivalent to non-parametric bootstrap percentages (BPs)[106], PPs have been shown to appear consistently higher than BPs in a wide range of empirical studies (see REF. [171] and references therein) and to provide an overestimate of accuracy in a phylogenomic dataset[172]. In fact, the posterior probability of a tree represents the probability that the tree is correct, assuming that the model is correct[173]. However, Bayesian methods can be sensitive to model misspecification[173,174], whereas the ML non-parametric bootstrap method appears to be more robust[173]. As a consequence, the non-parametric bootstrap remains the method of choice for assessing confidence, particularly as its computation can easily be performed in parallel.

"Divide-and-conquer". Given the rate of sequencing, the size of sequence-based phylogenomic datasets will soon make phylogenetic analysis computationally problematic using classical methods. The resolution of large phylogenetic problems can be tackled by using "divide-and-conquer" strategies. These methods break the dataset down into smaller subsets (i.e. a fraction of the species), infer optimal trees for these subsets, then finally combine these trees into a larger tree. The first implementations of this strategy have been based on quartets including four species[175] and **DISK-COVERING METHODS** using larger species subsets are currently being developed[176]. The combination of the supermatrix and supertree approaches (BOX 2) might thus represent a solution for reconstructing

phylogenomic trees with thousands of species in order to eventually obtain a full picture of the tree of life[40,177].

Fig 1 | **Phylogenomics and the tree of life**

A schematic representation showing recent advances and future challenges of the phylogenomic approach for resolving the major branches of the tree of life. This tree aims at representing a consensus view on evolutionary relationships within the three domains Bacteria, Archaea, and Eukaryota with hypothetical relationships indicated as broken lines. Major branches that have been identified (purple) or confirmed (orange) by phylogenomics are indicated. Blue coloured broken lines underline putative phylogenetic hypotheses that have been suggested by phylogenomic studies and need further investigations. The major uncertainties where the phylogenomic approach might provide future answers are pinpointed by red dots. Note that most of the progress brought about by the phylogenomic approach have been realised at a smaller taxonomic scale within land plants and placental mammals within metazoans (see main text). The two well recognized endosymbiotic events involving bacteria that gave rise to eukaryotic organelles (mitochondria and chloroplasts) are indicated by arrows. Note however, that other horizontal gene transfers and gene duplication events are not represented in this organismal tree, although they do constitute important aspects of genome evolution.

Fig 2 | **"Garbage in, garbage out": Inconsistency and the use of reliable characters**

If the characters used in phylogenomics are unreliable, even the most accurate tree reconstruction method can fail. Therefore, methods focusing on the most reliable characters have been developed in order to reduce the impact of inconsistency.

a | **RY-coding.** This strategy considers only transversion events between the two families of bases in DNA (purines and pyrimidines). In the case of arthropod phylogeny based on the four most conserved mitochondrial genes (3,729 nucleotides)[141], the Maximum Likelihood (ML) criterion using the GTR+Γ+I model recovers an incorrect but strongly supported (BP = 90) result (left panel): the four ticks (*Varroa destructor*, *Ornithodoros moubata*, *Ixodes hexagonus* and *Rhipicephalus sanguineus*) are nested within the insects as a sister-group of bees (*Melipona bicolor* and *Apis mellifera*), instead of clustering with the other chelicerate represented by the horseshoe crab (*Limulus polyfemus*). Mitochondrial genomes of some ticks and bees have converged towards high proportions of AT residues, and this is not accounted for by the model which assumes base composition homogeneity. RY-coding (right panel) allows the recovery of the correct phylogeny supporting the monophyly of Chelicerates (BP = 63) when analysed under ML using the 2-state CF+Γ+I model[178]. Therefore the removal of half of the original information by RY-coding (i.e. the number of parsimony steps is roughly halved) has removed the inconsistency of ML. All calculations used PAUP*[164]. Photos are copyright Biodic 2003 (http://www.ulb.ac.be/sciences/biodic/index.html) and are used with permission from Pr. Louis De Vos.

b | **Slow/Fast (SF) method.** The misleading effect of the long-branch attraction (LBA) artefact[104] is here tackled using the SF method[147] for a phylogenomic dataset from animals and fungi[23]. Using this method, different subsets of the data (S0, S1,…, Sn) are constructed containing sites that have experienced a total number of substitutions equal or less than 0, 1,…, n within predefined monophyletic groups. The “evolution” of the phylogenetic signal for a particular hypothesis is then monitored as fast evolving sites are progressively removed from the original dataset. The ML analysis under the JTT+Γ model

of the S5 matrix yields a tree supporting the Coelomata hypothesis (arthropods + deuterostomes) with a bootstrap support of BP = 88. However, the progressive removal of fast evolving positions results in a decrease of the support for the Coelomata hypothesis and a concomitant increase in the support for the Ecdysozoa hypothesis (arthropods + nematodes). The support in favour of the new animal phylogeny (Ecdysozoa) is reaching BP = 91 with the 13,947 slowest evolving sites for which no substitutions occurred within each of the four groups (S0). The initial support observed in favour of Coelomata is attributable to the fastest evolving sites likely causing an LBA artefact with fast evolving nematodes being attracted by the distant fungal outgroup[23]. All calculations used PHYML[165].

Fig 3 | **Phylogenomics and the resolution of phylogenetic trees**

This figure illustrates the differential increase in resolution provided by phylogenomics for different internal nodes of a given phylogenetic tree. A Maximum Likelihood (ML) tree (not shown) was reconstructed for a phylogenomic dataset of 141 genes representing a total of 31,731 amino-acid sites for 35 Eukaryotes (Naïara Rodriguez-Ezpeleta, pers. com.). The level of statistical support expressed as bootstrap percentages (BP) for internal nodes observed in the ML tree was plotted as a function of the number of jackknife resampled sites[179]. As expected, the resolution increased with additional characters and four types of profiles can be defined with respect to the amount of data that is needed to define different nodes: (1) numerous nodes (green) can be resolved with a small number of genes and often a single one (BP > 95 reached for less than 2,500 sites); (2) the majority of nodes (blue) are resolved with multigene datasets including less than 10,000 sites as recently achieved in plants[11] or mammals[85]; (3) few nodes (orange) can only be resolved by a phylogenomic approach considering a large amount of characters (i.e. more than 100

genes[17,23]); and (4) very rare nodes (red) are virtually irresolvable because the extrapolated number of sites required to reach BP = 95 exceeds the number of homologous sites that can be extracted from complete genomes. In this last case, only the improvement of the species sampling or of the tree reconstruction method could possibly change the picture.

Glossary

HOMOLOGOUS CHARACTERS Homologous characters are those that are descended from a common ancestor.

NODE Nodes of phylogenetic trees represent taxonomic units. Internal nodes (or branches) refer to hypothetical ancestors whereas terminal nodes (or leaves) generally correspond to extant species.

INCONSISTENCY A phylogenetic reconstruction method is statistically inconsistent if it converges towards supporting an incorrect solution with increasing confidence as more data is analysed.

HOMOPLASY Identical character states (for example, the same nucleotide base in a DNA sequence) that are not the result of common ancestry (not homologous), but arose independently in different ancestors by convergent mutations.

CONVERGENCE The independent evolution of similar features (such as genes) in evolutionarily distinct lineages.

REVERSAL The independent reacquisition of the ancestral character state in a given evolutionary lineage.

HOMOLOGY Two sequences are homologous if they share a common ancestor.

ORTHOLOGY Two sequences are orthologous if they share a common ancestor and are separated by speciation.

HORIZONTAL GENE TRANSFER The transfer of genetic material between the genomes of two organisms, which usually belong to different species and which is not via parent-descendant routes.

PARALLEL GENE LOSS The independent loss of homologous genes in evolutionary distinct lineages.

SATURATION Mutational saturation occurs when multiple changes at a given position have randomised the genuine phylogenetic signal.

ROOT The root of a phylogenetic tree represents the common ancestor of all taxa represented in the tree. The position of the root is often determined using an outgroup taxon to determine the order of evolution in the group of taxa of interest.

MONOPHYLY Monophyletic taxa include all the species derived from a single common ancestor.

STOCHASTIC OR SAMPLING ERROR The error in phylogenetic estimates caused by the finite length of the sequence used in the inference. As the size of the sequences increases, the magnitude of the stochastic error decreases.

SYSTEMATIC ERROR The error in phylogenetic estimates due to the failure of the reconstruction method to fully account for the properties of the data.

BOOTSTRAP ANALYSIS A type of statistical analysis to test the reliability of certain branches in an evolutionary tree. The non-parametric bootstrap proceeds by re-sampling the original data, with replacement, to create a series of bootstrap samples of the same size as the original data. The bootstrap percentage of a node is the proportion of times that a node is present in the set of trees that is constructed from the new data sets.

BAYESIAN POSTERIOR PROBABILITY In Bayesian phylogenetics, the posterior probability of a particular node of a tree is the probability that the node is correct, which is conditional on the data and the model used in the analysis both being correct.

COVARION MODEL OF MOLECULAR EVOLUTION In this model although some sites in a macromolecule are critical to function and can never change through time, most switch between being free to evolve in some species and being invariable in others.

METAGENOMICS The functional and sequence-based analysis of the collective microbial genomes contained in an environmental sample of uncultured organisms.

HEURISTIC A method of inference that relies on educated guesses or simplifications that limit the parameter space over which solutions are searched. This approach is not guaranteed to find the correct answer.

BREAKPOINT In the context of phylogenetic methods based on gene-order comparison between genomes, a breakpoint is defined when a pair of genes are adjacent in one genome but not in the other.

MARKOV CHAIN MONTE CARLO A computational technique for the efficient numerical calculation of likelihoods.

DISK-COVERING METHODS A family of divide-and-conquer algorithmic methods for large-scale tree reconstruction. These methods use graph theory to identify optimal decompositions of the input dataset into small overlapping sets of closely related species, reconstruct phylogenetic trees on these subsets, and then combine the subtrees into one tree including the entire set of species.

References

1. Darwin, C. *The origin of species by means of natural selection* (Murray, London, 1859).

2. Haeckel, E. *Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie* (Georg Reimer, Berlin, 1866).

3. Van Niel, C. B. in *Perspectives and Horizons in Microbiology* 3-12 (Rutgers University Press, New Brunswick, 1955).

4. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**, 357-366 (1965).

5. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. USA* **74**, 5088-5090 (1977).

6. Eisen, J. A. & Fraser, C. M. Phylogenomics: intersection of evolution and genomics. *Science* **300**, 1706-1707 (2003).

7. Philippe, H. & Laurent, J. How good are deep phylogenetic trees? *Curr. Opin. Genet. Dev.* **8**, 616-623 (1998).

8. Rokas, A. & Holland, P. W. Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* **15**, 454-459 (2000).

9. Gribaldo, S. & Philippe, H. Ancient phylogenetic relationships. *Theor. Popul. Biol.* **61**, 391-408 (2002).

10. Holder, M. & Lewis, P. O. Phylogeny estimation: Traditional and Bayesian approaches. *Nat. Rev. Genet.* **4**, 275-284 (2003).

11. Qiu, Y. L. et al. The earliest angiosperms: Evidence from mitochondrial, plastid and nuclear genomes. *Nature* **402**, 404-407 (1999).

12. Moreira, D., Le Guyader, H. & Philippe, H. The origin of red algae: Implications for the evolution of chloroplasts. *Nature* **405**, 69-72 (2000).

13. Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290**, 972-977 (2000).

14. Madsen, O. et al. Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**, 610-614 (2001).

15. Murphy, W. J. et al. Molecular phylogenetics and the origins of placental mammals. *Nature* **409**, 614-618 (2001).

16. Blair, J. E., Ikeo, K., Gojobori, T. & Hedges, S. B. The evolutionary position of nematodes. *BMC Evol. Biol.* **2**, 7 (2002).

17. Bapteste, E. et al. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc. Natl. Acad. Sci. USA* **99**, 1414-1419 (2002).
The first phylogenomic study based on the supermatrix approach including more than 100 genes for a relatively broad taxon sampling of eukaryotes.

18. Lerat, E., Daubin, V. & Moran, N. A. From gene trees to organismal phylogeny in Prokaryotes: The case of the γ-Proteobacteria. *PLoS Biol.* **1**, e19 (2003).

19. Rokas, A., Williams, B. L., King, N. & Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798-804 (2003).

An empirical study on the phylogenomics of yeasts showing that, for the same number of positions, a robust phylogenetic tree is recovered more rapidly with randomly selected positions than with entire genes.

20. Philippe, H. et al. Phylogenomics of eukaryotes: Impact of missing data on large alignments. *Mol. Biol. Evol.* **21**, 1740-1752 (2004).

21. Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. Coelomata and not Ecdysozoa: Evidence from genome-wide phylogenetic analysis. *Genome Res.* **14**, 29-36 (2004).

22. Driskell, A. C. et al. Prospects for building the tree of life from large sequence databases. *Science* **306**, 1172-1174 (2004).

References 20 and 22 demonstrate the robustness of the supermatrix approach to a surprisingly high amount of missing data in phylogenomic analyses.

23. Philippe, H., Lartillot, N. & Brinkmann, H. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa and Protostomia. *Mol. Biol. Evol.* **published online Feb 9** (2005).

This study demonstrates the impact of the long-branch attraction artefact in phylogenomics and provides evidence for the new animal phylogeny based on a relatively large species sampling.

24. Lecointre, G., Philippe, H., Le, H. L. V. & Le Guyader, H. Species sampling has a major impact on phylogenetic inference. *Mol. Phylogenet. Evol.* **2**, 205-224 (1993).

25. Graybeal, A. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **47**, 9-17 (1998).

26. Poe, S. & Swofford, D. L. Taxon sampling revisited. *Nature* **398**, 299-300 (1999).

27. Hillis, D. M., Pollock, D. D., McGuire, J. A. & Zwickl, D. J. Is sparse taxon sampling a problem for phylogenetic inference? *Syst. Biol.* **52**, 124-126 (2003).

28. Rosenberg, M. S. & Kumar, S. Taxon sampling, bioinformatics, and phylogenomics. *Syst. Biol.* **52**, 119-124 (2003).

References 27 and 28 present a recent exchange on the relative importance of character and taxon sampling for phylogenetic inference.

29. Philippe, H. Rodent monophyly: Pitfalls of molecular phylogenies. *J. Mol. Evol.* **45**, 712-715 (1997).

30. Lin, Y.-H. et al. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. *Mol. Biol. Evol.* **19**, 2060-2070 (2002).

31. Philip, G. K., Creevey, C. J. & McInerney, J. O. The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. *Mol. Biol. Evol.* **Published online Feb 9** (2005).

32. Sanderson, M. J., Driskell, A. C., Ree, R. H., Eulenstein, O. & Langley, S. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. *Mol. Biol. Evol.* **20**, 1036-1042 (2003).

33. Kluge, A. G. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Syst. Zool.* **38**, 7-25 (1989).

34. Felsenstein, J. *Inferring phylogenies* (Sinauer Associates, Inc., Sunderland, MA, USA, 2004).

35. Gatesy, J., Matthee, C., DeSalle, R. & Hayashi, C. Resolution of a supertree/supermatrix paradox. *Syst. Biol.* **51**, 652-664 (2002).

36. Wiens, J. J. Missing data, incomplete taxa, and phylogenetic accuracy. *Syst. Biol.* **52**, 528-538 (2003).

37. Bininda-Emonds, O. R. P., Gittleman, J. L. & Steel, M. A. The (Super)tree of life: Procedures, problems, and prospects. *Annu. Rev. Ecol. Syst.* **33**, 265-289 (2002).

38. Baum, B. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* **41**, 3-10 (1992).

39. Ragan, M. A. Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.* **1**, 53-58 (1992).

40. Bininda-Emonds, O. R. P. The evolution of supertrees. *Trends Ecol. Evol.* **19**, 315-322 (2004).

41. Liu, F. G. et al. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* **291**, 1786-1789 (2001).

42. Daubin, V., Gouy, M. & Perrière, G. A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. *Genome Res.* **12**, 1080-1090 (2002).

The first application of a supertree method in phylogenomics showing its utility for reconstructing bacterial phylogeny in the presence of horizontal gene transfer.

43. Gatesy, J., Baker, R. H. & Hayashi, C. Inconsistencies in arguments for the supertree approach: Supermatrices versus supertrees of Crocodylia. *Syst. Biol.* **53**, 342-355 (2004).

44. Salamin, N., Hodkinson, T. R. & Savolainen, V. Building supertrees: An empirical assessment using the grass family (Poaceae). *Syst. Biol.* **51**, 136-150 (2002).

45. Bininda-Emonds, O. R. P. Trees versus characters and the supertree/supermatrix "paradox". *Syst. Biol.* **53**, 356-359 (2004).

46. Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. Universal trees based on large combined protein sequence data sets. *Nat. Genet.* **28**, 281-285 (2001).

47. Brochier, C., Bapteste, E., Moreira, D. & Philippe, H. Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet.* **18**, 1-5 (2002).
A comprehensive study of bacterial phylogeny based on the supermatrix approach using statistical methods to detect and exclude genes likely affected by horizontal transfer.

48. Yang, Z. On the best evolutionary rate for phylogenetic analysis. *Syst. Biol.* **47**, 125-133 (1998).

49. Wolf, Y. I., Rogozin, I. B., Grishin, N. V. & Koonin, E. V. Genome trees and the tree of life. *Trends Genet.* **18**, 472-479 (2002).

50. Snel, B., Bork, P. & Huynen, M. A. Genome phylogeny based on gene content. *Nat. Genet.* **21**, 108-110 (1999).

51. Tekaia, F., Lazcano, A. & Dujon, B. The genomic tree as revealed from whole proteome comparisons. *Genome Res.* **9**, 550-557 (1999).

52. Clarke, G. D., Beiko, R. G., Ragan, M. A. & Charlebois, R. L. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a

distance matrix based on mean normalized BLASTP scores. *J. Bacteriol.* **184**, 2072-2080 (2002).

53. Korbel, J. O., Snel, B., Huynen, M. A. & Bork, P. SHOT: A web server for the construction of genome phylogenies. *Trends Genet.* **18**, 158-162 (2002).
This paper presents reconstruction of prokaryotic phylogenies based on gene content and the conservation of gene pairs with a critical view on the impact of horizontal gene transfer on their accuracy.

54. Dutilh, B. E., Huynen, M. A., Bruno, W. J. & Snel, B. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J. Mol. Evol.* **58**, 527-539 (2004).

55. Lin, J. & Gerstein, M. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res.* **10**, 808-818 (2000).

56. Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. & Koonin, E. V. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1**, 8 (2001).
A study of bacterial phylogenomics using five independent reconstruction methods to corroborate the emergence of a recurrent phylogenetic pattern.

57. Fitz-Gibbon, S. T. & House, C. H. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* **27**, 4218-4222 (1999).

58. House, C. H. & Fitz-Gibbon, S. T. Using homolog groups to create a whole-genomic tree of free-living organisms: An update. *J. Mol. Evol.* **54**, 539-547 (2002).

59. House, C. H., Runnegar, B. & Fitz-Gibbon, S. T. Geobiological analysis using whole genome-based tree building applied to the Bacteria, Archaea, and Eukarya. *Geobiology* **1**, 15-26 (2003).

60. Lake, J. A. & Rivera, M. C. Deriving the genomic tree of life in the presence of horizontal gene transfer: Conditioned reconstruction. *Mol. Biol. Evol.* **21**, 681-690 (2004).

61. Gu, X. & Zhang, H. Genome phylogenetic analysis based on extended gene contents. *Mol. Biol. Evol.* **21**, 1401-1408 (2004).

62. Huson, D. H. & Steel, M. Phylogenetic trees based on gene content. *Bioinformatics* **20**, 2044-2049 (2004).

63. Sankoff, D. et al. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. USA* **89**, 6575-6579 (1992).

64. Hannenhalli, S. & Pevzner, P. A. Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. *J. ACM* **46**, 1-27 (1999).

65. Blanchette, M., Kunisawa, T. & Sankoff, D. Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.* **49**, 193-203 (1999).

66. Moret, B., Tang, J. & Warnow, T. in *Mathematics of Evolution and Phylogeny* (ed. Gascuel, O.) 321-352 (Oxford Univ. Press, Oxford, 2005).

67. Koski, L. B. & Golding, G. B. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **52**, 540-542 (2001).

68. Philippe, H. & Douady, C. J. Horizontal gene transfer and phylogenetics. *Curr. Opin. Microbiol.* **6**, 498-505 (2003).

69. Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99-113 (1970).

70. Stanhope, M. J. et al. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* **411**, 940-944 (2001).

71. Sicheritz-Ponten, T. & Andersson, S. G. A phylogenomic approach to microbial evolution. *Nucleic Acids Res.* **29**, 545-552 (2001).

72. Campbell, A., Mrazek, J. & Karlin, S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **96**, 9184-9189 (1999).

73. Edwards, S. V., Fertil, B., Giron, A. & Deschavanne, P. J. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst. Biol.* **51**, 599-613 (2002).

74. Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* **13**, 145-158 (2003).

A study demonstrating that phylogenetic signal can be retrieved from the distribution of oligonucleotides in prokaryote genomes.

75. Qi, J., Wang, B. & Hao, B. I. Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. *J. Mol. Evol.* **58**, 1-11 (2004).

76. Nikaido, M., Rooney, A. P. & Okada, N. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements:

hippopotamuses are the closest extant relatives of whales. *Proc. Natl. Acad. Sci. USA* **96**, 10261-10266 (1999).

77. van Dijk, M. A. et al. Protein sequence signatures support the African clade of mammals. *Proc. Natl. Acad. Sci. USA* **98**, 188-193 (2001).

78. Venkatesh, B., Erdmann, M. V. & Brenner, S. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. *Proc. Natl. Acad. Sci. USA* **98**, 11382-11387 (2001).

79. Philippe, H. et al. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proc. R. Soc. Lond. B Biol. Sci.* **267**, 1213-1221 (2000).

80. Stechmann, A. & Cavalier-Smith, T. Rooting the eukaryote tree by using a derived gene fusion. *Science* **297**, 89-91 (2002).

81. Snel, B., Bork, P. & Huynen, M. Genome evolution. Gene fusion versus gene fission. *Trends Genet.* **16**, 9-11 (2000).

82. Bapteste, E. & Philippe, H. The potential value of indels as phylogenetic markers: Position of trichomonads as a case study. *Mol. Biol. Evol.* **19**, 972-977 (2002).

83. Krzywinski, J. & Besansky, N. J. Frequent intron loss in the white gene: A cautionary tale for phylogeneticists. *Mol. Biol. Evol.* **19**, 362-366 (2002).

84. Pecon-Slattery, J., Pearks Wilkerson, A. J., Murphy, W. J. & O'Brien S, J. Phylogenetic assessment of introns and SINEs within the Y chromosome using the cat family Felidae as a species tree. *Mol. Biol. Evol.* **21**, 2299-2309 (2004).

85. Murphy, W. J. et al. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348-2351 (2001).

86. Amrine-Madsen, H., Koepfli, K. P., Wayne, R. K. & Springer, M. S. A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. *Mol. Phylogenet. Evol.* **28**, 225-240 (2003).

87. Reyes, A. et al. Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods. *Mol. Biol. Evol.* **21**, 397-403 (2004).

88. Soltis, P. S., Soltis, D. E. & Chase, M. W. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* **402**, 402-404 (1999).

89. Barkman, T. J. et al. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. *Proc. Natl. Acad. Sci. USA* **97**, 13166-13171 (2000).

90. Pryer, K. M. et al. Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. *Nature* **409**, 618-622 (2001).

91. Soltis, D. E., Soltis, P. S. & Zanis, M. J. Phylogeny of seed plants based on evidence from eight genes. *Am. J. Bot.* **89**, 1670-1681 (2002).

92. Zanis, M. J., Soltis, D. E., Soltis, P. S., Mathews, S. & Donoghue, M. J. The root of the angiosperms revisited. *Proc. Natl. Acad. Sci. USA* **99**, 6848-6853 (2002).

93. Savolainen, V. & Chase, M. W. A decade of progress in plant molecular phylogenetics. *Trends Genet.* **19**, 717-724 (2003).

94. King, N. & Carroll, S. B. A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution. *Proc. Natl. Acad. Sci. USA* **98**, 15032-15037 (2001).

95. Lang, B. F., O'Kelly, C., Nerad, T., Gray, M. W. & Burger, G. The closest unicellular relatives of animals. *Curr. Biol.* **12**, 1773-1778 (2002).

96. Simpson, A. G. & Roger, A. J. The real 'kingdoms' of eukaryotes. *Curr. Biol.* **14**, R693-R696 (2004).

97. Rivera, M. C. & Lake, J. A. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**, 152-155 (2004).

98. Esser, C. et al. A genome phylogeny for mitochondria among α-Proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* **21**, 1643-1660 (2004).

99. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51**, 221-271 (1987).

100. Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124-2129 (1999).

101. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304 (2000).

102. Yang, S., Doolittle, R. F. & Bourne, P. E. Phylogeny determined by protein domain content. *Proc. Natl. Acad. Sci. USA* **102**, 373-378 (2005).

103. Matte-Tailliez, O., Brochier, C., Forterre, P. & Philippe, H. Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* **19**, 631-639 (2002).

104. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401-410 (1978).

105. Felsenstein, J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783-791 (1985).

106. Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310-2314 (2001).

107. Huelsenbeck, J. P. Performance of phylogenetic methods in simulation. *Syst. Biol.* **44**, 17-48 (1995).

108. Swofford, D. L. et al. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* **50**, 525-539 (2001).

109. Kolaczkowski, B. & Thornton, J. W. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**, 980-984 (2004).

A simulation study showing that the performance of current likelihood-based methods of phylogenetic reconstruction are noticeably affected by heterotachy.

110. Whelan, S., Lio, P. & Goldman, N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.* **17**, 262-272 (2001).

111. Steel, M. A., Lockhart, P. J. & Penny, D. Confidence in evolutionary trees from biological sequence data. *Nature* **364**, 440-442 (1993).

112. Hendy, M. & Penny, D. A framework for the quantitative study of evolutionary trees. *Syst. Zool.* **38**, 297-309 (1989).

113. Lopez, P., Casane, D. & Philippe, H. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19**, 1-7 (2002).

114. Mathews, S. & Donoghue, M. J. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* **286**, 947-950 (1999).

115. Goremykin, V. V., Hirsch-Ernst, K. I., Wolfl, S. & Hellwig, F. H. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that amborella is not a basal angiosperm. *Mol. Biol. Evol.* **20**, 1499-1505 (2003).

116. Goremykin, V. V., Hirsch-Ernst, K. I., Wolfl, S. & Hellwig, F. H. The chloroplast genome of *Nymphaea alba*: Whole-genome analyses and the problem of identifying the most basal angiosperm. *Mol. Biol. Evol.* **21**, 1445-1454 (2004).

117. Soltis, D. E. et al. Genome-scale data, angiosperm relationships, and "ending incongruence": A cautionary tale in phylogenetics. *Trends Plant Sci.* **9**, 477-483 (2004).

118. Stefanovic, S., Rice, D. W. & Palmer, J. D. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evol Biol* **4**, 35 (2004).

119. Adoutte, A. et al. The new animal phylogeny: Reliability and implications. *Proc. Natl. Acad. Sci. USA* **97**, 4453-4456 (2000).

120. Halanych, K. M. The new view of animal phylogeny. *Annu. Rev. Ecol. Evol. Syst.* **35**, 229-256 (2004).

121. Aguinaldo, A. M. et al. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**, 489-493 (1997).

122. Dopazo, H., Santoyo, J. & Dopazo, J. Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. *Bioinformatics* **20 Suppl 1**, I116-I121 (2004).

123. Keeling, P. J. & Fast, N. M. Microsporidia: Biology and evolution of highly reduced intracellular parasites. *Annu. Rev. Microbiol.* **56**, 93-116 (2002).

124. Sullivan, J. & Swofford, D. L. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Syst. Biol.* **50**, 723-729 (2001).

125. Huelsenbeck, J. P. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Mol. Biol. Evol.* **12**, 843-849 (1995).

126. Gaut, B. S. & Lewis, P. O. Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.* **12**, 152-162 (1995).

127. Siepel, A. & Haussler, D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21**, 413-428 (2004).

128. Whelan, S. & Goldman, N. Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics* **167**, 2027-2043 (2004).

129. Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N. & Thorne, J. L. Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.* **28**, 1692-1704 (2003).

130. Rodrigue, N., Lartillot, N., Bryant, D. & Philippe, H. Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, in press (2005).

131. Galtier, N. & Gouy, M. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* **15**, 871-879 (1998).

132. Foster, P. G. Modeling compositional heterogeneity. *Syst. Biol.* **53**, 485-495 (2004).

133. Fitch, W. M. Rate of change of concomitantly variable codons. *J. Mol. Evol.* **1**, 84-96 (1971).

134. Tuffley, C. & Steel, M. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* **147**, 63-91 (1998).

135. Penny, D., McComish, B. J., Charleston, M. A. & Hendy, M. D. Mathematical elegance with biochemical realism: The covarion model of molecular evolution. *J. Mol. Evol.* **53**, 711-723 (2001).

136. Galtier, N. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.* **18**, 866-873 (2001).

137. Huelsenbeck, J. P. Testing a covariotide model of DNA substitution. *Mol. Biol. Evol.* **19**, 698-707 (2002).

138. 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).

139. Pagel, M. & Meade, A. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.* **53**, 571-581 (2004).

References 138 and 139 explore promising mixture models to handle sequences that evolved under heterogeneous conditions.

140. Woese, C. R., Achenbach, L., Rouviere, P. & Mandelco, L. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. *Syst. Appl. Microbiol.* **14**, 364-371 (1991).

141. Delsuc, F., Phillips, M. J. & Penny, D. Comment on "Hexapod origins: Monophyletic or paraphyletic?" *Science* **301**, 1482 (2003).

142. Phillips, M. J. & Penny, D. The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol. Phylogenet. Evol.* **28**, 171-185 (2003).

143. Gibson, A., Gowri-Shankar, V., Higgs, P. G. & Rattray, M. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. *Mol. Biol. Evol.* **22**, 251-264 (2005).

144. Phillips, M. J., Delsuc, F. & Penny, D. Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* **21**, 1455-1458 (2004).
A cautionary tale for phylogenomic studies from the empirical demonstration that compositional bias can lead to inconsistency of some distance methods.

145. Lopez, P., Forterre, P. & Philippe, H. The root of the tree of life in the light of the covarion model. *J. Mol. Evol.* **49**, 496-508 (1999).

146. Ruiz-Trillo, I., Riutort, M., Littlewood, D. T. J., Herniou, E. A. & Baguna, J. Acoel flatworms: Earliest extant bilaterian metazoans, not members of Platyhelminthes. *Science* **283**, 1919-1923 (1999).

147. Brinkmann, H. & Philippe, H. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol.* **16**, 817-825 (1999).

148. Burleigh, J. G. & Mathews, S. Phylogenetic signal in nucleotide data from seed plants: Implications for resolving the seed plant tree of life. *Am. J. Bot.* **91**, 1599-1613 (2004).

149. Pisani, D. Identifying and removing fast-evolving sites using compatibility analysis: An example from the arthropoda. *Syst. Biol.* **53**, 978-989 (2004).

150. Miyamoto, M. M. & Fitch, W. M. Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.* **44**, 64-76 (1995).

151. Herniou, E. A. et al. Use of whole genome sequence data to infer baculovirus phylogeny. *J. Virol.* **75**, 8117-8126 (2001).

152. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: Genomic analysis of microbial communities. *Annu. Rev. Genet.* **38**, 525-552 (2004).

153. Philippe, H., Chenuil, A. & Adoutte, A. Can the cambrian explosion be inferred through molecular phylogeny? *Development* **120**, S15-S25 (1994).

154. Dobzhansky, T. Nothing in biology makes sense except in the light of evolution. *Am. Biol. Teacher* **35**, 125-129 (1973).

155. Saitou, N. & Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425 (1987).

156. Rzhetsky, A. & Nei, M. Statistical properties of the ordinary least-squares, generalized least- squares, and minimum-evolution methods of phylogenetic inference. *J. Mol. Evol.* **35**, 367-375 (1992).

157. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**, 368-76 (1981).

158. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).

159. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680 (1994).

160. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540-552 (2000).

161. Yang, Z. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42**, 587-596 (1996).

162. Pupko, T., Huchon, D., Cao, Y., Okada, N. & Hasegawa, M. Combining multiple data sets in a likelihood analysis: which models are the best? *Mol. Biol. Evol.* **19**, 2294-2307 (2002).

163. Springer, M. S., Amrine, H. M., Burk, A. & Stanhope, M. J. Additional support for Afrotheria and Paenungulata, the performance of mitochondrial versus nuclear genes, and the impact of data partitions with heterogeneous base composition. *Syst. Biol.* **48**, 65-75 (1999).

164. Swofford, D. L. PAUP*: Phylogenetic Analysis Using Parsimony and other methods (Sinauer, Sunderland, MA, 2002).

165. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696-704 (2003).

166. Philippe, H. & Lopez, P. On the conservation of protein sequences in evolution. *Trends Biochem. Sci.* **26**, 414-416 (2001).

167. Lockhart, P. J., Larkum, A. W., Steel, M., Waddell, P. J. & Penny, D. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. *Proc. Natl. Acad. Sci. USA* **93**, 1930-1934 (1996).

168. Philippe, H. & Germot, A. Phylogeny of eukaryotes based on ribosomal RNA: Long-branch attraction and models of sequence evolution. *Mol. Biol. Evol.* **17**, 830-834 (2000).

169. Inagaki, Y., Susko, E., Fast, N. M. & Roger, A. J. Covarion shifts cause a long-branch attraction artifact that unites Microsporidia and Archaebacteria in EF-1α phylogenies. *Mol. Biol. Evol.* **21**, 1340-1349 (2004).

170. Kishino, H., Miyata, T. & Hasegawa, M. Maximum likelihood inference of protein phylogeny, and the origin of chloroplasts. *J. Mol. Evol.* **31**, 151-160 (1990).

171. Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F. & Douzery, E. J. P. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* **20**, 248-254. (2003).

172. Taylor, D. J. & Piel, W. H. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. *Mol. Biol. Evol.* **21**, 1534-1537 (2004).

173. Huelsenbeck, J. P. & Rannala, B. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* **53**, 904-913 (2004).

174. Lemmon, A. R. & Moriarty, E. C. The importance of proper model assumption in bayesian phylogenetics. *Syst. Biol.* **53**, 265-277 (2004).

175. Strimmer, K. & von Haeseler, A. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964-969 (1996).

176. Roshan, U., Moret, B. M. E., Williams, T. L. & Warnow, T. Rec-I-DCM3: a fast algorithmic technique for reconstructing large phylogenetic trees. *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference* (2004).

177. Roshan, U., Moret, B. M. E., Williams, T. L. & Warnow, T. in *Phylogenetic supertrees: Combining information to reveal the Tree of Life* (ed. Bininda-Emonds, O. R. P.) 301-328 (Kluwer Academics, 2004).

178. Cavender, J. A. & Felsenstein, J. Invariants of phylogenies in a simple case with discrete states. *J. Classif.* **4**, 57-71 (1987).

179. Lecointre, G., Philippe, H., Le, H. L. V. & Le Guyader, H. How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. *Mol. Phylogenet. Evol.* **3**, 292-309 (1994).

Acknowledgments

We thank Nicolas Rodrigue, Naïara Rodriguez-Ezpeleta and Emmanuel Douzery for critical reading of early versions of the manuscript. Constructive comments from three anonymous referees also helped to make the manuscript more accurate. We apologise to our colleagues whose relevant work has not been cited because of space limitations. The authors gratefully acknowledge the financial support provided by Génome Québec, Canadian Research Chair and the Université de Montréal.

Online links

Genomes online: http://www.genomesonline.org/

NCBI National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov/

EMBL-EBI Nucleotide Sequence Database: http://www.ebi.ac.uk/embl/

The Tree of Life Web Project: http://tolweb.org/tree/

The Cyberinfrastructure for Phylogenetic Research (CIPRes) Project:

http://www.phylo.org/

The Protist EST Program (PEP): http://megasun.bch.umontreal.ca/pepdb/pep_main.html

Assembling the Tree of Eukaryotic Diversity (Eu-Tree):

http://www.biology.uiowa.edu/eu_tree/

Assembling the Fungal Tree of Life: http://ocid.nacse.org/research/aftol/

Higher-Level Arthropod Phylogenomics from the Cunnigham Lab:

http://www.biology.duke.edu/cunningham/DeepArthropod.html

The Nematode Genome Sequencing Center: http://www.nematode.net/

The Blaxter Lab Nematode Genomics Web Site: http://www.nematodes.org

Assembling the Dipteran Tree of Life (Fly-Tree):

http://www.inhs.uiuc.edu/cee/FLYTREE/

Assembling the Bird Tree of Life (Early Bird):

http://www.fieldmuseum.org/research_collections/zoology/zoo_sites/early_bird/

The Green Plant Phylogeny Research Coordination Group (DeepGreen):

http://ucjeps.berkeley.edu/bryolab/GPphylo/

The DeepTime Project: http://www.flmnh.ufl.edu/deeptime/

TreeBase: http://www.treebase.org/

Phylogeny Programs at Joe Felsenstein's Web Page:

http://evolution.genetics.washington.edu/phylip/software.html

The PhyCom Phylogenetic Community: http://www.yphy.org/phycom/

BIOGRAPHIES

Frédéric Delsuc studied evolutionary biology at the Université Montpellier II (France). His Ph.D. research was directed by Emmanuel Douzery. The focus of his dissertation was the application of probabilistic methods to the reconstruction of the evolutionary history of xenarthrans, one of the four major groups of placental mammals. His interest in phylogenetic methodology increased during a postdoctoral research position with David Penny at Massey University (New Zealand). Since 2004, he has been a postdoctoral researcher with Hervé Philippe at the Université de Montréal (Québec, Canada), working on biological applications and methodological aspects of the phylogenomic approach for reconstructing the tree of life.

Henner Brinkmann obtained his PhD in plant molecular biology at the Université Joseph Fourier (Grenoble, France). For several years he worked in the laboratory of Rüdiger Cerff on various aspects of the evolution of the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene system, with a special emphasis on the horizontally transferred genes of endosymbiotic origin present in the nucleus of all eukaryotic cells. Subsequently the focus of his research shifted towards prokaryotic evolution and the early evolution of life, including the relationships between the three domains. After an interlude dedicated to vertebrate evolution, especially the origin of tetrapods, he returned to study deep-level phylogenies, with a special emphasis on eukaryotic evolution in a phylogenomic framework, at the Université de Montréal in the laboratory of Hervé Philippe.

Hervé Philippe is professor of Bioinformatics and Evolutionary Genomics at the Centre Robert-Cedergren (Université de Montréal, Canada), and a fellow of the Evolutionary Biology program of the Canadian Institute for Advanced Research. He is interested in the general problem of methods for recovering evolutionary information from sequence data. This has led to many studies of the limitations and uses of sequence data. His specific interests are the nature of the last universal common ancestor (LUCA) and the functional information that can be extracted from the phylogenetic comparison of protein sequences.

Online Summary

- Understanding phylogenetic relationships among organisms is a prerequisite of evolutionary studies, as contemporary species all share a common history through their ancestry.
- The wealth of sequence data generated by large-scale genome projects is transforming phylogenetics — the reconstruction of evolutionary history — into phylogenomics.
- Traditional sequence-based methods of phylogenetic reconstruction (supermatrix and supertree approaches) can also be used at the genome level.
- New methods based on whole-genome features are also currently being developed to infer phylogenomic trees.
- Recent studies have revealed the potential of phylogenomic methods for answering longstanding phylogenetic questions.
- The supermatrix approach that analyses the concatenation of multiple gene sequences is the best-characterised method. Its potential relies on the increased resolving power provided by the use of a large number of sequence positions, which reduces the sampling error.
- Including large amount of data in phylogenomic analyses increases the possibility of obtaining highly supported but incorrect phylogenetic results due to inconsistency — that is, the convergence towards an incorrect solution as more data are added.
- Inconsistency arises because current phylogenetic reconstruction methods do not account for the full complexity of the molecular evolutionary process in their underlying assumptions.
- The risks of inconsistency in phylogenomics analyses can be reduced by the development of better models of sequence evolution, by the critically evaluation of data properties and by the use of only the most reliable characters.
- Corroboration of phylogenomic results is an important issue, as whole genomes represent the ultimate source of phylogenetically informative characters. Sources of corroboration

include the congruence of results obtained using different phylogenomic methods, and their robustness to taxon sampling.

- The very nature of the evolutionary process and the limitations of current phylogenetic reconstruction methods imply that parts of the tree of life may prove difficult, if not impossible, to resolve with confidence.

Figure 1

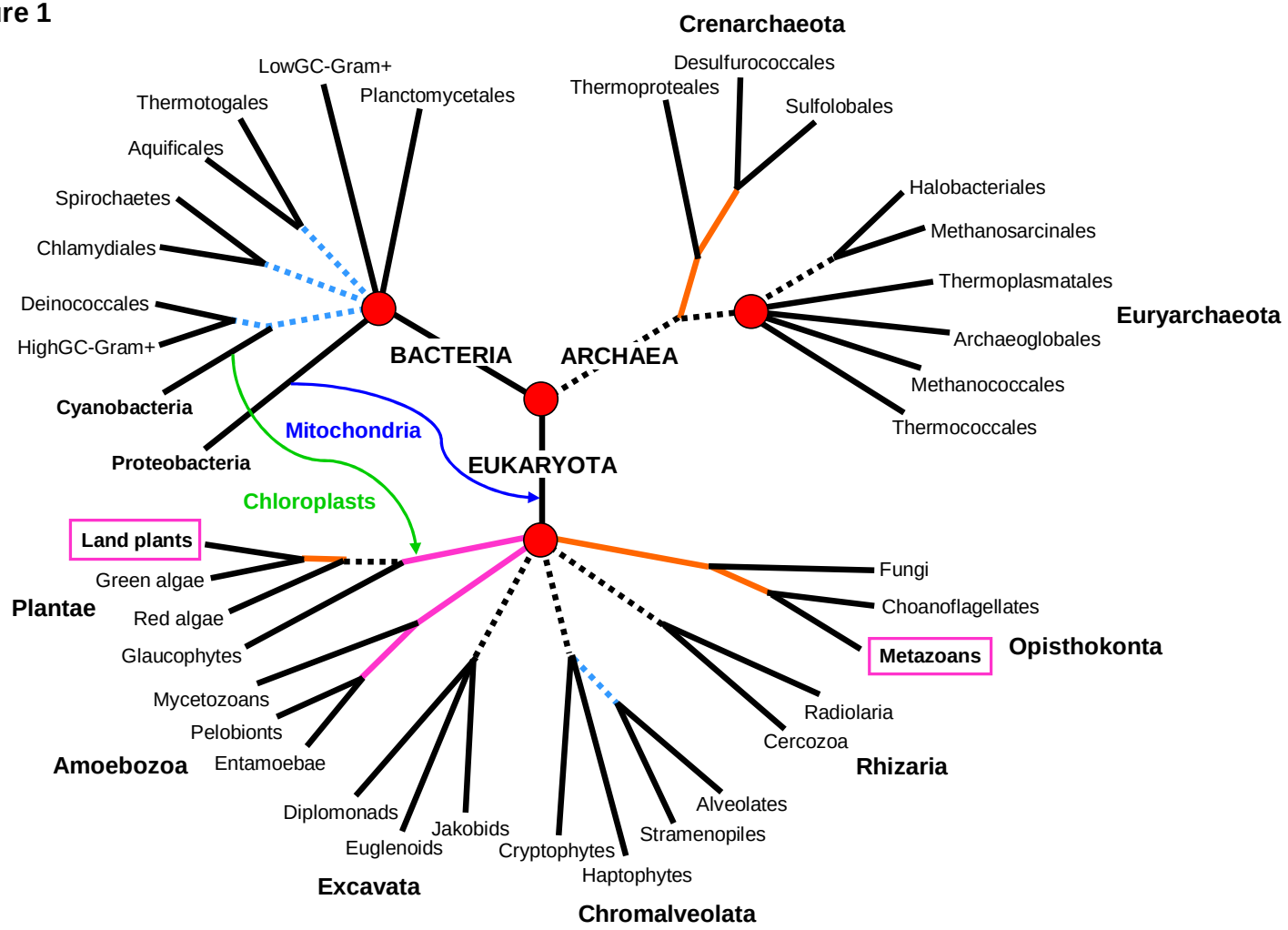

Crenarchaeota
Desulfurococcales
Thermoproteales
Sulfolobales
LowGC-Gram+
Planctomycetales
Thermotogales
Aquificales
Spirochaetes
Chlamydiales
Deinococcales
HighGC-Gram+
Halobacteriales
Methanosarcinales
Thermoplasmatales
Euryarchaeota
Archaeoglobales
Methanococcales
Thermococcales
BACTERIA
ARCHAEA
Cyanobacteria
Proteobacteria
Mitochondria
EUKARYOTA
Chloroplasts
Land plants
Green algae
Plantae
Red algae
Glaucophytes
Fungi
Choanoflagellates
Metazoans
Opisthokonta
Mycetozoans
Pelobionts
Entamoebae
Amoebozoa
Radiolaria
Cercozoa
Rhizaria
Diplomonads
Euglenoids
Jakobids
Excavata
Cryptophytes
Haptophytes
Alveolates
Stramenopiles
Chromalveolata

Figure 2a

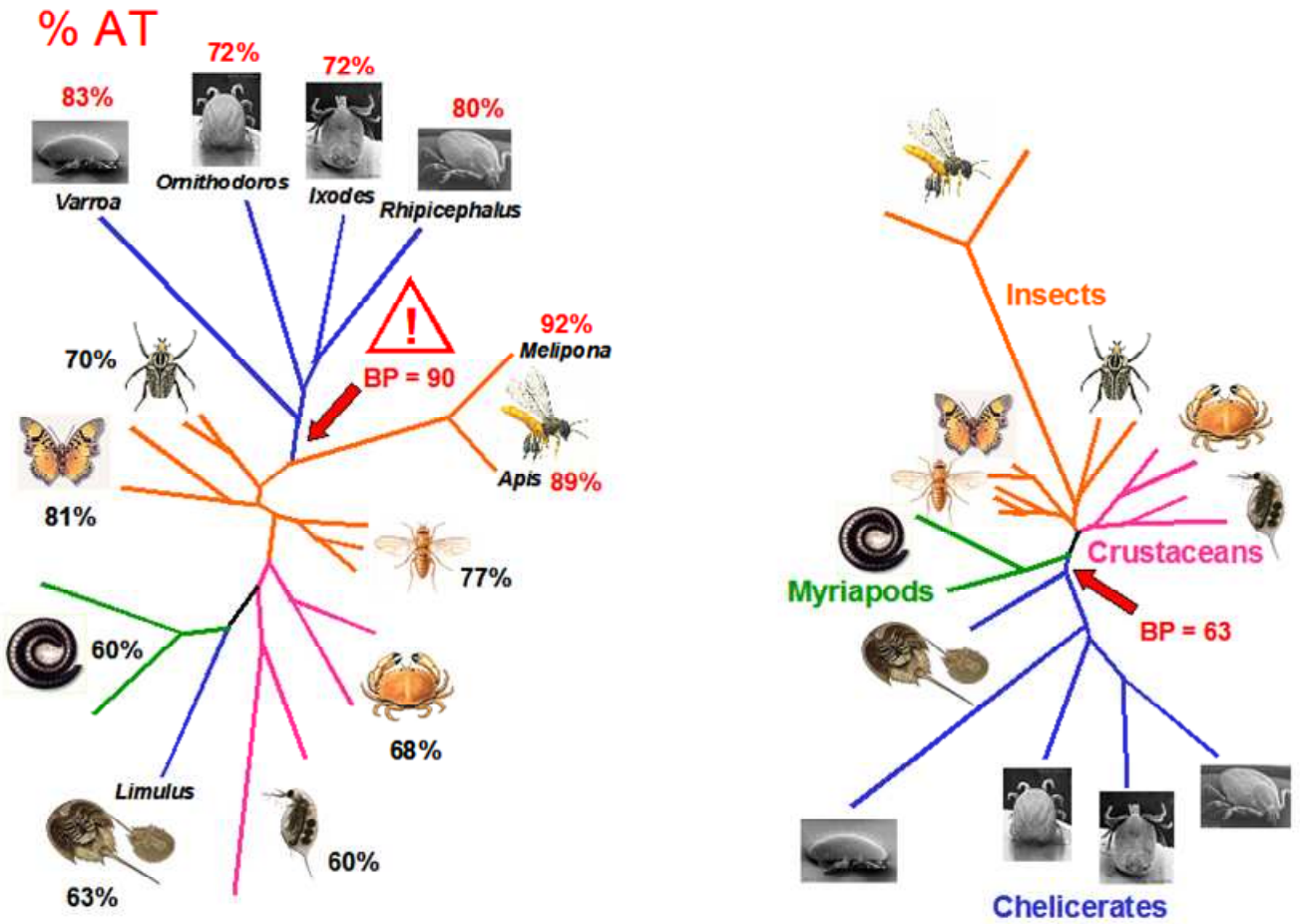

% AT
83%
72%
72%
80%
Varroa
Ornithodoros
Ixodes
Rhipicephalus
!
BP = 90
92%
Melipona
70%
Apis 89%
81%
77%
60%
68%
Limulus
60%
63%
Insects
Crustaceans
Myriapods
BP = 63
Chelicerates

Figure 2b

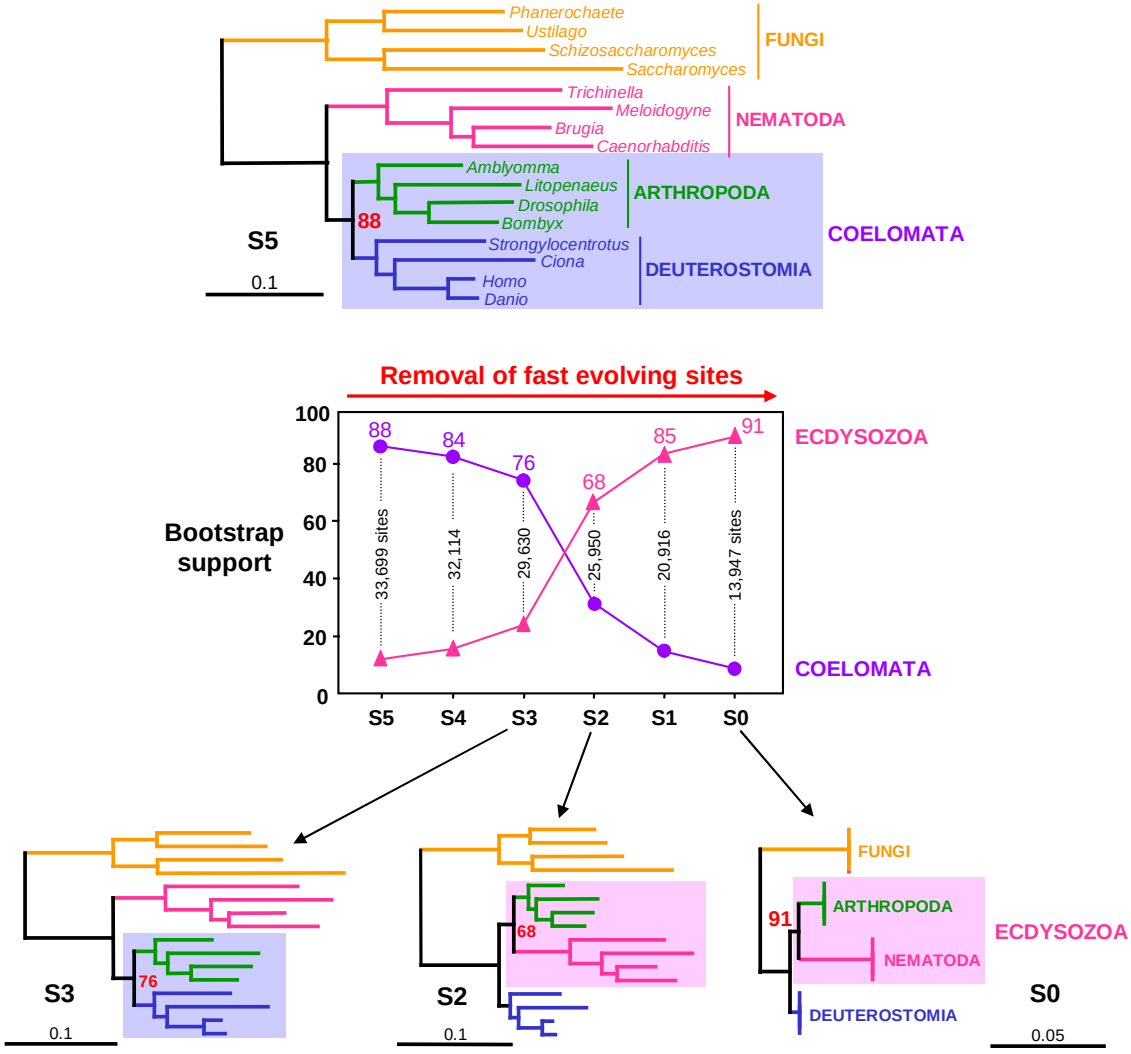
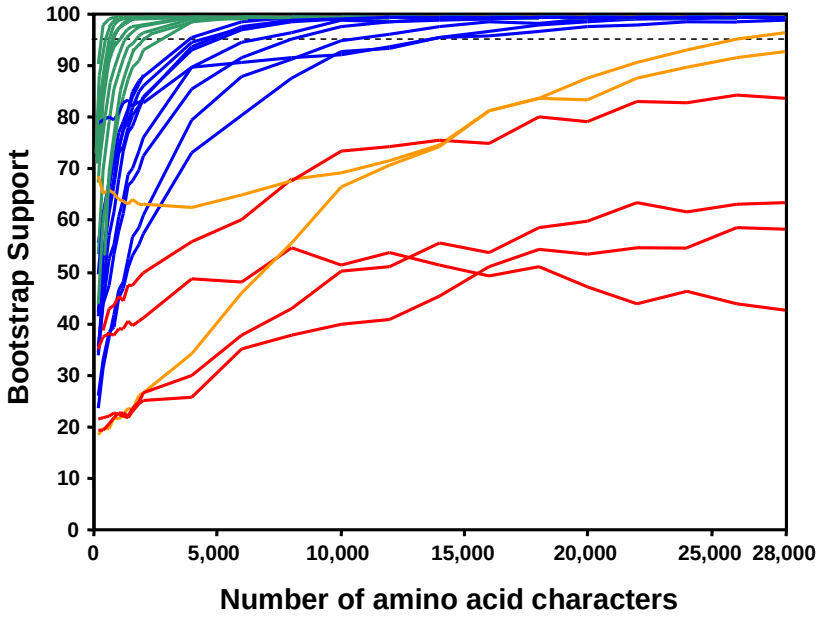
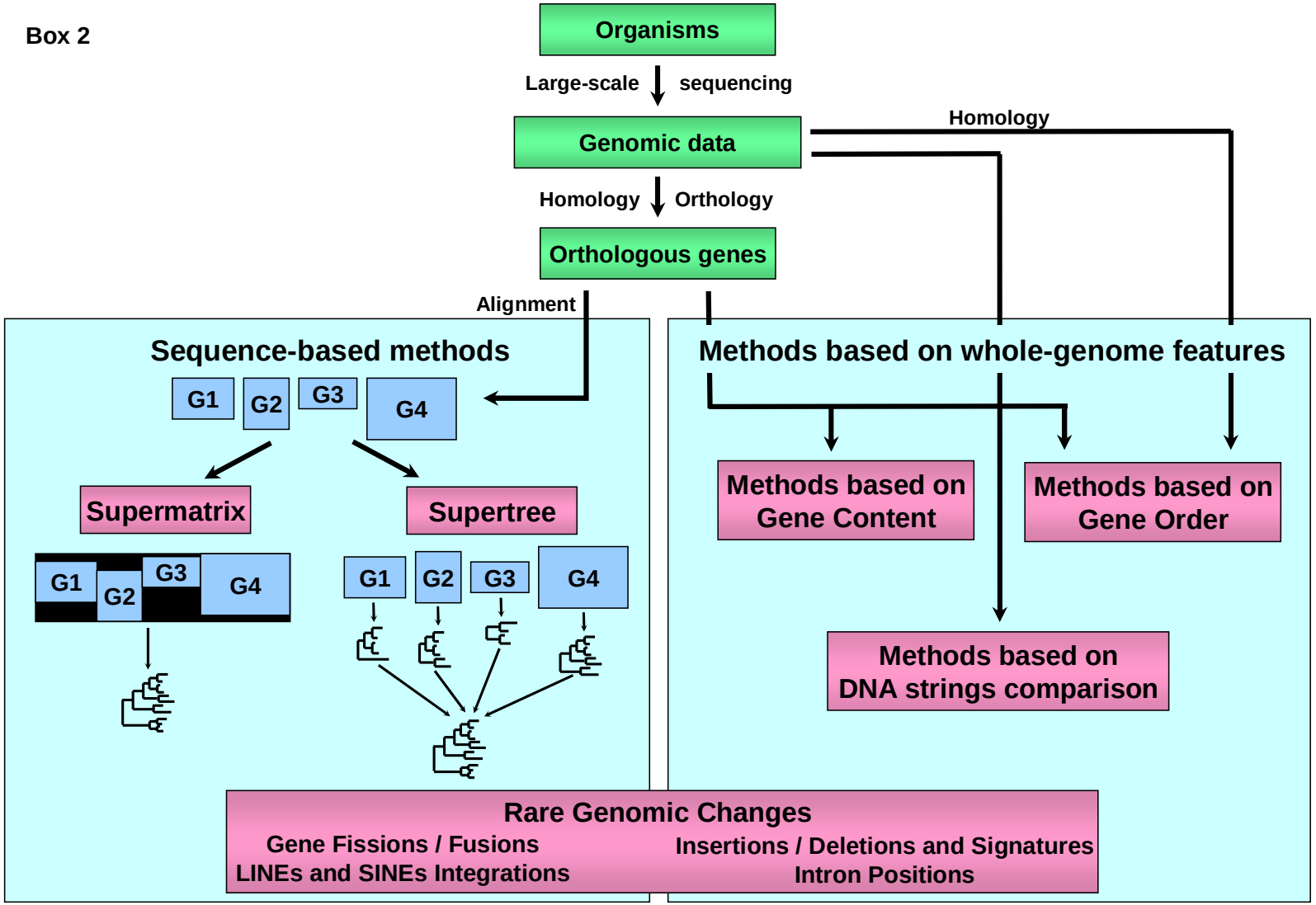

Phanerochaete
Ustilago
Schizosaccharomyces
Saccharomyces
FUNGI
Trichinella
Meloidogyne
Brugia
Caenorhabditis
NEMATODA
Amblyomma
Litopenaeus
Drosophila
Bombyx
ARTHROPODA
88
S5
0.1
Strongylocentrotus
Ciona
Homo
Danio
DEUTEROSTOMIA
COELOMATA
Removal of fast evolving sites
Bootstrap support
100
80
60
40
20
0
88
84
76
68
85
91
33,699 sites
32,114
29,630
25,950
20,916
13,947 sites
ECDYSOZOA
COELOMATA
S5
S4
S3
S2
S1
S0
S3
0.1
76
S2
0.1
68
FUNGI
91
ARTHROPODA
NEMATODA
DEUTEROSTOMIA
ECDYSOZOA
S0
0.05

Figure 3

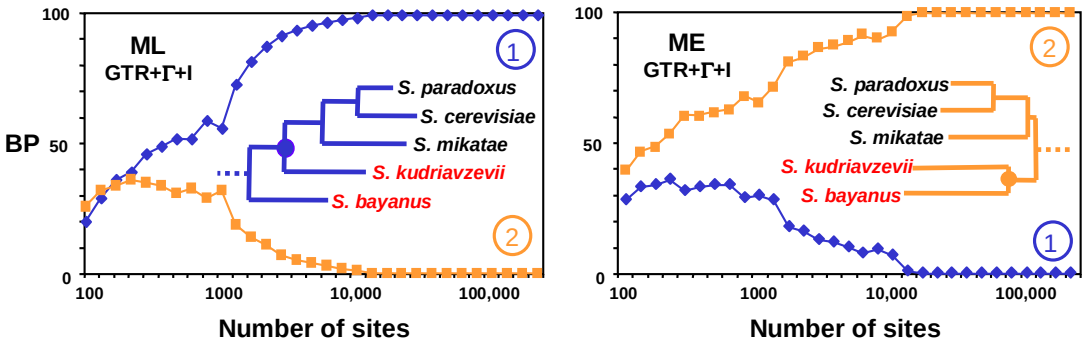

Bootstrap Support
100
90
80
70
60
50
40
30
20
10
0
0
5,000
10,000
15,000
20,000
25,000
28,000
Number of amino acid characters

Box 2

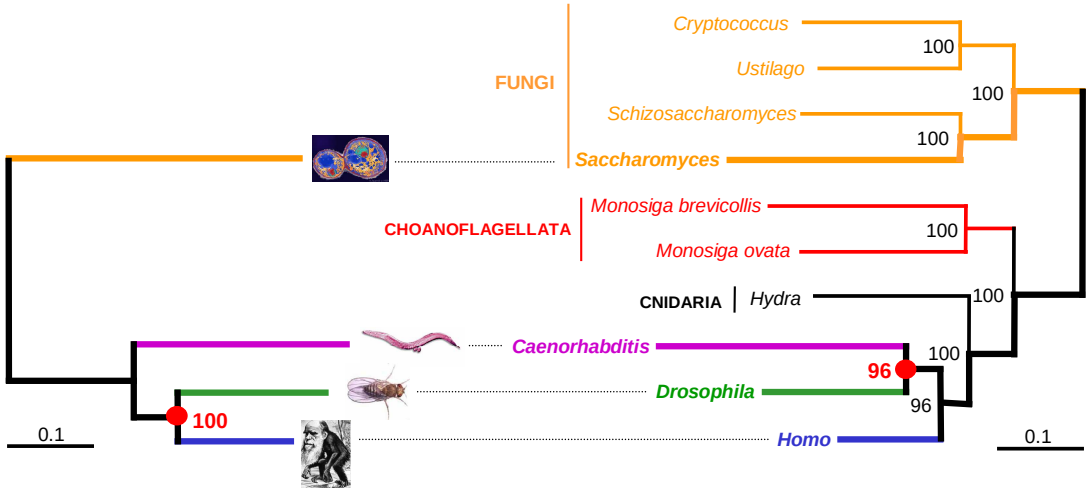

Organisms
Large-scale sequencing
Genomic data
Homology
Homology Orthology
Orthologous genes
Alignment
Sequence-based methods
G1
G2
G3
G4
Supermatrix
Supertree
Methods based on whole-genome features
Methods based on Gene Content
Methods based on Gene Order
Methods based on DNA strings comparison
Rare Genomic Changes
Gene Fissions / Fusions
LINEs and SINEs Integrations
Insertions / Deletions and Signatures
Intron Positions

Box 3

a



b



ML 4 taxa
Coelomata (BP = 100)

ML 10 taxa
Ecdysozoa (BP = 96)