

# THE ORIGIN AND EVOLUTION OF MODEL ORGANISMS

*S. Blair Hedges*

The phylogeny and timescale of life are becoming better understood as the analysis of genomic data from model organisms continues to grow. As a result, discoveries are being made about the early history of life and the origin and development of complex multicellular life. This emerging comparative framework and the emphasis on historical patterns is helping to bridge barriers among organism-based research communities.

Model organisms represent only a small fraction of the biodiversity that exists on Earth, although the research that has resulted from their study forms the core of biological knowledge. Historically, research communities — often in isolation from one another — have focused on these model organisms to gain an insight into the general principles that underlie various disciplines, such as genetics, development and evolution. This has changed in recent years with the availability of complete genome sequences from many model organisms, which has greatly facilitated comparisons between the different species and increased interactions among organism-based research communities. All fields have benefited from these advances, including evolutionary biology, in which the surge in molecular data has greatly clarified estimates of phylogenetic relationships and divergence times among taxa.

In past decades, the term “model organism” has been narrowly applied to those species — such as mouse or *Drosophila* — that, because of their small size and short generation times, facilitate experimental laboratory research. However, in the past decade, with the increase in the number of genome-sequencing projects, this definition has broadened. For example, researchers have focused attention on some organisms, such as the tiger pufferfish, because of unique aspects of their genome rather than their feasibility for experimental studies, and referred to them as “genomic” models<sup>1</sup>. In most cases, economics has had a large part in the choice of organism to study, such as the agriculturally important species (for example, rice) and those related to human health (for example, the malarial parasite *Plasmodium*). All

these species are receiving an unusually large amount of attention from the research community and fall under the broad definition of “model organism”.

Knowledge of the relationships and times of origin of these species can have a profound effect on diverse areas of research<sup>2</sup>. For example, identifying the closest relatives of a disease vector will help to decipher unique traits — such as single-nucleotide polymorphisms — that might contribute to a disease phenotype. Similarly, knowing that our closest relative is the chimpanzee is crucial for identifying genetic changes in coding and regulatory genomic regions that are unique to humans, and are possibly associated with traits such as intelligence<sup>3</sup>. Furthermore, knowing when humans and chimpanzees diverged from one another allows researchers to calibrate the rates of genetic change in modern humans and to estimate when populations migrated to different regions of the World<sup>4</sup>. In addition, genomic comparisons across species are fundamental to locating conserved gene sequences, which presumably reflect the constraints that are imposed by natural selection<sup>5</sup>. The methodology for generating phylogenetic trees from sequence data continues to be refined and expanded, and aids the above studies. The various methods that are, at present, used to generate phylogenetic trees and estimate divergence times among taxa are described in BOX 1.

Model organisms are found among the prokaryotes, protists, fungi, plants and animals. Therefore, a discussion of their origin and evolution necessarily concerns the pattern and timing of the “tree of life”. In the early 1990s, the tree of life was derived mostly from the small subunit ribosomal RNA (rRNA) gene<sup>6</sup>, and the timescale

NASA Astrobiology Institute  
and Department of Biology,  
208 Mueller Laboratory,  
The Pennsylvania State  
University, University Park,  
Pennsylvania 16802, USA.  
e-mail: sbh1@psu.edu  
doi:10.1038/nrg929

was on the basis of the geochemical and fossil record<sup>7,8</sup> (FIG. 1a). Eukaryotes and their genomes were considered to be closest relatives in the Archaeobacteria (and their genomes), just as any two species might be close relatives. The current view (FIG. 1b) is noticeably different, and it holds that eukaryotes are genomic hybrids of Eubacteria and Archaeobacteria: numerous genes were transferred to the eukaryote nucleus during symbiotic events, such as those that gave rise to mitochondria, leading to the

fusion of some evolutionary branches. The relative contribution of these major gene transfers, other HORIZONTAL TRANSFER events and the number of symbiotic events is highly debated<sup>9</sup>, with some favouring the existence of a eukaryotic root<sup>10</sup>. The times of divergence — as estimated from molecular clocks (FIG. 1c; BOX 1) — are also debated, generating attention because they are often older than the corresponding fossil dates. However, this is not surprising as fossil-based times are minimum

#### Box 1 | Methods for estimating molecular phylogenies and times of divergence

##### Sequence data

Both DNA and protein sequences are used for estimating phylogenetic relationships and times of divergence among taxa. Typically, DNA sequences are used for relatively recent events — for example, the human and chimpanzee split — when protein sequences are too conserved to be useful. Protein sequences are desirable for more ancient events — for example, human divergence from insects — when DNA sequences are usually too divergent to make accurate estimates on the basis of patterns of nucleotide substitutions. Unequal base or amino-acid composition among the genomes of different species is common and makes sequence change more difficult to estimate. In addition, sequence length is a limiting factor, in that the average gene (coding) or protein sequence (~1,000 nucleotides, ~350 amino acids) is usually not long enough to yield a robust phylogeny or time estimate, and therefore many genes and proteins must be used.

##### Phylogeny estimation

The general principle behind phylogenetic methods is to find a tree that minimizes sequence change. For example, if two species have a unique amino acid at a particular site and are joined in the tree, only one change (in their ancestor) is needed to explain this data. Conversely, an additional change would be required if the two species were not joined in the tree, making the other tree less likely to be the true tree. The two tree-building methods that are most

often used with molecular sequence data are minimum evolution, such as NEIGHBOUR JOINING, and MAXIMUM LIKELIHOOD<sup>117</sup>. These methods, and the BAYESIAN METHOD<sup>118</sup>, are flexible enough to include diverse information on the biological nature of molecular sequence change, such as rate variation among sites. A fourth method, MAXIMUM PARSIMONY, is also widely used. Although the various methods are quite different from one another, they often result in the same phylogenetic tree. Reliability can be tested in different ways, with the BOOTSTRAP METHOD<sup>119</sup> being the most widely used. Phylogeneticists often use and compare several methods in a single study to evaluate the robustness of their results.

##### Time estimation

On the basis of the observation that sequences diverge in a roughly clock-like fashion, the 'molecular clock' method is used to estimate divergence time<sup>120</sup>. Usually, the rate of divergence is determined by dividing the substitutional differences observed between two species, by the time elapsed since their divergence. This is based on a fossil calibration, with the 'calibration time' for the divergence of species 1 and 2 being shown in a. The rate obtained is then used to estimate the timing of unknown branch points on the tree (a). Divergence times that are based on fossils always yield overestimates of the true rates, so the measuring of one or a few robust calibration points might be more reliable than averaging the values that are obtained from many less-robust calibration points (b). The fact that different genes evolve at different rates is an advantage because it allows them to be used for different time periods and levels of phylogeny. Lineages that evolve at different rates can be detected by RELATIVE RATE TESTS<sup>62</sup>, and those comparisons can be either omitted from the analysis or accommodated by methods that allow the rate to be adjusted<sup>16,65,121</sup>. Divergence times can also be overestimated; for example, this can occur if paralogous comparisons, which measure earlier gene-duplication events rather than speciation events, are accidentally included in the analysis (c). In the example shown, the early duplication of genes 2 and 3 would lead to an overestimate of the time at which species 1 and 2 diverged.

##### HORIZONTAL TRANSFER

The transfer of genetic material between the genomes of two organisms, which are usually different species.

##### NEIGHBOUR JOINING

A method that selects the tree that has the shortest overall length (sum of all branch lengths).

##### MAXIMUM LIKELIHOOD

A method that selects the tree that has the highest probability of explaining the sequence data, under a specific model of substitution (changes in the nucleotide or amino-acid sequence).

##### BAYESIAN METHOD

A method that selects the tree that has the greatest posterior probability (probability that the tree is correct), under a specific model of substitution.

##### MAXIMUM PARSIMONY

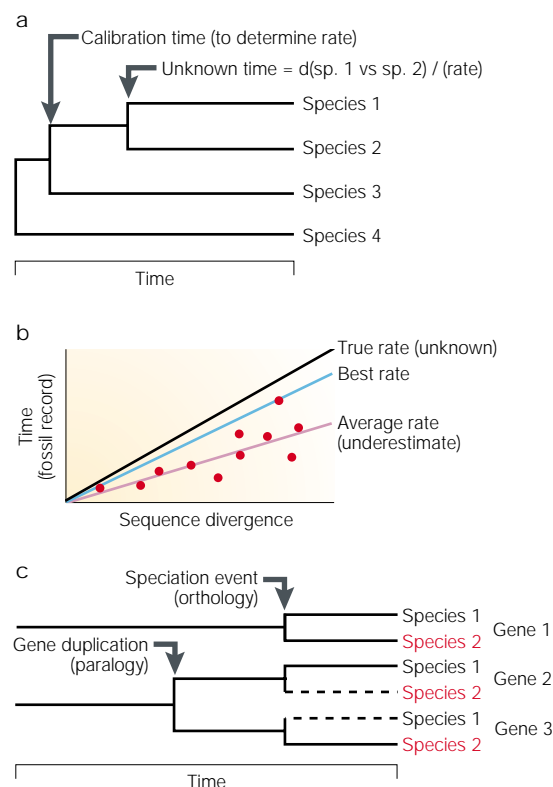
A method that selects the tree that requires the fewest number of substitutions.

##### BOOTSTRAP METHOD

As applied to molecular phylogenies. Nucleotide or amino-acid sites are sampled randomly, with replacement, and a new tree is constructed. This is repeated many times and the frequency of appearance of a particular node among the bootstrap trees is viewed as a support (confidence) value for deciding on the significance of that node.

##### RELATIVE RATE TESTS

Statistical tests that determine, at a given level of stringency, whether two or more branches in a tree have evolved at the same rate of sequence change.



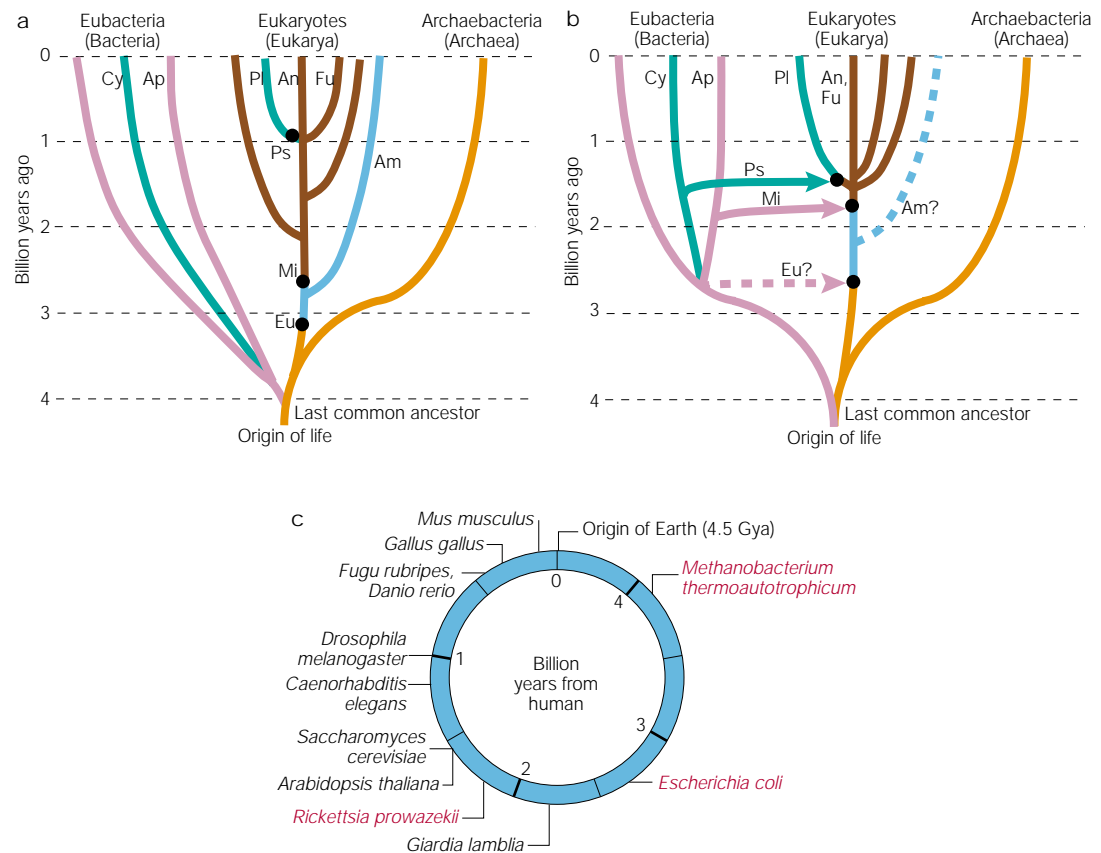


Figure 1 | **Changing views of the tree and timescale of life.** **a** | An early-1990s view, with the tree determined mostly from ribosomal RNA (rRNA) sequence analysis. This tree emphasizes vertical (as opposed to horizontal) evolution and the close relationship between eukaryotes and the Archaeobacteria. The deep branching (>3.5 Giga ( $10^9$ ) years ago, Gya) of CYANOBACTERIA (Cy) and other Eubacteria (purple), the shallow branching (~1 Gya) of plants (Pl), animals (An) and fungi (Fu), and the early origin of mitochondria (Mi), were based on interpretations of the geochemical and fossil record<sup>7,8</sup>. Some deeply branching amitochondriate (Am) species were believed to have arisen before the origin of mitochondria<sup>44</sup>. Major symbiotic events (black dots) were introduced to explain the origin of eukaryotic organelles<sup>42</sup>, but were not assumed to be associated with large transfers of genes to the host nucleus. They were: Eu, joining of an archaeobacterium host with a eubacterium (presumably a SPIROCHAETE) to produce an amitochondriate eukaryote; Mi, joining of a eukaryote host with an  $\alpha$ -proteobacterium (Ap) symbiont, leading to the origin of mitochondria, and plastids (Ps), joining of a eukaryote host with a cyanobacterium symbiont, forming the origin of plastids on the plant lineage and possibly on other lineages. **b** | The present view, based on extensive genomic analysis. Eukaryotes are no longer considered to be close relatives of Archaeobacteria, but are genomic hybrids of Archaeobacteria and Eubacteria, owing to the transfer of large numbers of genes from the symbiont genome to the nucleus of the host (indicated by coloured arrows). Other new features, largely derived from molecular-clock studies<sup>16,39</sup> (BOX 1), include a relatively recent origin of Cyanobacteria (~2.6 Gya) and mitochondria (~1.8 Gya), an early origin (~1.5 Gya) of plants, animals and fungi, and a close relationship between animals and fungi. Coloured dashed lines indicate controversial aspects of the present view: the existence of a premitochondrial symbiotic event and of living amitochondriate eukaryotes, ancestors of which never had mitochondria. **c** | The times of divergence of selected model organisms from humans, based on molecular clocks. For the prokaryotes (red), because of different possible origins through symbiotic events, divergence times depend on the gene of interest.

CYANOBACTERIA

A phylum of Eubacteria, formerly known as the “blue-green algae”. These prokaryotes are the only organisms known to be capable of oxygenic photosynthesis.

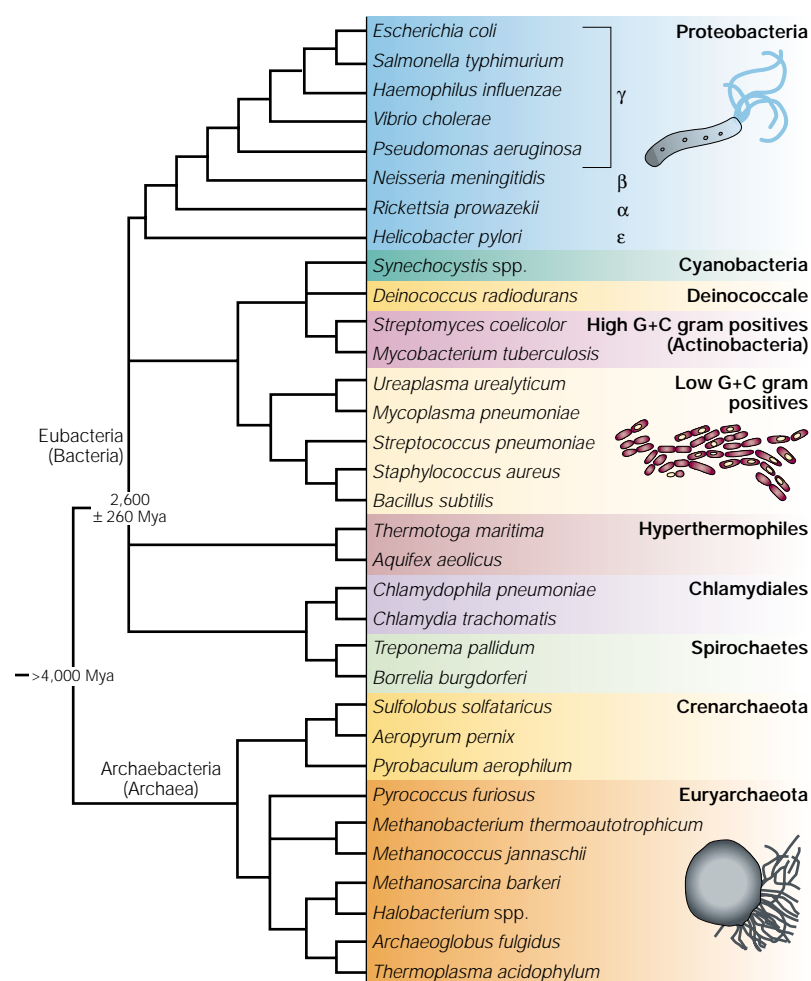
SPIROCHAETES

A phylum of Eubacteria that has a spiral or corkscrew-like appearance and axial filaments (similar to flagella). These prokaryotes are responsible for human diseases, such as Lyme disease and syphilis.

estimates, whereas molecular divergence begins as soon as two lineages separate.

Here, I give a brief overview of the current knowledge of the phylogeny and divergence times of model organisms, with an emphasis on recent developments. For ease of discussion, this review is organized into five sections — prokaryotes, protists, plants, fungi and animals — which correspond to the classical groups of organisms that are recognized by nearly all biologists. This is not meant to advocate the five-kingdom<sup>11</sup> over the three-domain<sup>6</sup> — namely, Eubacteria, Eukarya and Archaeobacteria — classification of life. Each is a useful

system in a different way, with the former emphasizing cytological and structural differences and the latter emphasizing genetic differences. The fact that some single-celled eukaryotes are more closely related to plants or animals than to other single-celled eukaryotes has not prevented the use of the word ‘protist’ to describe this group of organisms, or its utility. Similarly, the recognition that Archaeobacteria and Eubacteria have clear genetic differences has not caused biologists to abandon the practical words ‘bacteria’ and ‘prokaryote’ that aptly describe both forms, and these systems remain useful and compatible ways of discussing biodiversity.



**Figure 2 | A phylogeny of prokaryotes.** The relationships of selected prokaryote model organisms based on recent studies<sup>14–19</sup>. Times of divergence (million years ago (Mya)  $\pm$  one standard error) are indicated at nodes in the tree<sup>16,39</sup>. Branch lengths are not proportional to time. Phyla and phylum-level groupings are indicated on the right.

### Prokaryotes

Prokaryotes are organisms that lack both membrane-bound organelles and a nucleus. They include the classical model of molecular biology, *Escherichia coli*, the spore-forming *Bacillus subtilis* (which is widely used in biotechnology) and notorious agents of disease, such as *Borrelia burgdorferi* (Lyme disease), *Mycobacterium tuberculosis* (tuberculosis), *Mycoplasma pneumoniae* (pneumonia) and *Vibrio cholerae* (cholera). Also included are species that are highly resistant to radioactivity (*Deinococcus radiodurans*) and the oxygen-producing cyanobacteria (*Synechocystis* spp.). The number (~5,900) of described prokaryote species is probably a considerable underestimate, not to mention that the species concept for these organisms is highly debated<sup>12</sup>.

For the past two decades, most prokaryote phylogenies have been constructed by analysing the sequences of the small subunit rRNA gene. Such data led to the recognition of Eubacteria and Archaeobacteria as two distinct domains<sup>13</sup>. In this review, I use these original names instead of 'Bacteria' and 'Archaea', as later proposed<sup>6</sup>. This is because the former allude nicely to the

cytological similarities between the two domains and avoid the confusion between 'Bacteria' and 'bacteria', as well as recognizing their status as genetically distinct groups. However, the recent availability of genomic data has shifted the emphasis towards building protein phylogenies that are derived from the sequences of many genes<sup>14–19</sup>, the presence and absence of genes<sup>17,20–22</sup> and the combination of gene and protein trees<sup>23</sup>. The horizontal transfer of genes is often difficult to confirm by phylogeny alone because the short length of typical proteins (~300 residues) usually precludes the construction of a robust tree, and different methods of detection do not always agree<sup>24</sup>. Therefore, 'misplaced' species on a tree might be evidence of horizontal transfer or poor resolution<sup>18</sup>. Despite the methodological problems that arise from analysing highly divergent sequences, these genomic phylogenies have converged on several well-supported groups (FIG. 2). As well as horizontal transfer, the tree-building problems that are under the greatest scrutiny are: variation in the rate of nucleotide or amino-acid substitution among sites (for example, the influence of substitutional hot spots); rate variation among lineages; and different compositions of nucleotides (for example, high G+C content) and amino acids among sequences.

One debated phylogenetic question is whether archaeobacteria form a single MONOPHYLETIC group. The insertion of 11 amino acids in the ELONGATION FACTOR 1  $\alpha$ -protein indicated initially that some archaeobacteria (Crenarchaeota) were more closely related to eukaryotes than to the other group of archaeobacteria (Euryarchaeota)<sup>25</sup>. Since then, the insertion–deletion region of the  $\alpha$ -protein has been found to be more variable than anticipated and is no longer considered to be strong evidence for archaeobacterial PARAPHYLY<sup>26</sup>. In addition, analyses of whole-genome sequences of crenarchaeotans do not obviously support archaeobacterial paraphyly<sup>27</sup>. Recently, the sequence analyses of 19 proteins significantly (>95%) supported archaeobacterial monophyly, with none significantly favouring paraphyly<sup>16</sup>; and analysis of 23 combined proteins<sup>15</sup> also supported archaeobacterial monophyly. Nonetheless, a paraphyletic Archaeobacteria was obtained in another three studies<sup>15,28,29</sup>, although in one case the result was attributable to a tree reconstruction artefact. Because all of these studies accounted for complex models of evolution — such as rate variation among sites — the different results are not explained easily, thereby leaving open the question of whether Archaeobacteria are monophyletic or paraphyletic.

Another intensely debated phylogenetic question involves the position of hyperthermophiles. Earlier studies with rRNA placed these species near the root of the tree, implying that the common ancestor of all living organisms lived at high temperatures<sup>12</sup>. For some, this fuelled speculation that life might have arisen at hydrothermal vents, whereas others found it consistent with an early hot Earth environment and the survival of gigantic, ocean-boiling asteroid effects<sup>30</sup>. Recent scrutiny of prokaryote phylogenies has thrown cold water on these hypotheses<sup>31–33</sup>. In particular, the separate branching

#### MONOPHYLETIC

Includes all the descendants of a single common ancestor.

#### ELONGATION FACTOR 1

An enzyme that functions in the process of protein translation.

#### PARAPHYLETIC

Includes some, but not all, of the descendants of a single common ancestor.

of two eubacterial hyperthermophiles, *Aquifex* and *Thermotoga*, near the root of the tree has been challenged by independent analyses, which indicate that they group together, possibly at a higher location on the tree<sup>17,32,33</sup>. There is disagreement as to whether this grouping results

from a true phylogenetic signal<sup>32</sup> or noise<sup>15</sup>. The debate will certainly continue but, if confirmed, this new result will cause a rethinking of the early history of life and its environment.

Timing is the other half of the story. When did the last common ancestor of all life live and when did the principal groups of Eubacteria and Archaeobacteria arise? A robust timescale for prokaryotes has not yet been determined, but some clues have come from fossils, biomarkers in ancient rocks — such as breakdown products of cell membranes — and molecular clocks. Organic residue from some of the earliest rocks (~3.9 Giga (10<sup>9</sup>) years ago, Gya) might<sup>34</sup> or might not<sup>35</sup> indicate the presence of life, and fossils from ~3.5 Gya could be prokaryotes<sup>36</sup> or simply artefacts<sup>37</sup>. Timing the early splits in life with molecular clocks has also proved challenging. This is because of methodological problems, such as accounting for sequence changes that have been obscured by repeated substitutions at the same nucleotide or amino-acid position and, for reasons yet to be determined, the finding that eukaryotes evolve faster than prokaryotes<sup>16,38</sup>. In addition, the fidelity of genetic replication and repair systems in the early history of life is unknown, and the different environment of early Earth might have affected rates of molecular change. It is for these reasons that we have less confidence in the time estimates for the earliest splitting events. On the basis of independent attempts to date using many proteins, an early time (>4 Gya) was set for the last common ancestor, with surprisingly younger dates (2–3 Gya) for the origin of Cyanobacteria<sup>16,39</sup>. The late emergence of Cyanobacteria was a surprise, owing to the assumption that oxygenic photosynthesis had evolved by at least 3.5 Gya (REF. 30). However, the earliest biomarker evidence for Cyanobacteria is only 2.7 Gya (REF. 40), and oxygenic photosynthesis might have evolved later. Molecular-clock studies also indicate a rapid radiation of the main groups of Eubacteria around this time (~2.5 Gya) with no deeper side branches. The reason for this is unclear, but might reflect the origin of oxygenic photosynthesis at about that time, causing a mass extinction of many lineages, while opening new niches for the later ADAPTIVE RADIATIONS of prokaryotes<sup>16</sup>.

Protists

The single-celled eukaryotes, informally known as 'protists', do not form a group, but are instead widely known to be paraphyletic; that is, some protists are more closely related to non-protists (such as plants, animals and fungi) than to other protists. About 100,000 living protist species have been described<sup>41</sup>. They include genetic model organisms, such as *Dictyostelium* and *Volvox*, and parasites of humans, such as *Plasmodium* (malaria), *Trypanosoma* (trypanosomiasis, Chagas disease), *Leishmania* (leishmaniasis) and *Giardia* (giardiasis). Genome projects for many protists are well underway, although the phylogenetic relationships of protists remain controversial. The scheme shown in FIG. 3 is a synthesis of recent studies, with each focusing on a different segment of the tree.

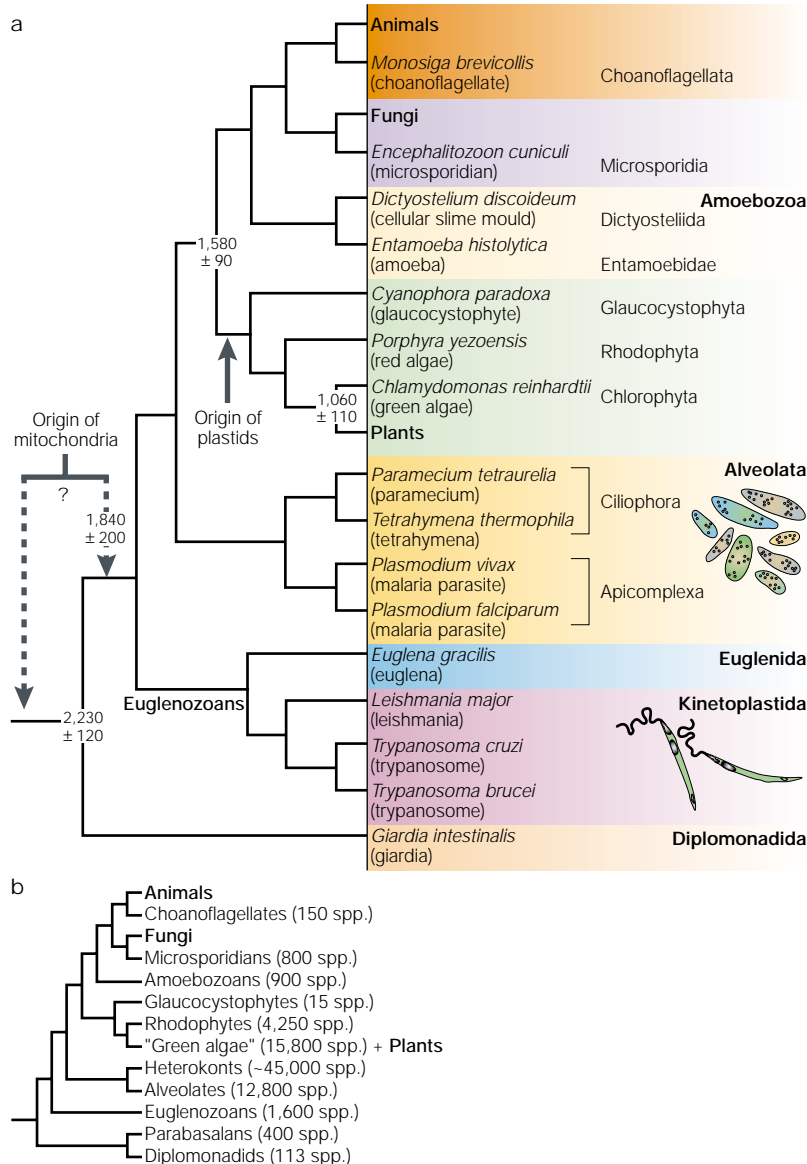


Figure 3 | **A phylogeny of protists.** **a** | The relationships and divergence times (millions of years ago (Mya) ± one standard error) of selected model protists and symbiotic events (mitochondria and plastids) are shown, based on recent studies<sup>16,51,52,54–56,67</sup>; branch lengths are not proportional to time. The dashed lines indicate the current debate over the origin of mitochondria, presumably a single event that occurred either before the last common ancestor of living eukaryotes or after the divergence of *Giardia intestinalis* (at 2,230 ± 120 Mya)<sup>16</sup>. **b** | The relationships and numbers of living species in the main groups of organisms<sup>41</sup>. Some groups, such as foraminiferans (single-celled organisms with shells; ~5,000 living species) are omitted because of their uncertain position. Molecular clocks<sup>16</sup> show an origin of eukaryotes at 2.7 Gya (biomarker evidence<sup>122</sup> has also indicated that eukaryotes were present at this time). The earliest fossil eukaryote appeared at 2.1 Gya (REF. 123), and other fossils show that protists, were cytologically and ecologically diverse by at least 1.5 Gya (REF. 124). The symbiotic event leading to the origin of mitochondria was dated at 1.8 Gya (REF. 16), and the split of the three main multicellular kingdoms (plants, animals and fungi) at about 1.5 Gya (REF. 51). Considering that red algae are on the plant lineage, and so arose after 1.5 Gya but before the earliest fossils of red algae at 1.2 Gya (REF. 78), the symbiotic event leading to the origin of plastids must have occurred during that interval (1.5–1.2 Gya).

Protistan phylogeny bears directly on the fundamental questions of eukaryote evolution, such as the origin of mitochondria. The symbiotic origin of this organelle is nearly indisputable<sup>11</sup>, although current debate centres on whether or not there was an earlier premitochondrial period in eukaryote evolution and if some living amitochondriate groups (for example, *Giardia*) arose during that time. The two positions are not inseparable and there might have been a premitochondrial period that left no living representatives<sup>16,42,43</sup>. For living eukaryotes, a premitochondrial period has been assumed, partly because several amitochondriate eukaryotes were found to have basal positions in phylogenetic trees<sup>44</sup>. However, some phylogenetic analyses have indicated that the most recent common ancestor of all living eukaryotes had a mitochondrion<sup>45–47</sup>. This problem is complex, and deciding among alternatives often requires more details than are found in the branching pattern of a tree<sup>43</sup>. Resolution of this debate might require the analysis of complete genomes of selected amitochondriate protists, placing particular attention on gene structure and the location and function of their proteins.

Among the several amitochondriate protists in question, the case for microsporidians — that is, intracellular parasites — being secondarily amitochondriate has received the most support, especially with the recent finding of tiny micro-organelles that seem to be remnants of mitochondria<sup>48</sup>. In addition, the nuclear genomes of microsporidia are greatly reduced and the rates of substitution are generally accelerated<sup>49</sup>. In phylogenies of individual proteins, microsporidia are either basal among eukaryotes — apparently because of substitution biases — or they cluster with fungi. The general consensus is that they are related to fungi, although the details of this relationship have yet to be established<sup>50</sup>. If microsporidia were related to fungi, they would provide a 'close relative' for comparative genetic research.

Remarkably, the relationships of plants, animals and fungi, have not been conclusively resolved. In traditional phylogenies, plants and fungi have been united, but analyses of individual and combined proteins have supported either an animal–fungus or animal–plant grouping, with the weight of evidence leaning towards the former<sup>51,52</sup> (FIG. 3). Nonetheless, those divergences that led to the three kingdoms were apparently closely spaced in time<sup>51</sup>. Even less certain are the positions of most lineages of protists relative to these three kingdoms. Analyses of several selected proteins, and a single gene-fusion event<sup>53</sup>, place the most diverse protist groups, including the Euglenozoa (euglenids and kinetoplasts), Alveolata (for example, apicomplexans, ciliates and dinoflagellates) and Heterokonta (including brown algae, golden algae and diatoms) on the lineage leading to plants, once it split from the animal–fungus lineage<sup>52</sup>. However, the combined sequence analyses of many proteins place these groups basal to the three kingdoms with the amoebae as the closest relatives of the animal–fungus group<sup>54,55</sup>. As predicted by morphology, choanoflagellates are the closest relatives of animals<sup>56</sup> (FIG. 3). Some analyses of individual genes and proteins

have indicated that the plastid-bearing rhodophytes (red algae) are basal to plants, animals and fungi<sup>57</sup>, although this group and the plastid-bearing GLAUCOCYSTOPHYTES are found on the plant lineage, basal to green algae, when many proteins are considered<sup>54</sup> (FIG. 3).

Besides these challenging phylogenetic questions, little is known of when the main branches of living protists split from each other (FIG. 3). The murky picture of protist evolution is certain to come into sharper focus in the next few years as protist genomes are completed and analysed. The details of relationships and timing will be of particular interest to those researchers who study the cytological, genetic and developmental implications of the transition from simple, unicellular organisms to complex multicellular life.

#### Plants

There are ~300,000 known species of land plants, but genomic models are only found in a handful of species and families (FIG. 4). These include the classical model of genetics and development, *Arabidopsis thaliana*, grasses such as rice (*Oryza sativa*), corn (*Zea mays*) and wheat (*Triticum aestivum*), which are food plants for most human populations, and other economically important crop plants, such as cotton (*Gossypium hirsutum*) and soybean (*Glycine max*).

Molecular phylogenetic studies of plants have focused on organellar genes (especially ribulose biphosphate carboxylase of the chloroplast, *rbcL*) and the nuclear small subunit ribosomal RNA gene<sup>58,59</sup>, with nuclear proteins being largely untouched (FIG. 4b). One current debate concerns the relationships of the vascular plants (tracheophytes) to the three lineages of 'BRYOPHYTIC' land plants (HORNWORTS, mosses and LIVERWORTS). The two primary hypotheses differ in that one designates liverworts<sup>60</sup> and the other hornworts<sup>58</sup> as the most primitive land plants. In the latter, there is some evidence that mosses and liverworts are close relatives. Knowing the correct phylogeny will help to understand how plants evolved and adapted to the terrestrial environment. For example, under the liverworts-basal hypothesis, the genes and structures shared by the model species *Physcomitrella patens* (moss) and *Marchantia polymorpha* (liverwort) would be considered primitive and present in the common ancestor of all land plants. By contrast, under the hornworts-basal hypothesis, such sequences or traits might simply have arisen on the side branch of land plants that led to liverworts and mosses.

Molecular clock studies in plants have been hampered by the availability of a relatively small number of genes that are expected to result in less-precise time estimates<sup>61</sup>. Moreover, phylogenetic trees of chloroplast and mitochondrial DNA sequences show large variation in branch lengths, which indicates a complex history of rate changes. Such rate variation complicates the estimates of branching order and the times at which the branches split. Methods are available to correct for these rate differences<sup>62–65</sup>, but the rate variation might be too great and complex to decipher<sup>66</sup>. An additional problem lies in the use of organellar sequence data, which

#### ADAPTIVE RADIATION

The rapid diversification of a group of species into various habitats over a relatively short period of geological time. However, the term is often used as a synonym for any large monophyletic group of taxa.

#### GLAUCOCYSTOPHYTES

A small group of freshwater algae, also called 'glaucophytes'. Species in this group have plastids with a peptidoglycan cell wall (peptidoglycan is the main component of bacterial cell walls).

#### BRYOPHYTES

A term that refers traditionally to non-vascular land plants, nearly all of which are quite small (1–2 cm high). Bryophytes include hornworts, liverworts and mosses; however, the term might also be used in a more restricted sense to refer to the mosses alone (Division: Bryophyta).

#### HORNWORTS

A group of small, non-vascular plants (Division: Anthocerotophyta) that are distinguished by their tall horn-like sporophyte (diploid generation) that grows on the more flattened gametophyte (haploid generation). They usually have a single, large chloroplast in each cell.

#### LIVERWORTS

A group of small, mat-like, non-vascular plants (Division: Marchantiophyta) that occur in diverse habitats but most commonly on the forest floor. Some species have lobe-shaped leaves that resemble a liver.

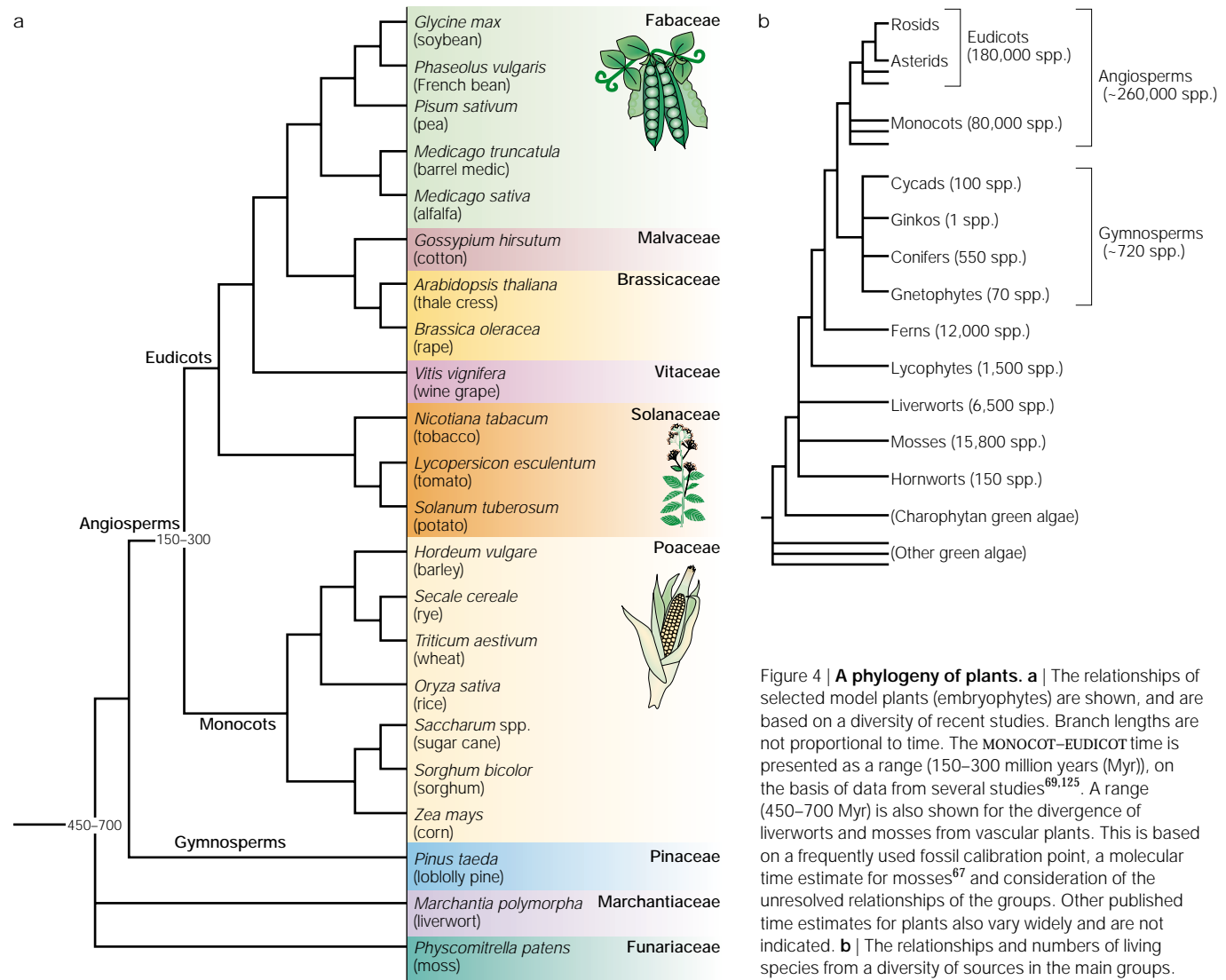


Figure 4 | **A phylogeny of plants.** **a** | The relationships of selected model plants (embryophytes) are shown, and are based on a diversity of recent studies. Branch lengths are not proportional to time. The MONOCOT–EUDICOT time is presented as a range (150–300 million years (Myr)), on the basis of data from several studies<sup>69,125</sup>. A range (450–700 Myr) is also shown for the divergence of liverworts and mosses from vascular plants. This is based on a frequently used fossil calibration point, a molecular time estimate for mosses<sup>67</sup> and consideration of the unresolved relationships of the groups. Other published time estimates for plants also vary widely and are not indicated. **b** | The relationships and numbers of living species from a diversity of sources in the main groups.

**MONOCOTS**  
(Monocotyledonous plants).  
Flowering plants with one cotyledon (or seed leaf).

**EUDICOTS**  
The largest clade of angiosperms, characterized by two cotyledons (seed leaves) and three symmetrically placed pollen apertures or aperture arrangements that are derived from this.

**ANGIOSPERMS**  
Flowering vascular plants that form seeds inside an ovary.

**GYMNASPERMS**  
Non-flowering vascular plants with naked seeds that are not enclosed in an ovary (for example, pine).

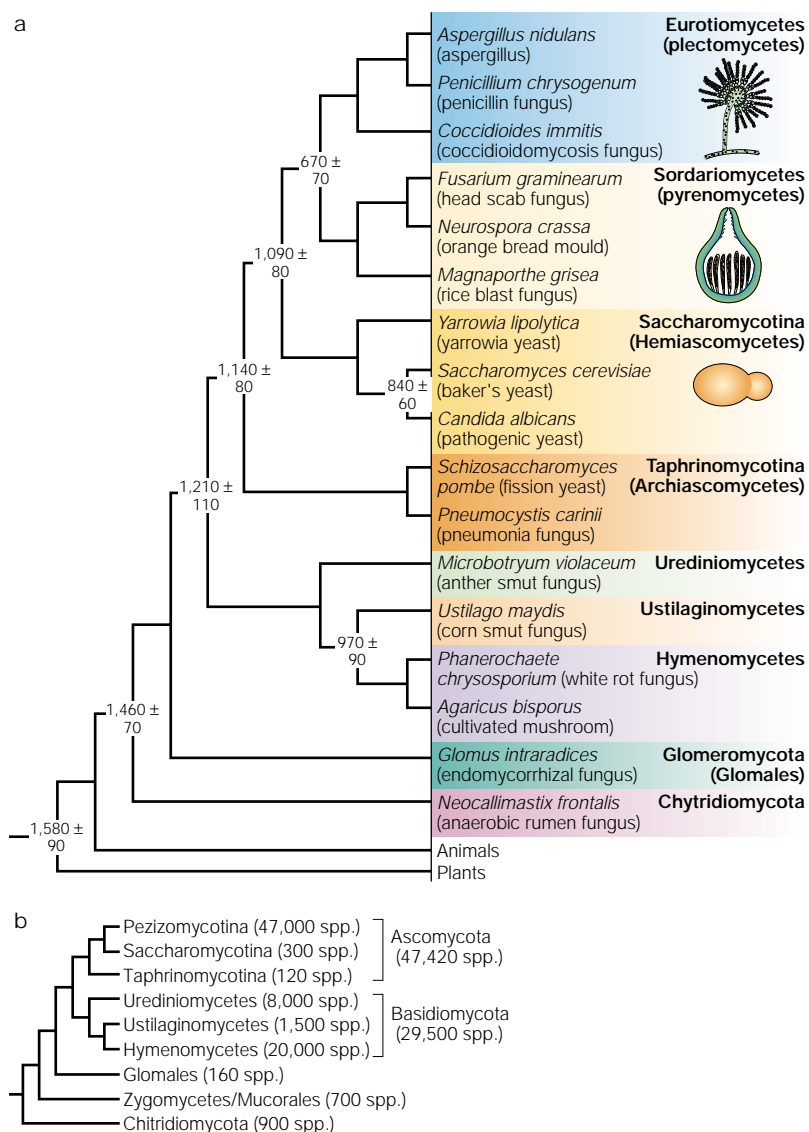
restricts calibrations to plants, whereas nuclear-protein data permits calibration with animals<sup>67</sup>, which have a more robust fossil record. Because of the current difficulties in timing divergences among plants, only selected times are shown in FIG. 4.

Three branch points in plants have been given more attention than others: the divergence between liverworts and vascular plants, ANGIOSPERMS and GYMNASPERMS, and monocots and eudicots. Usually, the first divergence time was used to set the calibration date which, in turn, was used to estimate the other two splitting events. However, there has been confusion in the literature as to which event is being timed. In animals, and most other groups, time estimates are typically made between two lineages, with the origin of each lineage being the point at which the two lineages split. The terminology is different in the plant literature, in which the origin of a group — defined as the split between the two most divergent living representatives — is often the time estimate in question<sup>68</sup>. The difficulty arises when the identity of these two lineages is not agreed on<sup>68,69</sup>.

If bryophytes split from vascular plants in the PRECAMBRIAN (>543 million years ago (Mya))<sup>67</sup>, this would indicate that divergence times in the literature, on the basis of the 400–450-Mya liverwort fossil calibration, are considerable underestimates. It has been argued that such early molecular time estimates are contradicted by the absence of fossil pollen grains that are typical of certain groups of plants<sup>70</sup>. However, pollen preservation is subject to the same biases as other fossil evidence<sup>71,72</sup>, such that a rare plant group that is restricted in distribution might not leave a pollen trail in the fossil record. Molecular clocks are likely to shake the plant tree in the next few years.

**Fungi**

There are thought to be millions of living species of fungi although, because of difficulties in identifying them, only ~80,000 have been described<sup>73</sup>. This group of organisms includes those models of great importance for genetics — *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Neurospora crassa* and *Aspergillus nidulans* — as well as



**Figure 5 | A phylogeny of fungi. a** | The relationships of selected model fungi are shown, based on recent studies<sup>75,126</sup>. Branch lengths are not proportional to time. The divergence times (millions of years ago (Mya)  $\pm$  one standard error) shown<sup>67</sup> were estimated assuming an equal rate of substitution in animals and fungi. Other evidence indicates that fungi have a faster average substitution rate, so slightly younger time estimates have been obtained using a LOCAL-CLOCK METHOD (S.B.H., unpublished observations). **b** | The relationships and numbers of living species<sup>73</sup> in the main groups.

**PRECAMBRIAN**

An informal geological time period that spans from the time the Earth was born, ~4,500 million years ago (Mya), until ~545 Mya.

**LOCAL-CLOCK METHOD**

A method for estimating divergence time by accounting for differences in the rate of substitution among lineages (branches) in a tree.

animal and plant pathogens and agricultural species (FIG. 5). Their fossil record is poor because they have soft bodies that readily decompose<sup>74</sup> and because their sexual stage, which is important for identification, is rarely preserved. For these reasons, it is not a surprise that molecular-clock analyses of fungi have revealed large gaps in the fossil record and have pushed back the origins of major groups deep into the Precambrian<sup>67,75</sup>. If these dates are correct, then it is possible that fungi and their associates, plants, affected the Precambrian climate by increasing the rate of land weathering and the burying of decay-resistant carbon, which would have lowered global temperatures<sup>67</sup>. The oxygen that was produced by this early flora might have allowed animals to increase their

body size and form hard parts, possibly explaining the CAMBRIAN EXPLOSION of animals<sup>67</sup>.

The yeasts, which are unicellular ASCOMYCOTAN fungi, are considered to be the best organisms for the study of basic eukaryotic genetics. Baker's yeast (*Saccharomyces cerevisiae*, 5,600 genes) was the first eukaryote genome to be sequenced fully, and fission yeast (*Schizosaccharomyces pombe*, 4,940 genes) was the second fungal genome to be completed<sup>76,77</sup>. Molecular-clock analyses indicate that these two species are separated by about 1 Gya (FIG. 5), which is ~25% of the age of the Earth and nearly as old as the oldest taxonomically resolved fossil eukaryote<sup>78</sup>. However, this old age, if correct, would help to explain other features of these two yeast species, such as their great sequence divergence and the lack of conserved gene order<sup>77,79</sup>. Even the two widely used species of budding yeasts, *Saccharomyces cerevisiae* and *Candida albicans*, have, according to molecular-clock analyses, been separated for about 840 million years (FIG. 5).

The multicellular (filamentous) ascomycotan fungi have apparently larger genomes than the yeasts. Genome projects include the recently completed *Neurospora* (13,000 estimated genes)<sup>80</sup> genome and several species of the PYRENOZYCE *Aspergillus*. This latter genus is of economic importance in agriculture, biotechnology and medicine. *Aspergillus fumigatus*, the primary agent of aspergillosis, is the most common killer of bone-marrow transplant patients worldwide. By contrast, other species in the same genus are used to produce human foods, such as citric acid in soft drinks and the flavour in soy sauce. Knowing the phylogeny and timing of these species will help to identify mutations that cause traits of interest, and help to decipher the history and genetic basis of pathogenesis<sup>81</sup>.

**Animals**

About 1.1 million of the 1.5 million recognized living species of all organisms are animals<sup>82</sup>. Despite the volume of sequence data for animal model organisms, especially vertebrates, the relationships among some of these species (FIG. 6) remain hotly contested. For example, hundreds of homologous genes have been available for primates (*Homo sapiens*), rodents (*Mus musculus*), rabbits (*Oryctolagus cuniculus*) and ARTIODACTYL (*Bos taurus*). As early as 12 years ago<sup>83</sup>, multi-gene analyses indicated that rodents are the most divergent (basal branching) of these four groups. Despite the availability of more DNA and protein sequence data, this understanding did not change until recently, when sequence analyses, using a modest number of sites but an expanded number of species, indicated that rodents and rabbits are the closest relatives, with artiodactyls being the most distant<sup>84</sup>. Because a rodent-rabbit group agrees with morphology (for example, both have ever-growing incisors), one might assume intuitively that the new result is correct. However, the results of TAXON SAMPLING simulations<sup>85</sup> place greater importance on adding more genes than taxa, so the reason for these pronounced differences is not yet understood. Moreover, empirical results indicate that taxon and gene sampling contribute to phylogenetic resolution<sup>85</sup>. This issue will probably be



settled in the next year or two, with more vertebrate genome data and refined methods. If confirmed, it would mean that the mouse and rat are even closer genetic models to humans than previously believed.

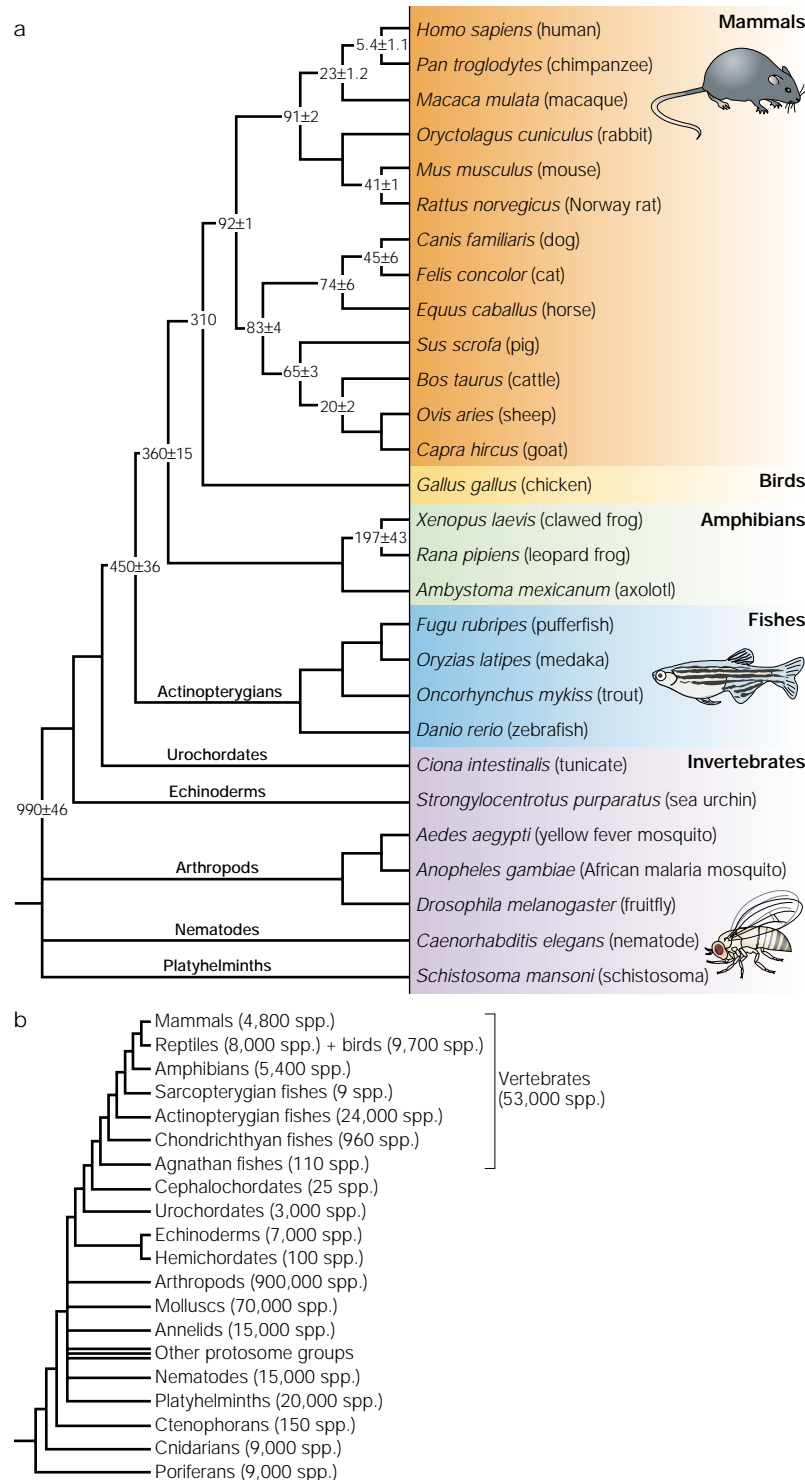


Figure 6 | **A phylogeny of animals.** **a** | The relationships and divergence times (millions of years ago (Mya) ± one standard error) of selected model animals are shown, based on recent multigene and multiprotein studies<sup>51,61,84</sup>. The fossil divergence time of birds and mammals (310 Mya) was used to calibrate the molecular clock. Branch lengths are not proportional to time. **b** | The relationships and numbers of living species, from a diversity of sources in most of the main groups.

Another area of disagreement concerns the relationships of three animals, the genomes of which have been completely sequenced — humans, fruitflies (*Drosophila melanogaster*) and nematodes (*Caenorhabditis elegans*). Historically, the nematode has been considered to be the most basal, partly because it has a PSEUDOCOELOM, whereas vertebrates and arthropods have a true COELOM. Sequence analyses typically resulted in the same grouping — in which fruitfly is together with human — termed ‘Coelomata’. However, in an analysis of nuclear small subunit rRNA sequences of various animals, a fruitfly–nematode group was obtained when a nematode (*Trichinella*), other than *C. elegans*, was used along with different tree-construction methods<sup>86</sup>. The ‘faster-evolving’ species of nematode (such as *C. elegans*) were thought to have artificially positioned the nematodes at the bottom of the tree. The new arthropod–nematode group was called Ecdysozoa (from the Greek, ‘to moult’) because it joined animals that shed their cuticle, although the cuticles of arthropods (chitin) and nematodes (collagen) are not homologous<sup>87</sup>. Subsequent studies have identified several ‘sequence signatures’ that support Ecdysozoa<sup>88,89</sup> and these have gained wide acceptance<sup>90,91</sup>. Nonetheless, comprehensive analyses of most gene and protein sequences, individually and combined, have supported Coelomata, not Ecdysozoa<sup>51,92,93</sup>. Recently, the Ecdysozoa hypothesis was tested using more than 100 protein alignments, which were ordered by rate of evolution, from slowest-evolving proteins to fastest-evolving proteins<sup>93</sup>. Supporters of the Ecdysozoa group predicted that the slowest-evolving proteins should support this hypothesis<sup>94</sup>, although the reverse was found, in which the slowest-evolving proteins supported Coelomata at high statistical confidence (100%). Moreover, other analyses that are designed to uncover substitution biases failed to support Ecdysozoa; this includes analyses that used only *Trichinella*, which is claimed to be a ‘slow-evolving species’ of nematode. If true, humans would provide a closer and better genetic and developmental comparison with *Drosophila* than the nematode. However, some prefer to wait until numerous sequences from large numbers of species become available before they consider the issue to be settled, in case taxon sampling is a factor.

Several time-related issues are also under debate. The split between humans and chimpanzees is of interest for, among other things, calibrating times of population divergence in our species<sup>4</sup>. However, even the most recent molecular-clock studies have resulted in widely spaced estimates for this split, ranging from 3.6–14 Mya<sup>95,96</sup>. The older dates, from mitochondrial DNA analyses, might have resulted from a methodological problem: slower-evolving non-primates were used to establish the rate (calibrate), which was then applied to the faster-evolving primate divergences, resulting in exaggerated times. Other mitochondrial analyses using primate rather than non-primate calibration points have obtained lower estimates (~5–6 Mya) of the divergence of humans and chimpanzees<sup>97</sup>. Time estimates based on nuclear genes and proteins, including the largest data sets, have also

## CAMBRIAN EXPLOSION

The sudden appearance, ~520 million years ago, of many major groups (phyla) of animals, as witnessed in the fossil record.

## ASCOMYCOTA

The largest phylum of fungi; also called ascomycetes or 'sac fungi'. They produce sexual spores in specialized sac-like cells called asci.

## PYRENOMYCETES

The largest subgroup of ascomycotan fungi, which are characterized by flask-shaped fruiting bodies.

## ARTIODACTYLS

Hoofed animals with an even number of digits. They belong to the mammalian Order Artiodactyla and include animals such as cattle, deer and pigs.

## TAXON SAMPLING

A term that indicates that the branching pattern of a tree might be influenced by the number or type of taxa (for example, species) included.

## PSEUDOCOELOM

Literally 'false cavity'; the body cavity of an animal, such as a nematode, that is not fully lined with mesodermal cells.

## COELOM

The body cavity of an animal, such as a vertebrate or insect, which is completely lined with mesodermal cells.

## CENOZOIC

The geological time period (era) that spans from 65 million years ago to the present day.

yielded lower time estimates (4.5–6.5 Mya) for this divergence<sup>61,97,98</sup>. These dates are compatible with the earliest known fossils of upright hominids (4.2 Mya)<sup>99,100</sup>, and older fossils (6–7 Mya) that might or might not be hominids<sup>101</sup>. The mouse–rat divergence time represents another point of disagreement, with molecular clocks indicating a deep split of 23–41 Mya (REFS 61,102). By contrast, evidence from fossils supports a 10–12 Mya split<sup>104</sup>. A three-fold increase in the rate of sequence change on the rodent lineage could reconcile these differences, but this has not yet been shown. However, accounting for base composition differences among lineages<sup>103</sup> reduces, rather than increases, the rate disparity. Resolving these rate differences will help to understand how point mutations typically accrue in the genome, either through errors during replication or in a time-dependent fashion.

The two main areas of disagreement between molecular clocks and the animal fossil record concern the radiation of mammal orders and animal phyla. The pattern is similar in each case: molecular clocks show much deeper divergences, indicating that either there are large gaps in the fossil record or that the clocks have run at different rates. In the case of placental mammals, fossils of most living orders first appear around the Cretaceous–Tertiary boundary (65 Mya), at about the time of the asteroid impact and dinosaur extinction<sup>104</sup>. This has long indicated that these mammals radiated in the CENOZOIC by filling niches vacated by the dinosaurs. The deep splits (80–100 Mya) among many orders, indicated by molecular clocks<sup>61,105</sup>, do not necessarily contradict an adaptive radiation in the Cenozoic, but indicate that their ancestors had already diverged from each other tens of millions of years earlier. The few fossils that existed during this time period<sup>106</sup> indicate that placentals were small, rodent- or rabbit-sized species and probably lived in humid forests and areas that are not conducive for preservation. The separation of a supercontinent into present-day continents during the Cretaceous (142–65 Mya) could have played a part<sup>105</sup>, as indicated by the association of ancient groups of mammals to continents<sup>107</sup>.

Molecular-clock analyses have indicated deep splits among animal phyla. For example, they indicate that a


split occurred 800–1,200 Mya between arthropods (*Drosophila melanogaster*) and vertebrates (*Homo sapiens*)<sup>51,108</sup>, despite the fossil record showing that the Cambrian explosion occurred 520 Mya. One interpretation is that lineage splitting occurred much earlier than the fossils indicate but that early animals were not well-preserved because they were much smaller, soft-bodied and perhaps restricted in distribution<sup>109</sup>. However, counter-arguments to that hypothesis have been made<sup>110–112</sup>. A presumed increase in atmospheric oxygen in the late Precambrian has been implied as the trigger that allowed animals to increase in size and form hard parts<sup>67,113,114</sup>. As with land plants and fungi<sup>67</sup>, these early molecular-clock dates have raised the possibility of a relationship with episodes of global glaciation (750–600 Mya)<sup>115</sup>; the existence of 'snowball Earths' remains a controversial topic among geologists. However, unlike the situation in mammals, no fossils, other than impressions of possible tracks<sup>116</sup> have been found from the early period of evolution (>650 Mya) that would support the molecular-time estimates. Therefore, the topic continues to engender lively debate that is certain to be stimulated in the future by genomic data from additional animal phyla.

## Conclusions

A significant leap in biological knowledge is underway and is driven by genomic data. The benefits have extended beyond medicine, industry and agriculture, to systematics and evolutionary biology. In particular, a framework for the tree and timescale of life is coming into focus through the analysis of genomic data from model organisms. In a reciprocal fashion, geneticists and others in the biological community are benefiting by this enhanced comparative framework. For example, the tree of life provides direction for finding genes and mutations of interest in close relatives, and the timescale of life is crucial for determining rates of substitution in genes and associations with environmental change. A focus on model organisms will continue to be productive, but the future is likely to see fusions of fields, cross-cutting initiatives and greater interdisciplinary exchange as a by-product of this renewed emphasis on comparative biology.

- Brenner, S. *et al.* Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* **366**, 265–268 (1993).
- Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).  
**Reviews how comparative biology methods that use phylogenies and molecular clocks can lead to remarkable insights into the evolution of life.**
- Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340–343 (2002).
- Ingman, M., Kaessmann, H., Pääbo, S. & Gyllenstein, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000).
- Hedges, S. B. & Kumar, S. Vertebrate genomes compared. *Science* **297**, 1283–1285 (2002).
- Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl Acad. Sci. USA* **87**, 4576–4579 (1990).
- Knoll, A. H. The early evolution of eukaryotes: a geological perspective. *Science* **256**, 622–627 (1992).
- Schopf, J. W. Microfossils of the early Archaean Apex chert: new evidence of the antiquity of life. *Science* **260**, 640–646 (1993).
- Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124–2128 (1999).  
**Describes how the finding of large amounts of horizontal gene transfer, as inferred from phylogenetic analyses of sequence data, has reshaped our view of the 'tree of life'.**
- Philippe, H. & Forterre, P. The rooting of the universal tree of life is not reliable. *J. Mol. Evol.* **49**, 509–523 (1999).
- Margulis, L. Archaeal–eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. *Proc. Natl Acad. Sci. USA* **93**, 1071–1076 (1996).
- Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
- Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74**, 5088–5090 (1977).
- Hansmann, S. & Martin, W. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *Int. J. Syst. Evol. Microbiol.* **50**, 1655–1663 (2000).
- Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. Universal trees based on large combined protein data sets. *Nature Genet.* **28**, 281–285 (2001).
- Hedges, S. B. *et al.* A genomic timescale for the origin of eukaryotes. *BMC Evol. Biol.* **1**, 4 (2001).
- 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).
- Brochier, C., Bapteste, E., Moreira, D. & Philippe, H. Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet.* **18**, 1–5 (2002).
- Matte-Tailliez, O., Brochier, C., Forterre, P. & Philippe, H. Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* **19**, 631–639 (2002).

20. Snel, B., Bork, P. & Huynen, M. A. Genome phylogeny based on gene content. *Nature Genet.* **21**, 108–110 (1999).
21. 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).
22. Tekala, F., Lazcano, A. & Dujon, B. The genomic tree as revealed from whole proteome comparisons. *Genome Res.* **9**, 550–557 (1999).
23. Daubin, V., Gouy, M. & Perriere, G. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.* **12**, 1080–1090 (2002).
24. Ragan, M. A. Detection of lateral gene transfer among microbial genomes. *Curr. Opin. Genet. Dev.* **11**, 620–626 (2001).
25. Rivera, M. C. & Lake, J. A. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* **257**, 74–76 (1992).
26. Cammarano, P., Creti, R., Sanangelantoni, A. M. & Palm, P. The Archaea monophyly issue: a phylogeny of translational elongation factor G(2) sequences inferred from an optimized selection of alignment positions. *J. Mol. Evol.* **49**, 524–537 (1999).
27. Faguy, D. M. & Doolittle, W. F. Genomics: lessons from the *Aeropyrum pernix* genome. *Curr. Biol.* **9**, R883–R886 (1999).
28. Tourasse, N. J. & Gouy, M. Accounting for evolutionary rate variation among sequence sites consistently changes universal phylogenies deduced from rRNA and protein-coding genes. *Mol. Phylogenet. Evol.* **13**, 159–168 (1999).
29. Katoh, K., Kuma, K. I. & Miyata, T. Genetic algorithm-based maximum-likelihood analysis for molecular phylogeny. *J. Mol. Evol.* **53**, 477–484 (2001).
30. Nisbet, E. G. & Sleep, N. H. The habitat and nature of early life. *Nature* **409**, 1083–1091 (2001).
31. Gallier, N., Tourasse, N. & Gouy, M. A nonhyperthermophilic common ancestor to extant life forms. *Science* **283**, 220–221 (1999).
32. Brochier, C. & Philippe, H. A non-hyperthermophilic ancestor for Bacteria. *Nature* **417**, 244 (2002).
33. Daubin, V., Gouy, M. & Perriere, G. Bacterial phylogeny using supertree approach. *Genome Informatics* **12**, 155–164 (2001).
34. Mojzsis, S. J. *et al.* Evidence for life on Earth before 3,800 million years ago. *Nature* **384**, 55–59 (1996).
35. Fedo, C. M. & Whitehouse, M. J. Metasomatic origin of quartz-pyroxene rock, Akilia, Greenland, and implications for Earth's earliest life. *Science* **296**, 1448–1452 (2002).
36. Schopf, J. W., Kudryavtsev, A. B., Agresti, D. G., Wdowiak, T. J. & Czaja, A. D. Laser-Raman imagery of Earth's earliest fossils. *Nature* **416**, 73–76 (2002).
37. Brasier, M. D. *et al.* Questioning the evidence for Earth's earliest fossils. *Nature* **416**, 76–81 (2002).
- Questions whether the 3.5-Gyr-old microfossils that were found in the Apex Chert rocks, in Australia (reference 8) are life forms. Reference 36 is a rebuttal to this paper and provides additional scrutiny of the same microfossils. These authors concur with one conclusion of reference 37, that the fossils are not of Cyanobacteria, but maintain that they are, nonetheless, fossils of microbes.**
38. Kollman, J. M. & Doolittle, R. F. Determining the relative rates of change for prokaryotic and eukaryotic proteins with anciently duplicated paralogs. *J. Mol. Evol.* **51**, 173–181 (2000).
39. Feng, D.-F., Cho, G. & Doolittle, R. F. Determining divergence times with a protein clock: update and reevaluation. *Proc. Natl Acad. Sci. USA* **94**, 13028–13033 (1997).
- An update of the influential 1996 Science paper from the laboratory of Russell Doolittle, one of the first to use large numbers of genes or proteins to date early events in the history of life.**
40. Summons, R. E., Jahnke, L. L., Hope, J. M. & Logan, G. A. 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature* **400**, 554–557 (1999).
41. Corliss, J. O. in *Nature and Human Society: the Quest for a Sustainable World* (ed. Raven, P. H.) 130–155 (The National Academy of Sciences, Washington DC, 2000).
42. Margulis, L. *Origin of Eukaryotic Cells* (Yale University Press, New Haven, Connecticut, 1970).
43. Gupta, R. S. Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among Archaeobacteria, Eubacteria, and eukaryotes. *Microbiol. Mol. Biol. Rev.* **62**, 1435–1491 (1998).
- Provides a detailed and often overlooked critique of the evidence bearing on the origin of mitochondria and on the number of symbiotic events (and gene transfers) that occurred in the origin of eukaryotes.**
44. Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A. & Peattie, D. A. Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*. *Science* **243**, 75–77 (1989).
45. Roger, A. Reconstructing early events in eukaryotic evolution. *Am. Nat.* **154**, S146–S163 (1999).
46. Horner, D. S. & Embley, T. M. Chaperonin 60 phylogeny provides further evidence for secondary loss of mitochondria among putative early-branching eukaryotes. *Mol. Biol. Evol.* **18**, 1970–1975 (2001).
47. Silberman, J. D. *et al.* Retortamonad flagellates are closely related to diplomonads: implications for the history of mitochondrial function in eukaryote evolution. *Mol. Biol. Evol.* **19**, 777–786 (2002).
48. Williams, B. A., Hirt, R. P., Lucocq, J. M. & Embley, T. M. A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*. *Nature* **418**, 865–869 (2002).
49. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414**, 450–453 (2001).
50. Keeling, P. J., Luker, M. A. & Palmer, J. D. Evidence from  $\beta$ -tubulin phylogeny that microsporidia evolved from within the fungi. *Mol. Biol. Evol.* **17**, 23–31 (2000).
51. Wang, D. Y.-C., Kumar, S. & Hedges, S. B. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B Biol. Sci.* **266**, 163–171 (1999).
52. 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).
53. Stechmann, A. & Cavalier-Smith, T. Rooting the eukaryote tree by using gene fusion. *Science* **297**, 89–91 (2002).
54. Moreira, D., LeGuyader, H. & Philippe, H. The origin of red algae and the evolution of chloroplasts. *Nature* **405**, 69–72 (2000).
- Provides strong evidence from several proteins that red algae and glaucocystophytes (glaucophytes) belong to the plant lineage, supporting a single origin of plastids.**
55. 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).
56. 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).
57. Stillier, J. W., Riley, J. & Hall, B. D. Are red algae plants? A critical evaluation of three key molecular data sets. *J. Mol. Evol.* **52**, 527–539 (2001).
58. Nickrent, D. L., Parkinson, C. L., Palmer, J. D. & Duff, R. J. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol. Biol. Evol.* **17**, 1885–1895 (2000).
59. Chaw, S. M., Parkinson, C. L., Cheng, Y., Vincent, T. M. & Palmer, J. D. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc. Natl Acad. Sci. USA* **97**, 4086–4091 (2000).
60. Oiu, Y.-L., Cho, Y., Cox, J. C. & Palmer, J. D. The gain of three mitochondrial introns identifies liverworts as the earliest land plants. *Nature* **394**, 671–674 (1998).
61. Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).
62. Takezaki, N., Rzhetsky, A. & Nei, M. Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.* **12**, 823–833 (1995).
63. Sanderson, M. J. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* **14**, 1218–1231 (1997).
64. Thorne, J. L., Kishino, H. & Painter, I. S. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**, 1647–1657 (1998).
65. Kishino, H., Thorne, J. L. & Bruno, W. J. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**, 352–361 (2001).
66. Sanderson, M. J. & Doyle, J. A. Sources of error and confidence intervals in estimating the age of angiosperms from *rbcl* and 18S rDNA data. *Am. J. Bot.* **88**, 1499–1516 (2001).
67. Heckman, D. S. *et al.* Molecular evidence for the early colonization of land by fungi and plants. *Science* **293**, 1129–1133 (2001).
68. Magallon, S. & Sanderson, M. J. Absolute diversification rates in angiosperm clades. *Evolution* **55**, 1762–1780 (2001).
69. Martin, W., Gierl, A. & Saedler, H. Molecular evidence for pre-Cretaceous angiosperm origins. *Nature* **339**, 46–48 (1989).
70. Crane, P. R., Friis, E. M. & Pedersen, K. R. The origin and early diversification of angiosperms. *Nature* **374**, 27–33 (1995).
71. Smith, A. B. *Systematics and the Fossil Record* (Blackwell Scientific, London, 1994).
72. Kenrick, P. & Crane, P. R. The origin and early evolution of plants on land. *Nature* **389**, 33–39 (1997).
73. Kirk, P. M., Cannon, P. F., David, J. C. & Stalpers, J. A. *Dictionary of Fungi* (CAB International, Surrey, UK, 2001).
74. Redecker, D., Kodner, R. & Graham, L. E. Glomalean fungi from the Ordovician. *Science* **289**, 1920–1921 (2000).
75. Berbee, M. L. & Taylor, J. W. in *The Mycota. VII. Systematics and Evolution* (eds McLaughlin, D. J. & McLaughlin, E.) 229–246 (Springer, New York, 2001).
76. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546–567 (1996).
77. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880 (2002).
78. Butterfield, N. J. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology* **26**, 386–404 (2000).
- Describes the oldest taxonomically resolved eukaryotic group (red algae): it arose 1.2 Gyr ago, and therefore has helped to constrain molecular clocks. The article also discusses the significance of this ancient group for understanding the origin of sex and multicellularity.**
79. Forsburg, S. L. The art and design of genetic screens: yeast. *Nature Rev. Genet.* **2**, 659–668 (2001).
80. Schulte, U., Becker, I., Mewes, H. W. & Mannhaupt, G. Large scale analysis of sequences from *Neurospora crassa*. *J. Biotechnol.* **94**, 3–13 (2002).
81. Berbee, M. L. The phylogeny of plant and animal pathogens in the Ascomycota. *Physiol. Mol. Plant Pathol.* **59**, 165–187 (2001).
82. May, R. M. in *Nature and Human Society: The Quest for a Sustainable World* (ed. Raven, P. H.) 30–45 (The National Academy of Sciences, Washington DC, 2000).
83. Li, W.-H., Gouy, M., Sharp, P. M., Ouhginn, C. & Yang, Y.-W. Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. *Proc. Natl Acad. Sci. USA* **87**, 6703–6707 (1990).
84. Murphy, W. J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348–2351 (2001).
- Presents a phylogenetic analysis of the most taxonomically diverse sequence data set for placental mammals.**
85. Rosenberg, M. S. & Kumar, S. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc. Natl Acad. Sci. USA* **98**, 10751–10756 (2001).
86. Aguinaldo, A. M. *et al.* Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**, 489–493 (1997).
- An influential paper that describes an analysis of sequences from the small subunit ribosomal RNA gene of animals. As a result, nematodes are placed together with arthropods in a controversial grouping dubbed 'Ecdysozoa' (see also reference 93).**
87. Adoutte, A. *et al.* The new animal phylogeny: reliability and implications. *Proc. Natl Acad. Sci. USA* **97**, 4453–4456 (2000).
88. deRosa, R. *et al.* *Hox* genes in brachiopods and priapulids and protosome evolution. *Nature* **399**, 772–776 (1999).
89. Manual, M., Kruse, M., Muller, W. E. G. & Parco, Y. L. The comparison of  $\beta$ -thymosin homologues among Metazoa supports an arthropod–nematode clade. *J. Mol. Evol.* **51**, 378–381 (2000).
90. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. *From DNA to Diversity* (Blackwell Science, Malden, Massachusetts, 2001).
91. Davidson, E. H. *Genomic Regulatory Systems* (Academic, San Diego, 2001).
92. Hausdorf, B. Early evolution of the bilateria. *Syst. Biol.* **49**, 130–142 (2000).
93. Blair, J. E., Ikeo, K., Gojobori, T. & Hedges, S. B. The evolutionary position of nematodes. *BMC Evol. Biol.* **2**, 7 (2002).
94. Mushegian, A. R., Garey, J. R., Martin, J. & Liu, L. X. Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genome Res.* **8**, 590–598 (1998).
95. Easteal, S. & Herbert, G. Molecular evidence from the nuclear genome for the time frame of human evolution. *J. Mol. Evol.* **44**, S121–S132 (1997).
96. Arnason, U., Gullberg, A., Burgeute, A. S. & Janke, A. Molecular estimates of primate divergences and new

- hypotheses for primate dispersal and the origin of modern humans. *Hereditas* **133**, 217–228 (2001).
97. Stauffer, R. L., Walker, A., Ryder, O. A., Lyons-Weiler, M. & Hedges, S. B. Human and ape molecular clocks and constraints on paleontological hypotheses. *J. Hered.* **92**, 469–474 (2001).
  98. Chen, F.-C. & Li, W.-H. Genomic divergences between humans and other hominoids and effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**, 444–456 (2001).
  99. Leakey, M. G., Feibel, C. S., McDougall, I. & Walker, A. New four-million-year-old hominid species from Kanapoi and Allia Bay, Kenya. *Nature* **376**, 565–571 (1995).
  100. Wood, B. Hominid revelations from Chad. *Nature* **418**, 134–135 (2002).
  101. Brunet, M. *et al.* A new hominid from the Upper Miocene of Chad, Central Africa. *Nature* **418**, 145–151 (2002).
  102. Adkins, R. M., Gelke, E. L., Rowe, D. & Honeycutt, R. L. Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. *Mol. Biol. Evol.* **18**, 777–791 (2001).
  103. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99**, 803–808 (2002).
  104. Benton, M. J. *Vertebrate Palaeontology* 452 (Blackwell Science, Oxford, 2000).
  105. Hedges, S. B., Parker, P. H., Sibley, C. G. & Kumar, S. Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**, 226–229 (1996).
  106. Archibald, J. D. Fossil evidence for a late Cretaceous origin of “hoofed” mammals. *Science* **272**, 1150–1153 (1996).
  107. Springer, M. S. *et al.* Endemic African mammals shake the phylogenetic tree. *Nature* **388**, 61–63 (1997).
- Sequence analyses define a superorder of mammals, now termed ‘Afrotheria’, that includes elephants, sea cows, hyraxes, aardvarks, golden moles and elephant shrews. After publication of this paper, tenrecs have also been added to this group. Support for the superorder continues to remain strong.**
108. Wray, G. A., Levinson, J. S. & Shapiro, L. H. Molecular evidence for deep Precambrian divergences among metazoan phyla. *Science* **274**, 568–573 (1996).
  109. Fortey, R. A., Briggs, D. E. G. & Wills, M. A. The Cambrian evolutionary ‘explosion’: decoupling cladogenesis from morphological disparity. *Biol. J. Linn. Soc. Lon.* **57**, 13–33 (1996).
  110. Valentine, J. W., Jablonski, D. & Erwin, D. H. Fossils, molecules and embryos: new perspectives on the Cambrian explosion. *Development* **126**, 851–859 (1999).
  111. Budd, G. E. & Jensen, S. A critical reappraisal of the fossil record of the bilaterian phyla. *Biol. Rev.* **75**, 253–295 (2000).
  112. Smith, A. B. & Peterson, K. J. Dating the time of origin of major clades: molecular clocks and the fossil record. *Annu. Rev. Earth Planet. Sci. Lon.* **30**, 65–88 (2002).
  113. Knoll, A. H. in *Early Life on Earth* (ed. Bengtson, S.) 439–449 (Columbia Univ. Press, New York, 1994).
  114. Knoll, A. H. & Carroll, S. B. Early animal evolution: emerging views from comparative biology and geology. *Science* **284**, 2129–2137 (1999).
  115. Hoffman, P. F., Kaufman, A. J., Halverson, G. P. & Schrag, D. P. A Neoproterozoic snowball Earth. *Science* **281**, 1342–1346 (1998).
  116. Rasmussen, B., Bengtson, S., Fletcher, I. R. & McNaughton, N. J. Discoidal impressions and trace-like fossils more than 1200 million years old. *Science* **296**, 1112–1115 (2002).
  117. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, New York, 2000).
  118. Rannala, B. & Yang, Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* **43**, 304–311 (1996).
  119. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
  120. Zuckerkandl, E. & Pauling, L. in *Horizons in Biochemistry* (eds Marsha, M. & Pullman, B.) 189–225 (Academic, New York, 1962).
  121. Sanderson, M. J. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19**, 101–109 (2002).
  122. Brocks, J. J., Logan, G. A., Buick, R. & Summons, R. E. Archean molecular fossils and the early rise of eukaryotes. *Science* **285**, 1033–1036 (1999).
  123. Han, T.-M. & Runnegar, B. Megascopic eukaryotic algae from the 2.1 billion-year-old Negaunee iron-formation, Michigan. *Science* **257**, 232–235 (1992).
  124. Javaux, E. J., Knoll, A. H. & Walter, M. R. Morphological and ecological complexity in early eukaryotic ecosystems. *Nature* **412**, 66–69 (2001).
  125. Wikström, N., Savolainen, V. & Chase, M. W. Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. Lond. B Biol. Sci.* **268**, 2211–2220 (2001).
  126. James, T. Y., Porter, D., Leander, C. A., Vilgalys, R. & Longcore, J. E. Molecular phylogenies of the Chytridiomycota supports the utility of ultrastructural data in chytrid systematics. *Can. J. Bot.* **78**, 336–350 (2000).
- Acknowledgements**  
I thank J. Blair, D. Geiser, S. Kumar, and D. Pisani for comments. I apologize to colleagues whose work could not be cited due to space constraints. Research in the author's laboratory is supported by the National Aeronautics and Space Administration (Astrobiology Institute) and National Science Foundation.
-  **Online links**
- FURTHER INFORMATION**  
**AlgaeBase:** <http://www.algaebase.org>  
**All species foundation:** <http://www.all-species.org/index.html>  
**Amphibian taxonomy:** <http://research.amnh.org/herpetology/amphibia>  
**Amphibiaweb:** <http://elib.cs.berkeley.edu/aw>  
**Angiosperm phylogeny:** <http://www.mobot.org/MOBOT/Research/APweb/welcome.html>  
**Animal diversity web:** <http://animaldiversity.ummz.umich.edu>
- Animal genome size database:** <http://www.genomesize.com>  
**Bird families:** <http://montereybay.com/creagrus/list.html>  
**Birds of the world:** <http://birdingonthe.net/sibmon/birdframe.html>  
**Blair Hedges laboratory:** <http://www.bio.psu.edu/faculty/hedges>  
**Deep hypha (phylogeny of Fungi):** <http://ocid.nacse.org/research/deephyphae>  
**Dictyostelium database (DictyBase):** <http://dictybase.org/dicty.html>  
**DOE microbial genomics gateway:** <http://www.microbialgenome.org>  
**Fishbase:** <http://www.fishbase.org/home.htm>  
**FlyBase (Drosophila):** <http://www.flybase.org>  
**Fossil record 2:** <http://palaeo.gly.bris.ac.uk/frwhole/FR2.html>  
**Generic model organism database:** <http://www.gmod.org>  
**Genome web:** <http://www.hgmp.mrc.ac.uk/GenomeWeb>  
**Human genome databases:** <http://www.hgmp.mrc.ac.uk/GenomeWeb/human-gen-db-genome.html>  
**Index to organism names:** [http://www.biosis.org.uk/free\\_resources/ion.html](http://www.biosis.org.uk/free_resources/ion.html)  
**International geologic timescale:** <http://micropress.org/stratigraphy/tscale.htm>  
**Land plants online:** <http://www.science.siu.edu/landplants>  
**List of bacterial names:** <http://www.bacterio.cict.fr>  
**Mammal species of the world:** <http://www.nmnh.si.edu/msw>  
**Model organism resources:** <http://www.cellbio.com/modelorgs.html>  
**NASA Astrobiology Institute:** <http://nai.arc.nasa.gov>  
**NASA evolutionary genomics web site:** <http://www.evogenomics.org>  
**NCBI genome databases:** <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome>  
**NCBI taxonomy browser:** <http://www.ncbi.nlm.nih.gov/Taxonomy>  
**NIH model organism database report:** <http://www.nhlbi.nih.gov/meetings/modeldb>  
**NIH model organism initiatives:** <http://www.nih.gov/science/models>  
**Phylogeny of life:** <http://www.ucmp.berkeley.edu/help/taxaform.html>  
**Phylogeny programs:** <http://evolution.genetics.washington.edu/phylip/software.html>  
**Sanger genome databases:** <http://www.sanger.ac.uk/Projects>  
**The EMBL reptile databases:** <http://www.embl-heidelberg.de/~uetz/LivingReptiles.html>  
**The global plant checklist:** <http://iopi.csu.edu.au/iopi/iopigpc1.html>  
**The international plant names index:** <http://www.ipni.org/index.html>  
**TIGR genome databases:** <http://www.tigr.org/tdb>  
**Tree of life web project:** <http://tolweb.org/tree/phylogeny.html>  
**Treebase:** <http://www.treebase.org/treebase/Index.html>  
**World taxonomist database:** <http://www.wetl.eti.bio.uva.nl/Database/WTD.html>  
**Wormbase (Caenorhabditis elegans):** <http://www.wormbase.org>  
Access to this interactive links box is free online.