We are family

Updating the tree of life needs both the skills of evolutionary biologists and the data from genome-crunchers — the two ignore each other at their peril. **John Whitfield** reports.

On 1 July 1858, in the Linnean Society of London's imposing neoclassical building on Piccadilly, biology changed for ever. That evening John Bennett, the society's secretary, read out papers by two biologists, Alfred Russel Wallace and Charles Darwin. From that point on, linnaean science was about more than describing and classifying living things. It was about using their traits to reveal their evolutionary relationships and assign them a position on life's family tree. Famously, such a tree was the only illustration in Darwin's *On the Origin of Species*, published the following year.

In the past 30 years, DNA sequencing has revolutionized this project. Comparing gene sequences has revealed not only a new domain of life, the archaea, but has affirmed that chimpanzees rather than gorillas are our closest relatives. And what was a trickle has become a flood — the major sequencing centres alone generate some 5,000 bases of DNA a second. About 1,000 species have now had their genomes sequenced, with more being published each month.

This torrent is set to reshape systematics — as the study of evolutionary relationships is known — once more. "Genomes are giving a much better view of the tree of life on Earth," says evolutionary biologist Mark Blaxter of the University of Edinburgh, UK. "The revolution is just starting — every new genome is causing rethinking." Blaxter is one of several researchers who has rallied under the banner of phylogenomics — the use of large quantities of genetic data (not just entire genomes) to build evolutionary trees, or phylogenies.

What has also become clear is that many problems cannot simply be battered into submission with more data. "Questions that are not resolved by a kilobase of sequence are seldom resolved by a megabase," says Jeffrey Boore of the Joint Genome Institute in Walnut Creek, California. At its best, phylogenomics is a two-way street — genome researchers need evolutionary biologists to help them work out the function of genes, and to avoid embarrassing mistakes.

Perhaps the most high-profile gaffe was the declaration by the Human Genome Project in 2001 that 100–200 genes in humans had come directly from bacteria. Analysis had revealed genes found in both humans and bacteria, but not in any species more closely related to humans. Project scientists concluded that such genes probably got into humans by lateral gene transfer from bacteria, a kind of inter-species sex where chunks of DNA cross from one cell to another. The result was heralded as one of the project's major revelations.

Jonathan Eisen first heard this finding while watching the press conference on TV at his workplace, the Institute for Genomic Research in Rockville, Maryland. "I felt sick to my stomach," he says. "They were talking about genes involved in brain development having come directly from bacteria. If you just think about that for a minute, it sounds so implausible." Sure enough, Eisen, and other similarly flabbergasted evolutionary biologists, rapidly shot the claim down, showing that the genes in question were more likely to have been present in the common ancestor of humans and bacteria but then lost in other lineages[1].

Updating Darwin: how do sequence data affect the family tree?

N. SPENCER/TEK IMAGE/SPL

This story is not the only example of the mistakes that can happen when genomics ignores evolution. "Molecular biologists are prone to not treating evolution as a particularly rigorous science," says Eisen, who now works at the University of California, Davis. "But you can't do good genome analysis without evolutionary analysis." Phylogenomics is about integrating the two, he says.

Form and function

Despite the shaky start, genome-sequencing centres now recognize the importance of evolution, and Eisen is one of several evolutionary biologists working at the heart of genomics. It's not just about keeping egg off faces. The techniques of evolutionary biology can also enhance other aspects of genome analysis — finding genes, and working out what is done by the proteins they encode.

Presented with a gene of unknown function, scientists can search for similar genes, of known function, in other sequences. A more rigorous approach is to use the sequence data to build a phylogeny of the different versions of that gene. This lets researchers track the gene's evolution, see how its function might have changed, and identify which other gene is most closely related to the target — which is not necessarily the one with the most similar sequence. Software packages can automate this process, and will predict protein function more accurately than doing a simple sequence comparison.

Tree-building is still the trickiest part. Systematists work out the relationships between genes, species or higher groups by searching for characters, be they morphological, biochemical or genetic, that arose in their last common ancestor and are shared by their descendants — lactation in mammals, for example. But even for just 10 species, there are more than 34 million possible evolutionary trees. Phylogenetic researchers

"The evolution of the animals has plagued us for years. We're hitting a wall." — Antonis Rokas

C. KENNEDY/KRT/NEWSCOM

Family strife: humans, fruitflies and nematodes were the first animals to have their complete genome sequenced — yet their exact evolutionary relationship still remains obscure.

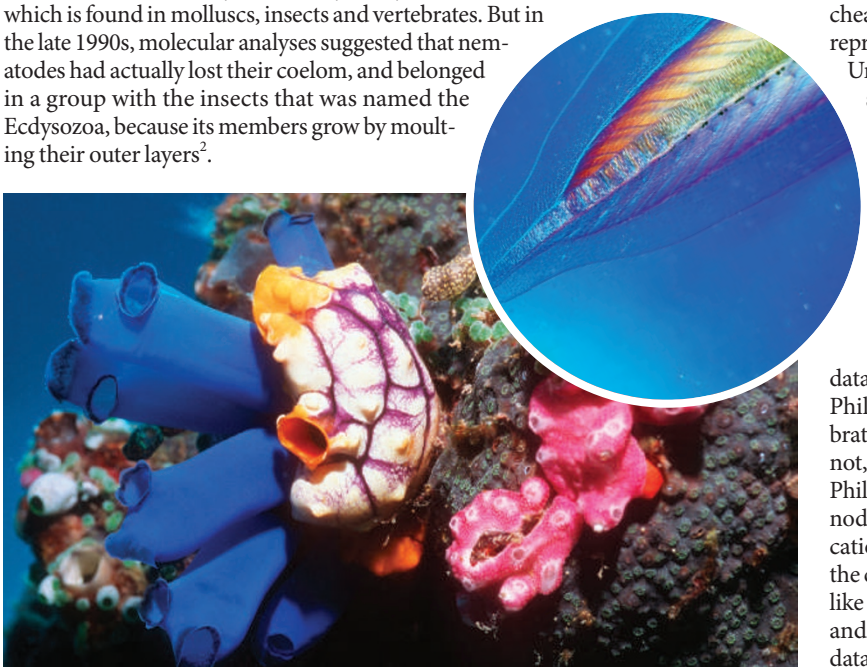
have developed complex algorithms to search among these possibilities, to find the tree most likely to reflect reality.

Picking the right tree is challenging. "For 15 years, people hoped for some software that could solve these problems at the push of a button, but sadly that's not come to pass. It's a very, very difficult problem algorithmically," says Boore. It is especially hard to say whether traits were inherited from a common ancestor, whether they look similar but evolved independently in unrelated groups — such as the wings of birds and bats — or whether they might have evolved but were then lost in later descendants, such as eyes in cave-dwelling fish. Often several different trees are equally good explanations of the data. And different genes from the same set of organisms often predict different trees.

Branching out

It is in the latter case that genome-sized datasets should be helpful. Analysing many genes at once allows the overall pattern of evolution to emerge, swamping the effect of quirky genes and confusing traits. One early success was in working out the relationships between groups of mammals, such as rodents, bats, carnivores and ungulates. These seem to have evolved very rapidly after the extinction of the dinosaurs, so their various common ancestors had little time to evolve unique characters that would mark their descendants. But multiple analyses, combining a few dozen genes from a cell's nucleus with the complete genomes of mitochondria, a cellular energy generator with its own small chromosome, have resulted in a robust consensus on who is related to whom.

Not everything is so tidy. Take the nematodes. For the past decade, a debate has swung back and forth about where this group of worms belongs. Anatomically simple organisms such as nematodes are a particular headache for systematists, as they offer few clues to their ancestral state. Based on morphology, nematodes were placed close to the root of the tree of animals, because they lack a body cavity called a coelom, which is found in molluscs, insects and vertebrates. But in the late 1990s, molecular analyses suggested that nematodes had actually lost their coelom, and belonged in a group with the insects that was named the Ecdysozoa, because its members grow by moulting their outer layers[2].



G. DOUWMA/SPL

Sea squirts briefly usurped cephalochordates (inset) as the closest relatives of vertebrates.

Where to put nematodes, and how they relate to insects, matters to genomics more broadly because the worm *Caenorhabditis elegans* and the fruitfly *Drosophila melanogaster* are two of the most important model animals. If the Ecdysozoa is the true group, both are equally distant from humans. If not, the fly is more like us than the worm. Humans, *C. elegans* and *D. melanogaster* were among the first animals to have their genomes sequenced. These complete genomes seemed to tell a different story from the earlier genetic analysis that put forward the Ecdysozoa. Or rather, they retold the old story. A tree built from the human, fly and worm genomes[3] makes the nematodes the outsiders.

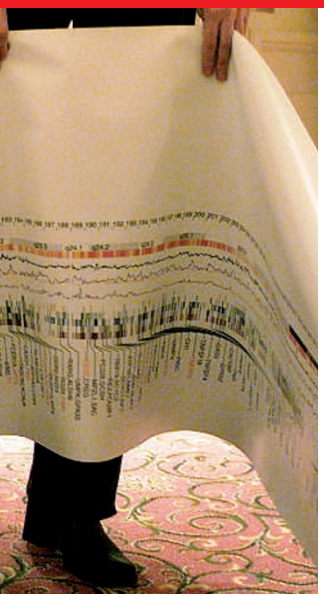
Express yourself

One way to choose between contradictory trees is to add more species to the analysis. With animals, one soon runs out of sequenced genomes to add — but there is a third way between having some data for lots of species and lots of data for a few species. This is to use expressed sequence tags (ESTs), which are DNA sequences made from the genes that are active in cells. EST libraries can provide data on several hundred genes for a species relatively quickly and cheaply, allowing more species to be analysed. Using ESTs representing nearly 150 genes, Hervé Philippe's team at the University of Montreal, Canada, built a tree containing 35 species that supported the existence of the Ecdysozoa[4].

C. NURIDSANY & M. PERENNOU/SPL

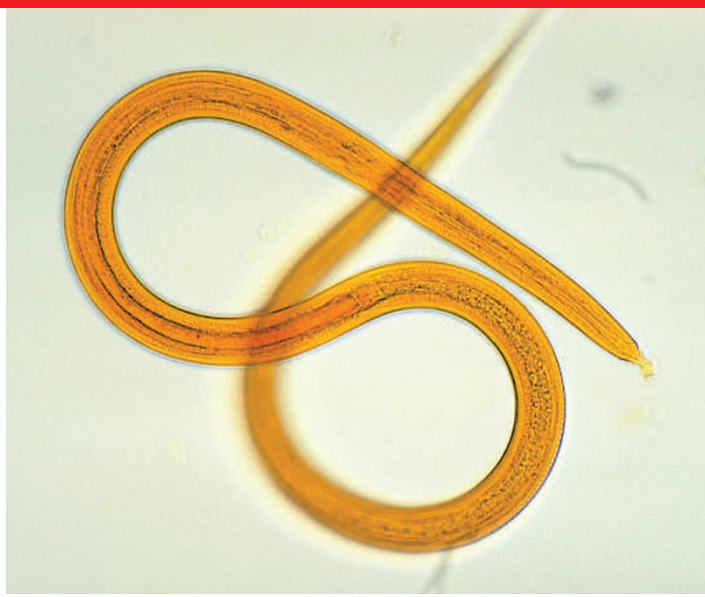
Most phylogenomics researchers are now inclined to believe the Ecdysozoa data. But enough analyses suggest the contrary for the question to remain open, says Antonis Rokas, a genomics researcher at the Massachusetts Institute of Technology in Cambridge. "I would be agnostic," he says. "Just looking at the phylogenetic evidence, I'm not convinced that either side has it right."

This seesawing has become common, as accumulating data tip an argument first one way, then the other. Last year, Philippe's team suggested that the closest relatives of vertebrates were the sea squirts, which resemble bags of jelly, and not, as previously thought, the fish-like cephalochordates[5]. Philippe's work grouped the cephalochordates with the echinoderms (sea urchins and the like) with the troubling implication that either fish-like creatures evolved twice, or that the common ancestor of vertebrates and sea urchins looked like a fish. In November, evolutionary biologist Max Telford and his colleagues at University College London, added a dataset of 35,000 amino acids from other groups, including starfish. The new tree put the cephalochordates back where



GRAPHIC SCIENCE/ALAMY

L. TELFORD



M. MELVIN/CDC

they started[6]. "There was a big sigh of relief," says Telford. "Not even Philippe had much faith in the first result."

In general, relations between the animal phyla — large groups such as molluscs or arthropods — remain a mystery. "The evolution of the animals has plagued us for years," says Rokas. "We're hitting a wall." The animal phyla all seem to have evolved very rapidly some 600 million years ago and the signal in DNA data has eroded. Rokas has compared evolutionary trees of fungi and animals built using the same genes[7]. Those from fungi, whose genes are thought to have changed at a steady pace, are well resolved. The animal trees are rickety.

Will more sequences help? "There are some problems that you could throw a whole genome at and will still probably be unresolved," says Telford. He gives the example of predatory marine worms called the chaetognaths, or arrow worms, which refused to reveal their evolutionary allegiance even when more than 70 genes were analysed[8]. "We had more data than you can shake a stick at, and we still weren't able to place them," he says.

Location, location, location

Where sequence fails, other features of the genome might help. The positions of genes, for example, are also inherited, and changes in gene order can mark evolutionary splits. Boore pioneered this approach, using the order of genes on mitochondria to show that insects and crustaceans are closer to each other than either is to spiders or centipedes[9].

Similarly, the points at which the 'jumping genes' called transposons insert themselves into sequences have been used to reconstruct the history of the mammals[10], and the positions of introns — gene pieces that are cut out before DNA is turned into protein — have been used to support the existence of the Ecdysozoa[11]. When a rare genetic event is shared among a group, it's a sure sign that they had a common ancestor, says Boore.

The usefulness of such approaches is not yet universally accepted. "I have found many genome features don't work well for evolutionary reconstruction — they're too prone to convergent evolution," says Eisen, explaining that the same gene order probably often arises independently in different groups. Philippe believes that sequence data are still the most reliable, but that other approaches will mature in time.

Part of the problem may be that most of the genomes sequenced so far, besides humans, have been lab models, diseases or economically important organisms. None is ideal for understanding evolution — model species and crops have quick generation times, so they are among the fastest evolving species. "The organisms chosen as models have got biological properties that mean, genomically, they're odd," says Blaxter.

"You could throw a whole genome at some problems and they will still probably be unresolved." — Max Telford

But evolutionary biologists are hoping that genome sequencers will soon switch focus. "A lot of genome projects are finishing and the centres are looking for something to do," says Tim Littlewood of the Natural History Museum in London. One target might be the eukaryote tree. The eukaryote domain covers organisms whose cells are divided into compartments, including such familiar kingdoms as animals, plants and fungi, but also a plethora of single-celled organisms called protists. No one has any idea which of these groups is the closest ancestor to the protist that made the leap to multicellularity and gave rise to the animals. "One of the most interesting questions that phylogenomics can address is of the major lines of eukaryotic evolution," says Philippe.

But evolutionary puzzles are not going to drive sequencing agendas on their own, says Boore. One fear is that once all the crops and pathogens are sequenced, the cash will dry up. "So far, the appetite for comparative genomics has been surprising," Boore says, "but it's hard to predict how things are going to go." When selecting organisms for sequencing, the potential for illuminating human biology and disease are strong arguments. Fussing over the lineage of obscure marine species is less compelling.

"Genomics people recognize the need for systematics," argues Littlewood. If systematists — whose discipline labours under a fusty, arcane image — show that voguish molecular research depends on their work, it should help reverse the decline in jobs and funding for the discipline, he says. "In many ways, genomics stands to learn more from systematics than vice versa — the onus is on systematists to point out that we can help." ■

John Whitfield is a science writer based in London.

1. Genereux, D. P. & Logsdon Jr, J. M. *Trends Genet.* **19,** 191–195 (2003).
2. Aguinaldo, A. M. A. *et al. Nature* **387,** 489–493 (1997)
3. Blair, J. E., Ikeo, K., Gojobori, T. & Hedges, S. B. *BMC Evol. Biol.* **2,** 7 (2002).
4. Philippe, H., Lartillot, N. & Brinkmann, H. *Mol. Biol. Evol.* **22,** 1246–1253 (2005).
5. Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. *Nature* **439,** 965–968 (2006).
6. Bourlat, S. J. *et al. Nature* **444,** 85–88 (2006).
7. Rokas, A., Krüger, D. & Carroll, S. B. *Science* **310,** 1933–1938 (2005).
8. Matus, D. Q. *et al. Curr. Biol.* **16,** R575–R576 (2006).
9. Boore, J. L., Collins, T. M., Stanton, D., Daehler, L L. & Brown, W. M. *Nature* **376,** 163–165 (1995).
10. Nishihara, H., Hasegawa, M. & Okada, N. *Proc. Natl Acad. Sci. USA* **103,** 9929–9934 (2006).
11. Roy, S. W. & Gilbert, W. *Proc. Natl Acad. Sci. USA* **102,** 4403–4408 (2005).