Michael J. Donoghue

33

Immeasurable Progress on the Tree of Life

In listening to the Assembling the Tree of Life (ATOL) symposium in New York, and in reading the manuscripts for this volume, I was overwhelmed by the enormous progress that we have made, over such a short time, on what Darwin so aptly called "the great Tree of Life." The word "immeasurable"—in the dictionary sense of "indefinitely extensive"—seems to apply perfectly to this situation. But what about the other, more literal, meaning of the word immeasurable? Is phylogenetic progress also "incapable of being measured"? This is the question I want to address. My sense is that there are many facets of "progress" that matter to us and that we would like to be able to measure. For some of these we can devise proper metrics, and we might even be able to provide concrete numbers. For others, as I'll argue, we aren't even entirely sure what we'd like to measure, and we're still a long way from being able to quantify how we are doing.

Let me back up, and ask, What are the ways we might think about expressing progress—to measure where we stand now in relation to where we were a decade ago and where we hope to end up? One possibility would be to tally the number of known species on Earth that have been included in bone fide phylogenetic analyses [in December 2003 there were almost 35,000 species represented in TreeBASE (available at http://www.treebase.org), but the real number might be more like 80,000], or maybe even the number that could potentially be included today if we harnessed all of the data in relevant databases [e.g., DNA sequences in GenBank (available at http://www.ncbi.nlm.nih.gov)]. Another possibility would be to chart trends in the number of phylogenetic papers published over the years (e.g., Sanderson et al. 1993; Hillis, ch. 32 in this vol.).

These are certainly interesting measures, and the numbers, insofar as we know them, certainly do bolster the gut-level feeling that we're making lots of progress. They don't, however, capture much about the nature and the quality of what's being learned. Maybe we should also be gauging our coverage of the Tree of Life in terms of the number of major lineages represented by some reasonable number of exemplars, or perhaps we should somehow represent the size and the variety of the data sets that are being analyzed. Or, perhaps a metric is needed to reflect changing levels of confidence in the clades being identified. Another worthy measure, for very obvious purposes, would gauge how many phylogenetic studies have provided solutions to practical problems. Success stories along these lines abound—identifying the source of an emerging infectious disease, pointing the way toward crop improvement, orienting the search for new pharmaceuticals, and so on (see Yates et al., ch. 1 in this vol.; examples of the practical importance of phylogenetic research are also highlighted in a brochure sponsored by the National Science Foundation (Cracraft et al. 2002). But how do we attach a number to such achievements? Patents pending, perhaps, although this would record only a small fraction of the successes.

Ultimately, I think we would all like a measure that captures how phylogenetic studies have affected our understand-

ing of life—how the living world is structured, how it works, and how it has come about. At first glance this truly does seems immeasurable, in the "not-capable-of-measurement" sense of the word. But on second thought, maybe there is a reasonably good proxy for this, which takes us back to Willi Hennig (e.g., Hennig 1966). What if we could faithfully tally up cases in which traditionally recognized taxonomic groups had been convincingly demonstrated to be paraphyletic? Paraphyletic groups are ones that contain an (inferred) ancestor and some, but not all, of its descendants. In practice, of course, paraphyly is "discovered" when a phylogenetic analysis identifies one or more new clades that unite some of the lineages previously assigned to the traditional group with one or more lineages placed outside of that group. In other words, the "negative" discovery of paraphyly *is precisely* the "positive" discovery of new "cross-cutting" clades.

Before we think about whether we could actually count up discoveries of paraphyly, let's contemplate why this might be a satisfying measure of phylogenetic progress. First of all, it's worth noting that this measure relates how changes in our knowledge of phylogenetic relationships have affected the application of taxonomic names, and as such, it can potentially be assessed everywhere in the Tree of Life, from the very base out to the tips, without needing to refer to particular groups or their characters. In this sense, it is a measure without units. Second, it registers a change in the language that we use to describe the structure of diversity, which can deeply (although often quite subtly) influence the way we perceive diversity, orient our research, and teach. Third, the discovery of paraphyly has immediate impacts on our understanding of character evolution. Some characters previously thought to have evolved convergently are seen instead to be homologous—to have evolved only once, in the inferred ancestor of a newly discovered cross-cutting clade. Even more generally, the recognition of paraphyly allows us to infer a sequence of evolutionary events, which helps fill in what appeared to be major gaps between traditional taxa. Often this is just the information we need to choose among competing evolutionary hypotheses about how and why major transitions occurred. In many of the same ways, of course, such discoveries also help us make sense of biogeography. Fourth, such discoveries generally change the way we perceive shifts in diversification, especially by accentuating differences in the number of species between sister groups.

Putting the third and fourth points together, my guess is that discoveries of paraphyly will eventually have even more profound impacts on how we view the connection between character change and diversification. In particular, I think we'll be forced to develop a more nuanced (and more productive) view of "key innovations." It will become increasingly natural to think from the outset about a series of changes culminating in a combination of traits that ultimately affected diversification. Rather than simply moving the causal explanation down a node or two in the phylogeny, this distributes the causation across a series of nodes and character changes. Also, increasingly we'll focus on how apparently subtle changes early in such a chain rendered new morphological designs accessible, which in turn enabled the evolution of the traits that we most often associate with the success of clades, with ecological transitions, and so forth.

To illustrate these points, let's look at green plants. Figure 33.1 provides an overview of our present knowledge of phylogenetic relationships among the major lineages—highly simplified, of course, and consciously pruned (rendered pectinate) to serve my purposes (see O'Hara 1992 for a general discussion of such simplifications). Several widely known traditional groups are supported as monophyletic in all recent analyses, including the entire green plant clade (viridophytes), land plants (embryophytes), vascular plants (tracheophytes), seed plants (spermatophytes), flowering plants (angiosperms), and monocotyledons (monocots). A number of other traditionally recognized groups have repeatedly been determined to be paraphyletic, confirming suspicions that they represent grades of organization, diagnosed only by ancestral features of the more inclusive clades to which they belong. Specifically, "green algae," "bryophytes," "pteridophytes," "gymnosperms," and "dicotyledons" all appear to be paraphyletic. In each case, one or more new clades were discovered that linked some lineages traditionally assigned to the group to related taxa. So, for example, the streptophyte and charophyte clades (as circumscribed here; for an alternative, see Delwiche et al., ch. 9 in this vol.) include lineages that used to be assigned to the green algae (the Charophyta in the traditional sense) along with the land plant clade. Likewise, the euphyllophyte clade unites all extant lineages of seedless vascular plants, except the lycophytes, with the seed plants, and so on. In the case of the "bryophytes" and the "gymnosperms," names were proposed for new cross-cutting clades ("stomatophytes" and "anthophytes," respectively), but recent analyses have cast doubt on their existence (see Nickrent et al. 2000, Donoghue and Doyle 2000). Nevertheless, in both cases it remains quite clear that these traditional groups are paraphyletic (see Delwiche et al., ch. 9, and Pryer et al., ch. 10 in this vol.).

The impact of these discoveries on our understanding has been enormous. The most obvious and immediate effect was on our ability to dissect the evolutionary sequence of events surrounding the greatest transformations in plant history. For example, take the transition from living in water to living on land (see Graham 1993). Before we recognized the paraphyly of green algae and of bryophytes, this shift appeared to entail a large number of steps, which we had no real basis for putting in order. This implied either many extinctions and, consequently, gaps in our knowledge, or else some sort of wholesale transformation from one life form to another. Under these circumstances, alternative theories emerged and remained viable. What kind of environment did the immediate ancestors of the land plants live in, and what did they look like? After all, "green algae" live in saltwater or in freshwater; may be unicells, colonies, filaments, or more complex

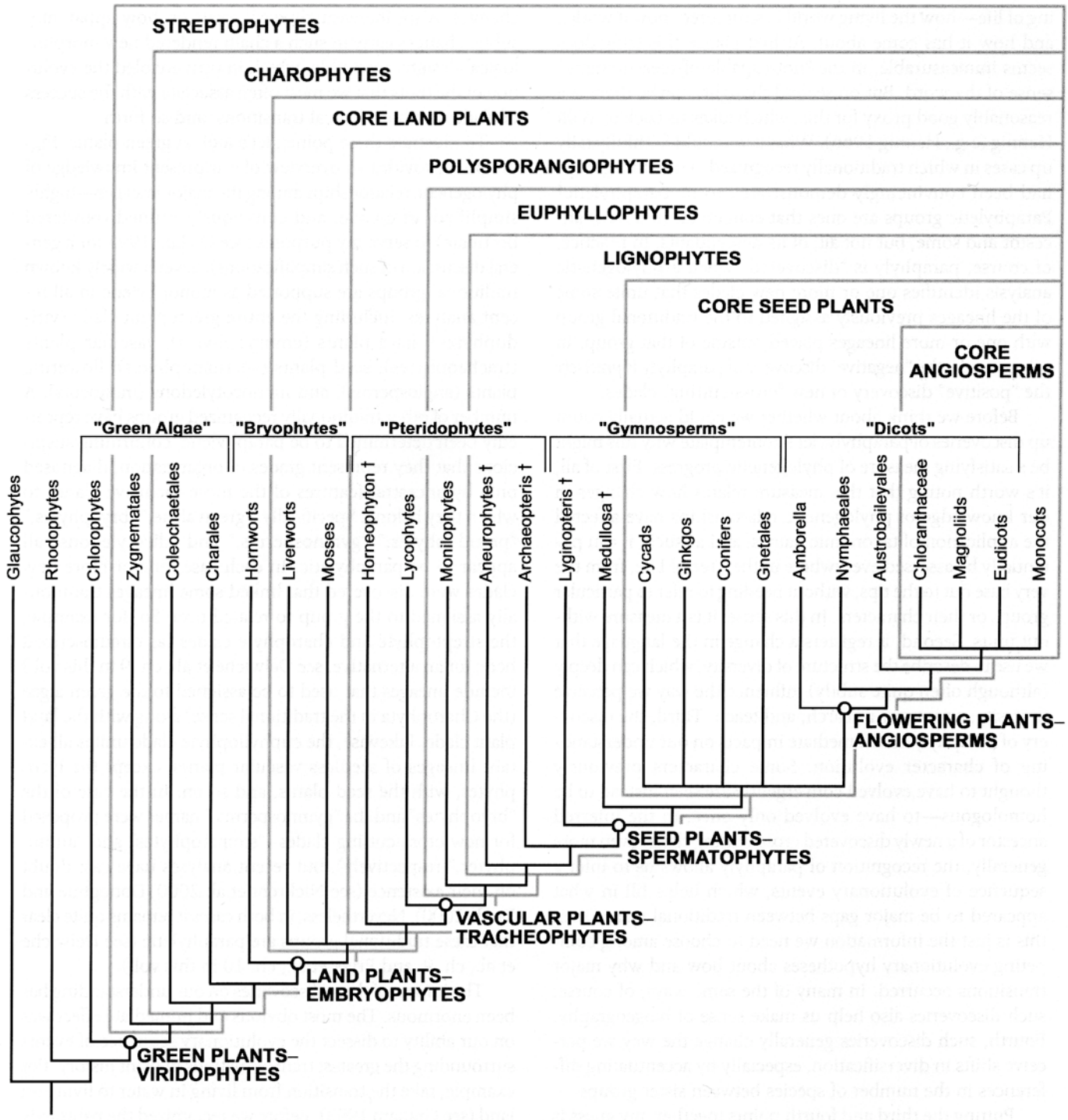

Figure 33.1. An overview of green plant phylogeny, illustrating progress through the recognition and abandonment of paraphyletic groups (e.g., "green algae" and "bryophytes") with the discovery of new major clades (e.g., streptophytes and euphyllophytes). For references to the primary literature, underlying evidence, levels of support, outstanding controversies, and additional evolutionary implications, see Kenrick and Crane (1997), Doyle (1998), Donoghue (2002), Judd et al. (2002, ch. 7), and chapters 9–11 in this volume. Note that Delwiche et al. (ch. 9 in this vol.; also Karol et al. 2001) use the name "Charophyta" for the clade here referred to as the streptophytes. The usage adopted here may better reflect original intentions (e.g., Bremer and Wanntorp 1981) and subsequent usage (e.g., Kenrick and Crane 1997); in any case, such nomenclatural problems highlight the desirability of providing explicit phylogenetic definitions for clade names.

forms; may or may not have cell walls separating the nuclei; and so on. And what about the evolution of the land plant life cycle—alternating between multicellular haploid (gametophyte) and diploid (sporophyte) phases? In short, the transition to land largely remained a mystery.

With the discovery of a series of intervening clades (fig. 33.1; Karol et al. 2001; see Delwiche et al., ch. 9 in this vol.), we're now able to infer a sequence of events from the first green plants through the transition to land. We can be quite certain that their immediate ancestors lived in freshwater, probably quite close to the shore; had rather complex parenchymatous construction; and bore eggs (and zygotes) on the parent plant in specialized containers. Likewise, we can finally put to rest the debate about the life cycle: the land plant life cycle originated through the intercalation of a multicellular diploid phase (through delayed meiosis) into an ancestral life cycle in which the diploid zygote underwent meiosis directly to yield haploid spores.

This example is meant only to illustrate the sorts of insights that can follow the discovery of paraphyly, and so to justify such a measure of progress. What can we say, then, about the number of these discoveries in recent years, or about our expectations in the future? In *The Hierarchy of Life* (Fernholm et al. 1989), the last major attempt to take stock of phylogenetic progress, Gareth Nelson remarked: "Paraphyly, it would seem, is the most common discovery of modern systematic research" (Nelson 1989: 326). This may well be true, but is there a way to put a number on it? Sadly, aside from asking experts on each major clade to come up with a list (or an account along the lines of fig. 33.1), we aren't really able to do this. We haven't been keeping track in any systematic way and, as I will argue, we haven't developed the necessary informatics tools.

Let us suppose that we wanted to be able to tally up those changes in knowledge of phylogeny that significantly changed our view of the world, and that for this purpose we wanted to focus on discoveries that changed the way that taxonomic names are used. Specifically, we would be looking for cases in which the name of a paraphyletic group had been abandoned altogether, or the circumscription had been adjusted so that the name again referred to a hypothesized clade. These are what might be called "meaningful" taxonomic changes, to distinguish them from other sorts of name changes. We would want to avoid, for example, changes only in the Linnaean taxonomic rank that a group is assigned (e.g., a shift from Family to Order). As things now stand, such rank assignments are fundamentally arbitrary, yet our nomenclatural codes are intimately tied to them, and in some cases a cascade of name changes can be required without any underlying advance in our knowledge of phylogeny. Also, it's important to note that quite a few clades are discovered and named that don't contradict the monophyly of any previously named taxon—instead, they resolve bits of the Tree of Life that were more or less unresolved and to which taxonomic names had not been applied. The point is that the problem is not as "simple" as just tracking changes in the names being used in the taxonomic literature.

What we really are talking about is tracking changes in the relationship between taxonomic names and hypothesized clades. If we knew how taxonomic names mapped onto a tree at some initial time, we could see at a later time how many names applied to the same clades versus how many no longer applied to clades but to paraphyletic groups. To do this in practice, one would need, first of all, a database that recorded changes in our knowledge of phylogeny. TreeBASE is designed for this purpose, but unfortunately, it still isn't used consistently enough by the authors of phylogenetic papers. One presumes that this will improve (probably driven by more journals requiring the submission of phylogenetic data and results), in which case we will automatically develop the record we need to make solid tree comparisons over time.

But tracking trees is only one part of the problem. The other is to understand how names have been used at different times. Although for some groups of organisms there are databases that keep track of all the names that have ever been published (e.g., the International Plant Names Index, available at http://www.ipni.org), or even of the accepted names and synonyms (e.g., Species 2000, available at http://www.sp2000.org), it's hard to say exactly how these names correspond to hypothesized clades at any one time, much less at different times. The problem is that taxonomic names have not traditionally been defined in such a way that we can be sure whether they were even meant to refer to clades (sometimes, mostly in the past, names were knowingly applied to paraphyletic groups) or, if so, which lineages were intended to be included (even assuming complete agreement on phylogenetic relationships). Of course, we could get better about designating how names are meant to coincide with clades by, for example, consistently labeling clades in TreeBASE. This would be a step in the right direction, but it would be even better to adopt a nomenclatural system in which the connection between a taxonomic name and a hypothesized clade needed to be precisely defined at the outset. Here I am referring to "node-based" and "stem-based" definitions and other conventions discussed in relation to the PhyloCode (available at http://www.phylocode.org). Interestingly, taxonomic names under such a system tend to be maintained in the face of changes in phylogenetic knowledge, although with a different composition of lineages. Specifically, the name of a taxon discovered to be paraphyletic might well be retained for a more inclusive clade, unless it happened to become synonymous with a preexisting name. Overall, it is hard to say how the turnover of names would compare between the PhyloCode and our traditional nomenclature codes, where names are neither defined with respect to a tree nor fixed in terms of content.

The conclusion I draw from the above is that the actual abandonment of the names of paraphyletic groups is probably not going to be a very sensitive measure (under either traditional nomenclature or under the PhyloCode). Names

can be retained and reconfigured in various ways, and in any case it would be hard to judge when a particular name had finally been dropped by the relevant taxonomic community. In the end, what we really want, regardless of "abandonment," is a database designed such that we can identify those phylogenetic discoveries that change how names map onto trees—whether a name refers to the same clade at different times or whether it can be made to refer to a clade only by changing the content to include lineages previously viewed as being outside the group. This would be a pretty sophisticated database, but I see no reason why it couldn't be developed.

My point is that it's time we attended to the business of naming clades and to the informatics issues surrounding the Tree of Life project. As Hennig stressed, "Investigation of the phylogenetic relationship between all existing species and the expression of the results of this research, in a form which cannot be misunderstood, is the task of phylogenetic systematics" (Hennig 1965: 97). Progress on the first of these goals—understanding phylogenetic relationships—has certainly been impressive. By comparison, progress on the second goal—expressing the results in a form that cannot be misunderstood—has been rather pathetic. Much of what we have learned about relationships has not been translated into the taxonomic language used to describe the diversity of Life. And much of what we have learned has not been properly incorporated into databases, so the effort is effectively wasted. I hope we have made real progress along these lines before we take stock again of the Tree of Life.

In summary, at this moment it strikes me that phylogenetic progress is immeasurable in both senses of the word—phylogenetic knowledge is expanding at a mind-boggling rate *and* we don't yet have the tools to measure this in the ways we would like. When we are eventually able to make measurements of the sort I have described, we will have achieved something truly monumental. We will certainly have charted much more of the Tree of Life, but we will also have changed the language we use to communicate about biological diversity and, therefore, how we think about the world. Perhaps most important, we will have rendered this knowledge widely accessible and prepared it for the queries that will propel the Tree of Life project to the next level. "Indefinitely extensive" will have become the only applicable meaning of "immeasurable."

Acknowledgments

I am grateful to Joel Cracraft for his leading role in organizing the symposium and editing the proceedings, and to the other speakers in the session on plants—Chuck Delwiche, Kathleen Pryer, and Pam Soltis. I have benefited from discussion of these issues with Susan Donoghue and Kevin de Queiroz. For their help with my presentation at the symposium and with figure 33.1, I am indebted to Brian Moore and Mary Walsh. Yale University, through Provost Alison Richard, generously supported the symposium and the participation of Yale students.

Literature Cited

Bremer, K., and H.-E. Wanntorp. 1981. A cladistic classification of green plants. Nord. J. Bot. 1:1–3.

Cracraft, J., M. Donoghue, J. Dragoo, D. Hillis, and T. Yates (eds.). 2002. Assembling the tree of life: harnessing life's history to benefit science and society. National Science Foundation. Available: http://ucjeps.berkeley.edu/tol.pdf. Last accessed 25 December 2003.

Donoghue, M. J. 2002. Plants. Pp. 911–918 *in* Encyclopedia of evolution (M. Pagel, ed.), vol. 2. Oxford University Press, Oxford.

Donoghue, M. J., and J. A. Doyle. 2000. Demise of the anthophyte hypothesis? Curr. Biol. 10:R106–R109.

Doyle, J. A. 1998. Phylogeny of the vascular plants. Annu. Rev. Ecol. Syst. 29:567–599.

Fernholm, B., K. Bremer, and H. Jörnvall (eds.). 1989. The hierarchy of life. Nobel Symposium 70. Elsevier, Amsterdam.

Graham, L. E. 1993. Origin of the land plants. Wiley, New York.

Hennig, W. 1965. Phylogenetic systematics. Annu. Rev. Entomol. 10:97–116.

Hennig, W. 1966. Phylogenetic systematics. University of Illinois Press, Champaign-Urbana.

Judd, W. S., C. S. Campbell, E. A. Kellogg, P. F. Stevens, and M. J. Donoghue. 2002. Plant systematics: a phylogenetic approach. 2nd ed. Sinauer, Sunderland, MA..

Karol, K. G., R. M. McCourt, M. T. Cimino, and C. F. Delwiche. 2001. The closest living relatives of land plants. Science 294:2351–2353.

Kenrick, P, and P. R. Crane. 1997. The origin and early diversification of land plants: a cladistic study. Smithsonian Institution Press, Washington, DC.

Nelson, G. 1989. Phylogeny of the major fish groups. Pp. 325–336 *in* The hierarchy of life (B. Fernholm, K. Bremer, and H. Jörnvall, eds.). Nobel Symposium 70. Elsevier, Amsterdam.

Nickrent, D., C. L. Parkinson, J. D. Palmer, and R. J. Duff. 2000. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. Mol. Biol. Evol. 17:1885–1895.

O'Hara, R. J. 1992. Telling the tree: narrative representation and the study of evolutionary history. Biol. Philos. 7:135–160.

Sanderson, M. J., B. G. Baldwin, G. Bharathan, C. S. Campbell, D. Ferguson, J. M. Porter, C. Von Dohlen, M. F. Wojciechowski, and M. J. Donoghue. 1993. The growth of phylogenetic information and the need for a phylogenetic database. Syst. Biol. 42:562–568.