CEB Journal Club: Price et al. (2012)

James E. Allen

28th March 2012

Members of the Computational and Evolutionary Biology (CEB) group at the University of Manchester participate in a monthly journal club, where a paper of broad interest is discussed. Here, I briefly describe the paper and its context, and summarize our conclusions about the methodology and results presented. (I have attempted to represent the discussion and consensus of the group, but any inaccuracies are my own.)

[If referring to this document, please cite its location on the Monkeyshines website: http://www.monkeyshines.co.uk/blog/archives/1627]

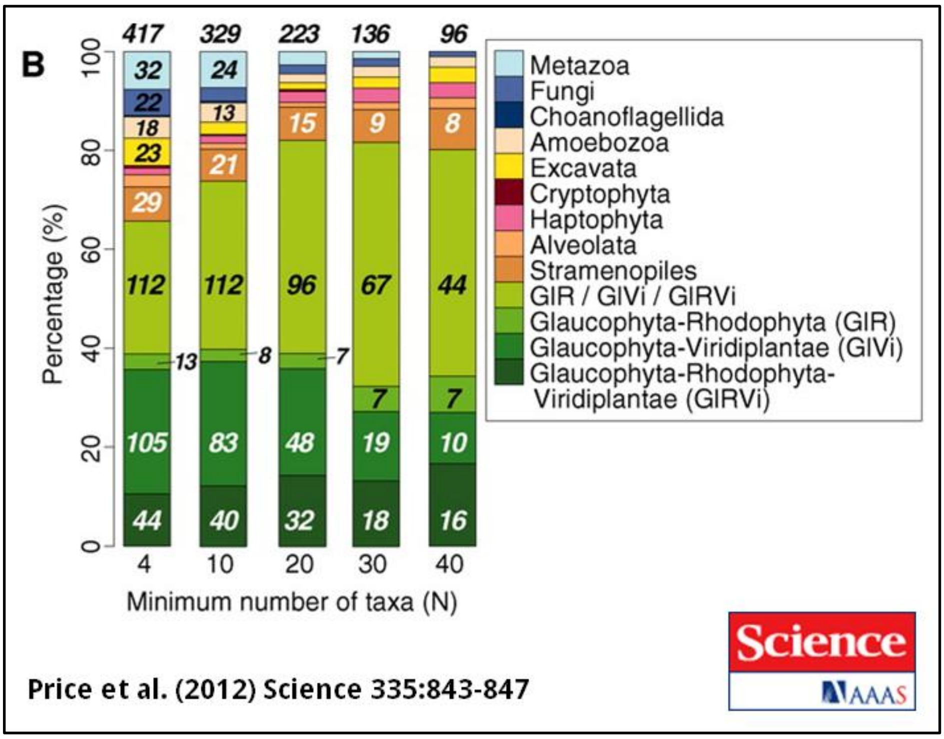
***Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants.** Dana C. Price *et al.* (2012) *Science* 335: 843-847. PubMed: 22344442 (Presented by James Allen, on Pi Day, 14th March 2012)

The paper in a sentence: Red algae, green algae and land plants, and glaucophytes (i.e. Plantae) are monophyletic; the photosynthetic ability of all plants derives from a single primary cyanobacterial endosymbiosis in their common ancestor.

Background: Photosynthesis is possible in plants due to the presence of a plastid, the remnant of an ancient endosymbiosis between a eukaryotic cell and a cyanobacteria. Until recently it was believed that a single, primary, endosymbiosis occurred in the ancestor of all plants and algae, and analysis of the plastid genome confirmed this. However, some phylogenomic analyses, enabled by increasing volumes of sequence data, provide either weak or no support for the monophyly of Plantae, suggesting the possibility of multiple endosymbioses. Resolving these issues is interesting because it sheds light (pun intended) on the nature of the first photosynthetic algae, and illuminates (OK, I'll stop now) the fascinating, billion-year-old, events which gave rise to, ultimately, the daffodils in my front garden.

The paper in detail: There are three divisions in the Plantae, two of which (red and green algae) have been well studied; until now, relatively little information has been available for glaucophytes, a species-poor group of algal protists that constitute the third division. Price *et al.* sequence the genome of a glaucophyte, *Cyanophora paradoxa* in an attempt to gain sufficient data to convincingly confirm or refute the monophyly of Plantae. To move onto the more interesting stuff, I assume that the sequencing was done sufficiently well to provide reliable data (the authors give enough detail in the supplementary material to justify this assumption).

From a set of almost 28,000 predicted proteins, almost 4,500 had prokaryotic or eukaryotic homologs and were of passable quality. The authors state that they "generated 4,445 maximum likelihood trees from the *C. paradoxa* proteins and found that >60% support a sister-group relationship between glaucophytes and red and/or green algae with a bootstrap value ≥90%". There are, I believe, a few problems with this sentence, if my interpretation of Figure 1B (reprinted below) is correct. In the first column of the figure, a total of 417 is given (subsequent columns are subsets of this first column, so we can ignore those for now); this is the number of maximum likelihood (ML) trees which contain 3 or more phyla, and in which the branch that places the



Price et al. (2012) Science 335:843-847

glaucophytes in a monophyletic group has a bootstrap value ≥90%. So, >60% of "4,445" trees do not support Plantae monophyly; >60% of 417 trees do. But, in fact, only 44 of these trees contain all three divisions of Plantae; 118 pair glaucophytes with either red or green algae and a further 112 trees show evidence of endosymbiotic gene transfer (EGT), and are assumed to support monophyly. The evidence of these last two sets is certainly consistent with monophyly, but is weaker than cases where all three divisions are present.

The approach taken here, to discard a large amount of data that does not meet a relatively arbitrary bootstrap criteria, seems wasteful. There are established methods for combining information in multiple gene trees to generate a species tree ('supertree' methods), and although these are not always straightforward to use, I would have expected at least some discussion on why they were not applied. Another option for phylogenomic analyses is the supermatrix approach, in which protein sequences are concatenated before tree inference. Supermatrices may not be effective if the proteins do not share approximately the same evolutionary history, i.e. if EGT or HGT has occurred; but since the authors are able to fairly confidently detect these events (e.g. Figure 1C and 1D in the paper), these proteins could have been excluded from the analysis. Even if the authors' approach is taken to be valid, >60% support for Plantae monophyly is not terribly convincing (incidentally, it is curious that the authors understate their case here, since the support value is actually 66% - why round it down?).

A final issue with the analysis of ML trees (before we move into more positive territory) is the lack of detail in the description of ML tree inference. The authors state that they are ‘using phylogenomics’ and reference a previous paper, in which a similar analysis is done; but the materials and methods section in that paper lacks detail, and some of it clearly does not apply. A concrete example of why this matters: RAxML can generate bootstrap replicates in different ways, which often does not affect any conclusions, but might be important here, where the analysis relies heavily on a particular cut-off value; it probably doesn’t make a difference, but the reader lacks the information needed to appropriately interpret the results.

I’ve gone into some detail about one aspect of the paper (the bit most pertinent to my area of research), and I am aware that I have been rather critical; often, in posts such as this, there is an understandable tendency to hem and haw, and obliquely imply that ‘perhaps the authors might have considered this or that’ and so on. But, I think my criticisms are fair, and are probably more interesting to read than the impending paragraph about the remainder of the paper, in which more convincing evidence of monophyly is presented…

The authors describe a number of proteins that are essential to eukaryotic photosynthesis, and demonstrate that these strongly suggest a common origin for red algae, green algae, and glaucophytes. The biological underpinning of these arguments, based on relatively few gene trees, is far more persuasive than the preceding phylogenomic analysis. There are also some interesting details on the gain and loss of fermentative enzymes, which would be clarified further with a greater number of species in the analysis.

Journal club conclusion: In places, the paper lacked sufficient, unequivocal, detail about the methods for us to wholly trust their conclusions. Some lines of evidence, however, were quite convincing, and we tend to believe that the Plantae are indeed monophyletic. Wider discussion of phylogenomics revealed a growing distrust of the results of such analyses; different researchers can arrive at contradictory answers to the same question, depending on how a dataset is selected and the exact nature of the analysis. Phylogenomics is necessarily complex, which makes it crucial for researchers to be meticulous when describing their data and methodology, so that we are able to decide to accept or reject their conclusions.