Deep Green Plant Phylogenetics: novel analytical methods for scaling data from genomics to morphology

PROJECT SUMMARY

The tree of life is inherently fractal. Look closely at one lineage of a phylogeny and it dissolves into many separate lineages, and so on down to a very fine scale. There is now a great body of phylogenetic research that has provided numerous tools applicable at particular, usually fairly constrained, scales. These tools have left many phylogenetic questions unanswered. We think these questions will remain unanswered until it is possible to do analyses across multiple scales.

We believe that the green plant lineage represents the most suitable system for such research. It is one of the oldest and most diverse branches of the tree of life, and it contains good examples of the known phylogenetic problems. Investigations on it may draw on a tradition of interdisciplinary collaborative research, facilitated by the Green Plant Phylogeny Research Coordination Group (GPPRCG or "Deep Green").

Many interesting questions remain to be tested in the green plants, once a better resolved phylogeny is available, such as: How many times was land colonized from the water by "green algae?" Where did the key adaptive features for life on land come from? How many times has multicellularity arisen in the green plants? Did multicellularity ever reverse? How many times did alternation of generations and diploid-dominant life-cycles arise? How has tempo and mode of macroevolution changed during diversification?

One could take two different approaches to broad phylogenetic studies such as this, either developing data sets with relatively few exemplars, but a very large number of comparable characters, or data sets with many exemplars but a smaller set of comparable characters. Both approaches have advantages, and both have their advocates. The two are not mutually exclusive: the compartmentalization approach taken here uniquely allows both approaches to be followed. A backbone phylogeny will be developed with a global data set and then local phylogenies with many more OTUs, but fewer and different characters, will be connected in.

Our overall objective for the work proposed here is resolve the primary pattern of evolutionary diversification among green plants and establish a model for doing so that will be applicable to other groups of organisms with long evolutionary histories. A solid backbone based on genomic and ultrastructural data for relatively few taxa will enable the integration of previous and ongoing studies of many more taxa into a comprehensive picture of green plant phylogeny.

To achieve this objective, we will:

* complete a matrix of whole genome sequences for chloroplasts and mitochondria and develop Bacterial Artificial Chromosome (BAC) nuclear genome libraries (where feasible given genome size) for ca. 50 representatives of the critical deep-branching lineages of green plants.

* produce a comprehensive set of comparable morphological and ultrastructural data for these same taxa;

* incorporate inferences from across the phylogenetic hierarchy in green plants using methods designed to permit scaling across studies.

We shall indicate how this work will link to other research being conducted on green plants at various scales, especially the concatenation of our datasets with theirs. We shall propose training, education, and outreach strategies by which the activities of our group, and the progress and results of our research, will be distributed to the scientific community and beyond.

1. INTRODUCTION

The tree of life is inherently fractal. Look closely at one lineage of a phylogeny and it dissolves into many separate lineages, and so on down to a very fine scale. The nature of both **OTU's** ("operational taxonomic units", the "twigs" of the tree in any particular analysis) and **characters** (markers that serve as evidence for the past existence of a lineage) change as one goes up and down this fractal scale. A robust reconstruction of the whole tree of life will require strategies that are powerful and flexible enough to encompass these phenomena. Although a great body of phylogenetic research has provided numerous tools applicable at particular (usually fairly constrained) scales, these tools have left many phylogenetic questions unanswered. We think they will remain unanswered until problems associated with the "scaling" have been addressed and applied to management and analysis of large datasets.

Our goal is to develop and test tools for phylogenetic reconstruction that address "scaling" and other large-dataset issues. To do this, we need a suitable system. "Suitable" implies that the system, a lineage of organisms, has sufficient diversity and a sufficiently long evolutionary history to provide a variety of different phylogenetic scales for examination. The system should be adequately studied to provide a reasonable phylogenetic framework, should be based on studies at scales for which the existing tools are relevant, and should identify discrete, unresolved domains for which hypotheses can be tested using new approaches. The system should interest a body of informed and networked investigators who are competent to tackle the various tasks associated with generating and analyzing large datasets for addressing important phylogenetic questions.

We argue here that the green plant lineage is the most suitable system at present, and the people who have gathered to study it within the framework of the Green Plant Phylogeny Research Coordination Group (GPPRCG, or "Deep Green") are best placed to develop and test general new tools needed to resolve the Tree of Life:

- This branch of the Tree is one of the most diverse in number of taxa (ca. 5 x 10^5 species), habitats, morphological types, reproductive strategies, and secondary chemistries;
- At a minimum age of ca. 10^9 years, it is one of the oldest lineages of "crown" eukaryotes;
- It contains good examples of the known phylogenetic problems, including deep and shallow branches, pulses of radiation/asymmetric extinction, heterogeneous evolutionary rates, and horizontal gene transfer;
- It has a better fossil record than most other branches of comparable depth and diversity;
- Its living representatives are of great importance to all aspects of human affairs;
- It has already been the focus of much coordinated phylogenetic research - the GPPRCG is an interactive, cooperative community that can productively address the several outstanding phylogenetic and methodological questions.

In the pages that follow, we describe the classes of phylogenetic problems that require attention. We identify several unresolved "deep" nodes of green plant phylogeny that represent selected examples of these problems, and detail the hypotheses to be tested in relation to them. We describe the procedures by which exemplars will be selected for analysis, and by which large datasets of morphological/ ultrastructural and molecular/genomic characters will be assembled, annotated, and archived. We set down what computational tools will be developed for analyzing these datasets, and how we will use them. We indicate how this work will link to other ongoing work on green plants at various scales, and will lead to concatenation of our datasets with theirs and the exploration of whether our scaling tools are adequate to generate robust phylogenetic reconstructions from these concatenated datasets. Finally, we propose training, education, and outreach strategies that will distribute the activities of our group and the progress and results of our research to the scientific community and the public.

Overall Objectives – To resolve the primary pattern of evolutionary diversification among green plants and establish a model for doing so that will be applicable to other groups of organisms with long evolutionary histories. A solid backbone based on genomic and ultrastructural data for relatively few taxa will enable the integration of previous and ongoing studies of many more taxa into a comprehensive picture of green plant phylogeny.

In the course of obtaining this objective, we will achieve the following:

Genomic characterization. We will complete a matrix of whole genome sequences for chloroplasts and mitochondria and develop Bacterial Artificial Chromosome (BAC) nuclear genome libraries (where feasible given genome size) for ca. 50 representatives of the critical deep-branching lineages of green plants.

Morphological characterization. We will produce a comprehensive set of morphological data for these same taxa, with emphasis on global cellular and ultrastructural features.

Integration of existing phylogenetic research. We will incorporate inferences from across the phylogenetic hierarchy in green plants using methods designed to permit scaling across studies.

2. PROBLEMS IN DEEP PHYLOGENETIC RECONSTRUCTION

Shallow versus deep phylogenetics. The challenges associated with reconstruction of "shallow" relationships are fundamentally different from those of "deep" ones [1]. In "shallow" reconstruction problems, branching events happened a relatively short time ago and the set of lineages resulting from these branching events is relatively complete (extinction has not had a major effect). In these situations, the relative lengths of internal and external branches are similar, giving less opportunity for long branch attraction. However, at this level an investigator often has to deal with the confounding effects of reticulation and lineage sorting. Characters at the morphological level may be quite subtle, and at the nucleotide level require very careful analysis to find rapidly evolving genes. (However, note that such genes are likely to be relatively neutral, thus less subject to adaptive constraints which can lead to non-independence).

In contrast, in "deep" reconstruction problems, the branching events happened a relatively long time ago and the set of lineages resulting from these branching events is relatively incomplete (extinction has had a major effect). In these situations, the relative lengths of internal and external branches are often quite different, thus there is a greater likelihood of long branch attraction. Conversely there are few problems with reticulation and lineage sorting, since most of the remaining branches are old and widely separated in time. Due to all the time available on many branches, a myriad of morphological characters should be available, yet they may have changed that homology assessments are difficult; the same is true at the nucleotide level, where multiple mutations in the same region may make alignment difficult. Thus very slowly evolving genes must be found, but such conservatism is caused by strong selective constraints that increase the danger of convergence leading to character dependence.

Structural vs. DNA sequence characters. How intrinsically useful are different categories of characters at these different scales? Clearly, structural and DNA sequence data have different and complementary strengths and weaknesses. Especially in "deeper" comparisons, structural characters such as morphological or genomic markers are more information-rich, allowing a temporal axis of comparison not possible with DNA sequence data. Structural characters often change in an episodic pattern, which is necessary for evidence of deep, short branches to remain detectable (clock-like markers are the worst kind of data for those sorts of branches). The number of possible character states is usually much higher in morphological character systems (and in genomic rearrangements) than in DNA sequence data and this makes long-branch attraction less problematic [2]. On the other hand, objectively defining character states in morphological comparisons can be difficult, particularly in "shallow" reconstructions, whereas the states are usually clear-cut in DNA sequence data. DNA sequence markers are also much more numerous, thus increasing the chance that sufficient markers can be found for all branches of a tree.

Dealing with heterogeneous data types. Deep phylogenetic reconstructions are inherently difficult, so all characters should be developed and used if they meet the criteria of good potential markers [1]. However, it remains controversial how data from different sources are to be evaluated and integrated with each other [3]. Some have argued that data sets derived from fundamentally different sources should be analyzed separately, and only common results taken as well-supported (i.e., consensus tree approaches), or at least that only data sets that appear to be similar in the trees they favor should be combined [4]. Others have argued that all putative homologies should be combined into one matrix. Theoretical arguments now favor the latter approach (i.e., "total evidence;" [5-8, 2, 9]). If characters have been independently judged to be good candidates for phylogenetic markers, then they are equivalent and should be analyzed together.

There is one major exception to our preference for a "total evidence" position: data should not be combined if there is evidence that some of it had a different branching history than the rest. However, there are several sources of homoplasy other than different branching history, including evolutionary convergence. If several data partitions show different highly discordant trees due to convergence, the only way to see the true tree topology is to combine them. The only weapon a systematist has against convergence is the likelihood that truly independent characters will be subject to different confusing factors and thus the true history may emerge when these independent characters are combined. Probably all character systems are influenced by constraints that tend to bias phylogeny reconstruction one way or another, yet a combination of very different character sets can allow the "noise" to cancel out revealing the historical signal.

Therefore, observing a particular data partition exhibiting serious conflict with another is not sufficient reason to reject combining them. There must also be additional evidence, outside of the phylogentic analysis, of reticulation or lineage sorting. The best examples of such discordance are in "shallow" analyses, where organellar genomes may have different phylogenies than those of associated nuclear genomes and morphologies [10-12]. Barring that sort of clearly explainable discordance, all appropriate data should be used, especially in "deep" analyses because as argued above, reticulation and lineage sorting are much less likely to be problems in "deep" analyses, while convergence is likely to be a greater problem.

Global versus local approaches. How will we ultimately connect "deep" and "shallow" analyses, each with their own distinctively useful data and problems? Some hold out hope for eventual global analyses, once enough universally comparable data are amassed and computer programs are efficient enough to deal with all extant species simultaneously. Others would go to the opposite extreme, and use a "supertree" approach, where the "shallow" analyses are simply grafted onto the tips of the "deep" analyses. An intermediate approach, "compartmentalization" [13, 2], uses the "shallow" topologies (that are based on analyses of the characters useful locally) to constrain "deep" analyses (that are based on analyses of characters useful globally).

The task at hand. We need to address how characters can be selected, interpreted, and most effectively analyzed at various scales. The primary advantages of using the green plant lineage for this work are that a wealth of "shallow" analyses are published and ongoing, and many of the methods for collecting the data for "deep" analyses (particularly, genome-level molecular data) are being developed. This enables us to evaluate unresolved "deep" nodes by developing a large dataset that encompasses characters derived from both genomic and morphological analyses. Given the requested funds, we will link our framework to existing "shallow" analyses and use these linkages to test which of our scaling approaches works best.

3. "DEEP" PROBLEMS IN GREEN PLANT PHYLOGENY

Current topology of the green plant tree (Fig. 1) arises from considerable morphological and ultrastructural data that have accumulated over the last three decades (e.g. [14-26]), and from the molecular tools that have been applied at many levels (e.g., [27-41]). Many recent advances reflect innovations in data gathering and analysis that have resulted from increased coordination of effort among different laboratories (see Management section). Despite these advances, several nodes in the "deep" phylogeny” remain unresolved (Fig. 1). We propose to resolve these problematic nodes.

Ambiguity in the origin of green plants. It is generally accepted that there are two principal clades of green plants: the streptophytes, consisting of the land plants and the green algae most closely related to them, and the chlorophytes, containing most of the remaining green algae. At the base of these two clades is a "residuum" consisting of unicellular "prasinophytes". These algae are regarded as ancestral within the green plants, on the basis of morphology and ultrastructure (e.g. [42]), nuclear gene sequences (e.g. [43, 32, 44]) and organelle genome features [45-48]. However, the relationships among the prasinophytes have been difficult to recover. One species only, *Mesostigma viride*, has been placed with some confidence at the base of the streptophytes [49, 50]. Most of the remaining prasinophytes have been placed at the base of the chlorophytes, but without clear affinities to any other chlorophyte lineage [44].

Which are the most ancestral prasinophytes? The "phycomate" prasinophytes (those with large, thick-walled floating stages, or "phycomata") are candidates. They have ultrastructural features common to both the chlorophyte and streptophyte lines [42], they are the only green algae with mixotrophy (nutrition by both photosynthesis and phagotrophy [51]), a presumptive precondition for endosymbiosis of chloroplasts, and they have a fossil record extending to the latest Precambrian and perhaps much earlier [52]. One extant genus, *Tasmanites*, has a fossil record dating back ≥600 million years, making it the oldest of all green plants [53, 54]. To test this result, we will incorporate a phycomate prasinophyte, *Pterosperma*, into our primary analysis, and attempt concatenation of our results with ongoing research on living (collaboration with Fawley) and fossil (collaboration with Knoll) prasinophytes.

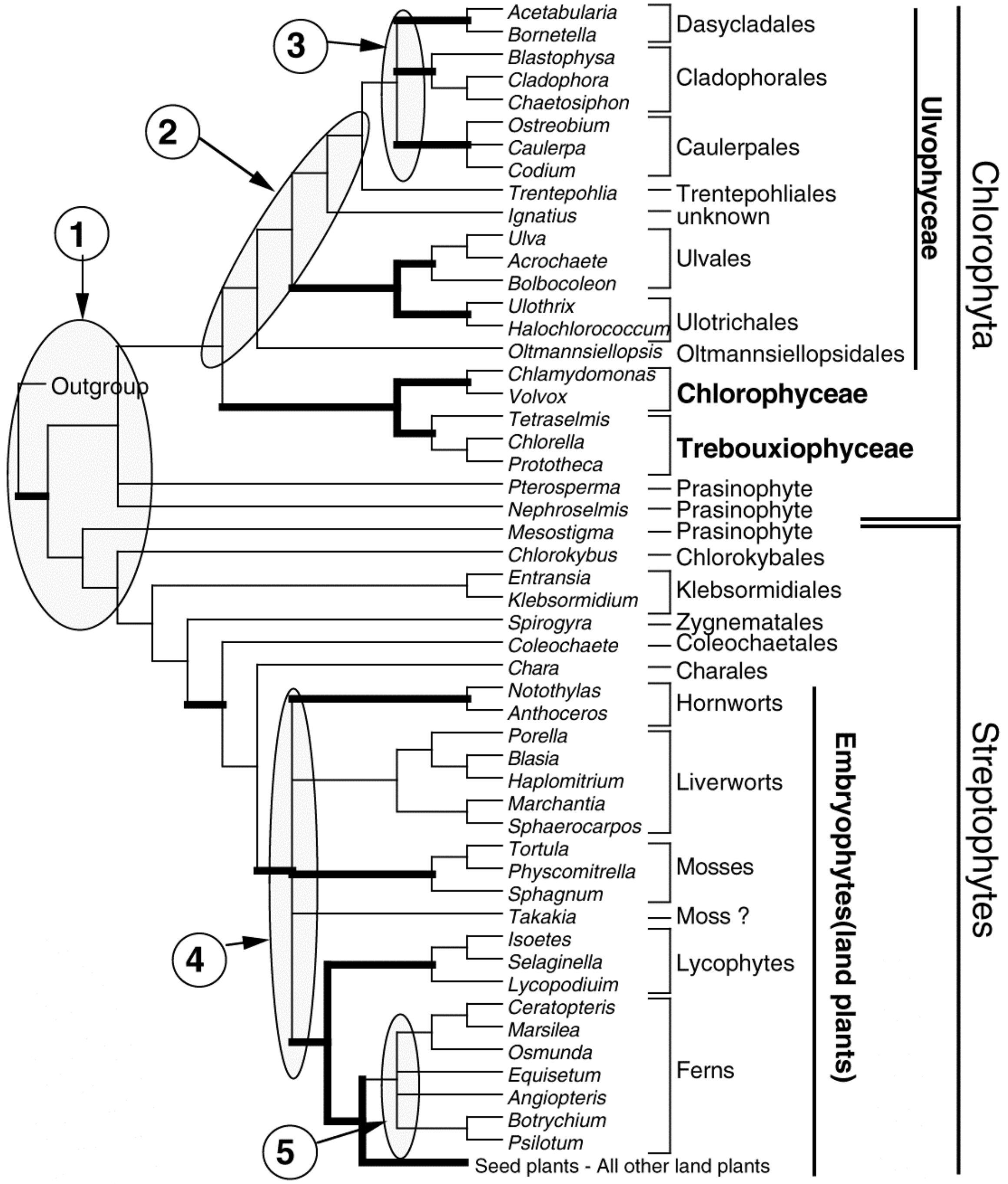


Figure 1. Outline phylogeny of the green plants, indicating the currently understood phylogenetic position for 51 candidate exemplar organisms (represented by generic names; see Table 1) and the higher-level taxa to which they are thought to belong. All branches are subject to further testing, but the best-supported branches with current data are indicated with a thick line, branches with some support are indicated with a thin line, and areas of the tree that remain controversial are indicated by labelled ovals. See text for further explanation: **1.** Base of green plants: outgroup relationships and basal branching among prasinophytes. **2.** Relationships among major lineages of Chlorophyta (some groupings are more firmly established, e.g. Ulvales/Ulotrichales, Chlorophyceae/Trebouxiophyceae). **3.** Relationships among the siphonous algae and their placement in the Chlorophyta. **4.** Base of the land plants. **5.** Base of the ferns (moniliforms).

To identify the ancestral prasinophyte, we need to define outgroups for green plants. Many recent gene-sequence trees indicate that the green plants are most closely related to the red algae and the glaucocystophytes, the other two groups of algae that appear to have gained their chloroplasts through a primary endosymbiosis with a cyanobacterium [55]. However, red algae and glaucocystophytes differ significantly from each other and from green plants in key morphological, reproductive and ultrastructural features, so much so that homologies are difficult to establish (e.g. [42]), and the molecular results have been questioned (e.g. [56]). The most ancient green algae have organellar genomes that may be more ancestral (more like eubacteria) than those in the red algae examined to date [57, 45-48]. We will test the idea that red and glaucocystophyte algae represent the proximal outgroups for green plants by incorporating published data from selected representatives of these outgroups into our analyses, and by interacting with colleagues who will be conducting research on potential outgroups from among both algae and protozoa (collaboration with Lang).

The Ulvophyceae. The Ulvophyceae is one of the 3 classes currently recognized in the chlorophyte lineage, the others being Chlorophyceae and Trebouxiophyceae (formerly Pleurastrophyceae; [58]). Most members are marine, and the majority of green "seaweeds", including well-known species of *Ulva, Acetabularia* and *Caulerpa*, are placed in this class. Conversely, the Chlorophyceae and Trebouxiophyceae, and the streptophytes, consist almost entirely of non-marine organisms. In classifications based on morphology and ultrastructure [59, 60, 42], the Ulvophyceae have been separated from other chlorophytes mostly on the basis of characters associated with mitosis, cytokinesis, and the flagellar apparati of zoospores and gametes. In molecular analyses, however, the relationships among these three classes are less clear [61]. Moreover, the "siphonous" orders of Ulvophyceae (Cladophorales, Dasycladales, Caulerpales) are difficult to resolve vis-à-vis each other and with other Ulvophyceae (orders Ulotrichales and Ulvales); phylogenetic trees based on single gene sequences reveal long branch lengths between "siphonous" sequences and those of other chlorophytes [59, 62] (O'Kelly unpublished, Friedl unpublished).

Are Ulvophyceae as conceived by Floyd and O'Kelly [64] monophyletic, and how are ulvophyte clade(s) related to other chlorophytes? We think it possible that the "siphonous" "ulvophyceae" represent a clade separate from, and basal to, the remaining chlorophytes (Chlorophyceae, Trebouxiophyceae and non-siphonous "Ulvophyceae"). We will test this idea by incorporating both siphonous and non-siphonous Ulvophyceae, together with representative Chlorophyceae and Trebouxiophyceae, into our large-scale analysis, and attempt concatenation of our results with those from other ongoing research on ulvophytes (O'Kelly ulvophyte grant). Some algae in our large-scale investigations, particularly *Acrochaete, Blastophysa, Bolbocoleon, Halochlorococcum, Ignatius, Ostreobium,* and *Trentepohlia*, we selected because very recent research (O'Kelly and Friedl, unpublished) suggests they may break up some of the long ulvophyte branches.

The "bryophytes" - early embryophyte radiation. There is no consensus as to the primary branching patterns at the base of the land plants. Molecular and morphological evidence [63-67, 25] suggests that either hornworts or liverworts are the oldest living lineage of land plants. In the “hornworts-basal” hypothesis, a moss + liverwort clade is typically supported. In the “liverworts-basal” hypothesis, all three main bryophyte lineages are paraphyletic, with either hornworts or mosses sister to the vascular plants. Fossil evidence supports the liverworts-basal hypothesis, as the first unambigous hornwort fossils date to the Cretaceous and liverwort fossils to the Devonian. However, ornamentation of Paleozoic spores from the Silurian (>410 Mya) are comparable to that of *Anthoceros*, raising the possibility that hornworts were the first bryophyte lineage to appear. Monophyly of the mosses is widely accepted, but the interrelationships among the four major moss lineages are unresolved. The hornworts are undoubtedly monophyletic but within-group phylogeny has not been fully explored. Liverworts are highly diverse and interrelationships are ambiguous to the extent that monophyly of the group is questionable. DNA sequence data have been equivocal, supporting a number of conflicting branching orders, in part due to poor taxon sampling or limited sequence lengths [68-70, 66, 35, 24, 39]. Our large-scale analysis is likely to resolve the bryophyte tangle.

The basal vascular plant radiation. Within the tracheophytes, the lycophytes are sister to all other tracheophytes (e.g., [71, 72]), a result supported by analyses of both morphological and DNA sequence data. A comprehensive analysis of morphological and molecular characters in basal tracheophytes [36] produced the topology: (lycophytes (((*Psilotum* + ophioglossoid ferns) + (*Equisetum* + marattioid ferns + leptosporangiate ferns)) + seed plants)). This novel topology unites horsetails together with all ferns as a monophyletic group that is sister to seed plants and refutes the earlier view that horsetails are transitional evolutionary grades between bryophytes and seed plants. This conclusion was supported consistently by data from morphology and spermatogenesis [73, 24, 25]. The sister relationship of *Psilotum* with ophioglossoid ferns, which was suggested previously (e.g., [74, 70, 75, 76]), is now strongly supported [36].

4. NEW DATA ACQUISITION: GENOMICS, MORPHOLOGICAL AND SEQUENCE DATA

Our goal is to build a robust phylogenetic reconstruction across multiple scales. To do this, we will generate a comprehensive dataset for at least 51 exemplar taxa. Taxa were chosen for their postulated phylogenetic position relative to the unresolved nodes in green plant evolution (Fig. 1, Table 1). This dataset will incorporate characters from morphology, ultrastructure, and organellar genome and nuclear gene sequences. We will generate, annotate and archive these data (note taxa already done: Table 1). These data will be used to reconstruct the "deep" phylogeny of green plants, and will serve as the backbone for concatenating "deep" analyses with many ongoing shallower analyses in green plants.

Table 1: Top strategy for obtaining genomes and genome sizes (1 x = 100 Mb = 0.1 pg).

Species (reference for nuclear genome size)	Nuclear genome size (1C)	Target genomic data				Top strategy to obtain organellar genomes (see Table footnote)		
		Chloroplast genome		Mitochondrial genome				
		Done	Us	Done	Us	Isolate	BAC	oBAC
Nephroselmis olivacea	1x^	OG		OG				
Pterosperma sp.	1x^	OG		OG				
Chlorella sp. [77-80]	0.4x		✓		✓		✓	
Prototheca wickerhamii	0.4x+			OG			✔	
Tetraselmis striata	1x$	OG		OG				
Volvox carteri [81, 82]	1x	B		B			B	
Chlamydomonas reinhardtii [83, 84]	1x	J		J				
Oltmannsiellopsis viridis	1x*		✔		✔		✔	
Ignatius tetrasporus	1x*		✔		✔		✔	
Halochlorococcum moorei	1x*		✓		✓		✓	
Ulothrix sp.	1x*		✓		✓		✓	
Acrochaete endozoica	1x*		✓		✓		✓	
Bolbocoleon proliferum	1x*		✓		✓		✓	
Ulva lactuca [85, 86]	1x		✓		✓		✓	
Trentepohlia sp.	1x*		✓		✓		✓	
Blastophysa rhizopus	5x@		✓		✓			✓
Chaetosiphon moniliformis	5x@		✓		✓			✓
Cladophora sp. [87-89]	5x		✓		✓			✓
Caulerpa taxifolia [90]	2x!	B		B			B	
Codium decorticatum [91]	6x		✓		✓		✓	
Ostreobium queketii	6x%		✓		✓		✓	
Acetabularia acetabulum [92]	9x	B?		B?			B?	✓
Bornetella sphaerica	9x&		✓		✓			✓
Mesostigma viride [93]	1x	B, OG		B, OG			B	
Coleochaete orbicularis [93]	1x	B, D		B, D			B	
Chara aspera [94]	72X	B		B			B	
Klebsormidium flaccidum	1x*	OG		OG			✓	
Entransia sp.	5x@		✓		✓			✓
Chlorokybus atmophyticus	1x*		✓		✓		✓	
Spirogyra sp. [95]	5x@		✓		✓			✓
Anthoceros sp. [93]	4x	B		B			B	
Notothylas orbicularis [96]	2x		✓		✓			✓
Sphaerocarpos sp.	2-20x#						✓	

Marchantia polymorpha [93]	3x	B		B			B	
Blasia pusilla [96]	5x		✓		✓			✓
Bazzania trilobata [96]	10x		✓		✓			✓
Haplomitrium sp.	2-20x#		✓		✓		✓	
Sphagnum palustra [97, 96, 98]	5x		✓		✓			✓
Takakia ceratophylla [96]	4x		✓		✓			✓
Physcomitrella patens [99]	6x	OG		OG				✓
Andreaea sp. [97, 100]	2x		✓		✓			✓
Tortula ruralis	4x		✓		✓			✓
Selaginella kraussiana [101]	0.5x	B?		B?			B?	
Isoetes englemanii [93]	5x	B?		B?			B?	
Lycopodium lucidulum [101]	20x	B		B			B	
Equisetum hyemale	120x		✓		✓			✓
Psilotum nudum [101]	20x	B		B			B	
Botrychium sp. [101, 102]	3x		✓		✓			✓
Angiopteris evecta [93]	4x	B		B			B	
Osmunda cinnamomea [103]	9x		✓		✓			✓
Ceratopteris richardii [104]	40x	B		B			B	
Marsilea quadrifolia [101]	4x	B		B			B	

Isolate - isolate organellar genomes using traditional gradients (1) or FACs (2);
BAC - make a Bacterial Artificial Chromosome library biased for nuclear genome;
oBAC - make an organellar Bacterial Artificial Chromosome library biased for organellar genomes.

Nuclear genome estimates based on:

* - *Ulva lactuca*
@ - *Cladophora albida*
+ - *Chlorella sp.*
^ - *Mesostigma viride*
& - *Acetabularia acetabulum*
% - *Codium decorticatum*
$ - *Chlamydomonas reinhardtii*
# - full range of bryophytes [97, 100]
! –4 *Caulerpa* taxa, range 0.1-0.15 pg/1C genome [90]

Genome sources:

✓ - this grant
OG - Organelle Genome Megasequencing Prog.
J – Joint Genome Institute
B - BAC grant (Mandoli, PI) confirmed taxa
B? - BAC grant (Mandoli, PI) in contract neogiations

Criteria for selection of taxa. Our primary criterion for selection of our 51 taxa was their hypothesized phylogenetic position in relation to the nodes we want to resolve (Fig. 1, Table 1). We selected among the many possible exemplars on the basis of four subsidiary criteria. **1**) To complement sampling of other studies that are developing genomic resources for comparative study in green plants, including the NSF-funded Collaborative Grant on Plant and Algal BACs (D. Mandoli, Project Director), the Organelle Genome Megasequencing Program (M. W. Gray and B. F. Lang, Program Directors), and Jansen's seed plant chloroplast seqeuncing project (see letters B. F. Lang & R. Jansen, collaborators). **2**) We added taxa that will facilitate concatenation of published and ongoing studies to the backbone phylogeny that we will develop here. **3**) The taxon must be easy to obtain through collection or cultivation. **4**) Taxa that are important models for research in various fields. When alternatives exist within the constraints of these criteria, the organism with the smaller nuclear genome size was chosen to maintain cost-efficiency of BAC production. If in the course of our work, an species proves to be intractable or we find another one that seems even more suitable or has a smaller genome size than one selected initially, we will make the appropriate replacements.

Morphological, ultrastructural, and other non-molecular data. A major component of this project is accumulation and interpretation of morphological data. Accurate detailing of anatomical, developmental and ultrastructural features is critical to all future morphological inquiry. Though they made crucial contributions to our understanding of green plant phylogeny, until recently studies on morphological components tended not to be conducted systematically. Differences in methodological approach, available technologies and investigator biases made them subject to discordances. Our studies are designed to provide reliable contemporary morphological data that will correct errors, clarify ambiguity and augment information available in the literature. To use "discrete" rather than "composite" OTUs, we will detail the morphological and ultrastructural features of all exemplars that we examine at the genomic level (Fig. 1, Table 1). In this way we will build a comprehensive dataset based on temporally and methodologically consistent approaches and maximally discrete OTUs. These data will allow us to critically evaluate morphological datasets compiled from the literature, for fossil as well as living specimens, and will contribute to analyses across "deep" and "shallow" scales by maximizing our ability to interpret homologies, paralogies and convergences in the evolution of morphological characters.

We will concentrate on 1) anatomical features that can be derived from light microscope observation of living, preserved and dried material and 2) ultrastructural, developmental and physiological data that require tissue preparation and observation in the TEM, SEM, fluorescence or light microscope. We will begin with recently composed coherent character matrices, including the 132-character bryophyte set [19, 24], the 75-character matrix for spermatogenesis in land plants [25], and the 77-character matrix for pteridophytes ([105, 76]; see http://www.science.siu.edu/landplants/Morphological/MorphData.html). These characters will serve as a baseline for data collection and will be substantially modified as characters are evaluated and character states defined. A major focus will be to construct like datasets for the chlorophyte algae, which have seldom been compiled in forms comparable to those cited above [66]. Acquisition of crucial ultrastructural and morphological characters will identify potential homologies and will significantly enhance resolution of morphological data.

In addition to accumulating general information on plant morphology, we will conduct intensive studies of key structural features and processes that are common to all or most taxa. This will provide data at all fractal scales and enable global comparisons. The available data are restricted to cellular features and so we will conduct thorough studies of cell division, especially mitosis, analyze cell wall constituents and examine motile cell structure and differentiation using standard TEM, fluorescent labels and immunolabeling protocols for TEM, fluorescent and light microscopy (e.g. [106, 25]).

Genomic data. Whole organellar genomes provide two distinct sorts of data for phylogenetic inference. Gene and intron losses, inversions, and other structural changes in the genome occur infrequently and can provide powerful phylogenetic markers (e.g., [107, 108, 72]; but see [109] for example of homoplasy). Complete chloroplast and mitochondrial genome sequences will also provide two important sequence data sets. In addition to structural genomic data, we will assemble chloroplast and mitochondrial datasets from all coding regions of sufficient size and conservation to permit confident sequence alignment. The tremendous amount of organellar sequence data should permit unambiguous reconstruction of organellar phylogenies for all taxa sampled. We will also sequence a few nuclear genes that are either single-copy or from small multi-gene families which are appropriate for analysis at this scale. BAC libraries will facilitate probing for (on filter arrays) and amplification of the desired sequences (from individual BAC clones).

Four approaches will be used to obtain organellar genomes (Table 2). The order in which we will execute these options reflects the relative costs per genome and the probability of working most easily.

Table 2: Comparison of four methods to obtain the organellar genome data.

Traditional isolation of organelles	FACs to purify organellar genomes	BAC library for 100Mb nuclear genome, 5X coverage	oBAC library biased for organellar genomes, ~19x coverage
$100-5,000	$850	$1,923	$192

We will determine the size of those nuclear genomes that have not been directly measured using flow cytometry (see Arumuganathan cv). For genomes ~100MB, we will make a standard BAC library (17 taxa in Table 1) because this is relatively inexpensive and will provide us with all three genomes. Average insert size per clone in the Wing lab is 130-150 Mb and we will aim for ≥5X coverage which is considered a minimum BAC library standard by NSF. Quality control of all libraries will be done by the Wing lab (CUGI standards). For genomes >100Mb we have three options to get the organellar genomes. Our first option will be to create an “organellar bacterial artificial chromosome” or oBAC library. During normal BAC library construction, tissue from which the cell wall has been digested is embedded in agar. Proteins, carbohydrates and organellar genomes are removed *in situ* to preserve intact chromosomes. Normally, a Triton-X step is included to reduce the organellar genome representation from 10-15% to 2-3% in the final BAC library. We will omit the Triton-X step, essentially capitalizing on old technology for a new purpose, and make a very small library, 384 clones, that will nevertheless represent each organellar genome ~19-times. The libraries will be arrayed and probed with standard genes to identify those clones containing organellar genomes (mito: atpA, cob, atp9, cox1; chloro: ndhA, rbcL, psbA). Many clones will contain the entire organellar genome. Our oBAC and BAC procedure may reveal nuclear regions that contain organellar DNA such as has been found in rice (Wing, unpub.). Not only is this method cost effective (Table2), but it is automated (http://www.genome.clemson.edu/), produces arrayed filters and –80C glycerol stocks of all clones, is the best chance of preserving fragile organelles from some of the more ancient taxa (Delwiche, pers. comm.), and will foster data and bioinformatics exchange into the genomic community via the CUGI/AZ website. Should this protocol fail for a particular organism, we will use a Fluorescence Activated Cell Sorter (FACS) to separate mitochondria and chloroplasts from fractionated cells. Standard DNA extractions will be made from the sorted organelles. If both oBAC and FACS fail for the larger genomes, we will fall back on centrifugation protocols for organellar separation and extraction, the classical approach [110, 111]. With four alternatives, all taxa should be feasible, but we will draw from our pool of "alternate" organisms if any taxa prove intractable.

The purified genomes or surrogate templates will be sheared randomly into fragments of ~3 kb using a Hydroshear device, end-repaired, and gel purified. Routine quality control measures ensure that shearing produces fragments of narrow size distribution (important in the later sequence assembly phase), with 1 s.d. ≤8% of the intended fragment length. These fragments will be blunt-ligated into pUC18, transformed into *E. coli* DH5α, and plated onto large format bacterial plates under conditions that allow for blue-white color selection. Colonies will be grown overnight, then processed robotically through creation of glycerol stocks, extracted and amplified using rolling circle amplification, separation for forward and reverse primer sequencing, and setting up of the sequencing reactions. Sequence determination will be on 96-capillary automated sequencers. For each genome, 96 clones will be sequenced to determine purity (based on BLAST searches of sequences). Sequencing will continue until approximately 8-fold redundancy, when gaps in the gene-rich genomes should be minimal. Gap filling and sequence completion will be done by returning to archived plasmid preps, or if necessary through amplification of genomic DNA. Gap filling will be done in collaboration between JGI and Utah State University. The goal will be to achieve a total of approximately 9 Mb of final sequence data.

At both CUGI/AZ and JGI all cloning and analysis steps are tracked using bar-code readers. The data are automatically entered into a workflow database for statistical analysis of each phase of the operation. Sequencing machines automatically output their data into a UNIX-based folder system, where they are assembled into contigs. The JGI software is unique in that it uses paired-plasmid ends to guide contig assembly. Gene annotation uses both standard and custom software which has been successful for many whole genomes sequenced at JGI. All sequence data will be deposited in GenBank.

Primary sequence characters. Sequences of chloroplast genomes are complete for 24 organisms, including four green algae, *Marchantia*, *Psilotum*, and numerous seed plants. From analyses of these genomes, we infer that the best source of characters will be protein-encoding genes and genes for the 16S and 23S ribosomal RNAs. Gene content ranges from 69 protein-coding genes in *Pinus* to 84 in *Marchantia* (78 in *Chlorella*; 76 in *Nicotiana*) so gene losses are likely to be important phylogenetic markers [108]. A strategy using nucleotide sequences of 17 protein-coding chloroplast genes exhibiting low synonymous substitution rates and site-to-site rate variation has been applied successfully to studies of basal angiosperms and land plants [112]. Additional results (R. Olmstead, unpubl.) suggest that this strategy can be used successfully at much deeper phylogenetic levels in green plants. Stoebe et al. [113] analyzed 46 protein-coding genes totaling >11,500 aligned amino acids positions in a study of 9 taxa representing all chloroplast genomes then available and including non-green plant taxa. Restricting our study to green plants will enable us to use 60 genes and ~50,000 nucleotides of DNA sequence. Characters will be defined at both the nucleotide and amino acid levels and analyses will be carried out where most appropriate given alignments and levels of nucleotide sequence divergence.

Complete mitochondrial genomes have been sequenced for fewer green plants than chloroplast genomes. However, two green algae, *Marchantia*, and at least one seed plant have been sequenced. Green plant mitochondrial genomes also contain small and large subunit rRNA genes, but contain far fewer protein coding sequences than do chloroplast genomes. We will conduct combined multi-gene analyses for green plants as we described for chloroplast genomes. Mitochondrial DNA substitution rates are slower than those of either chloroplast or nuclear genomes [114]. Various mtDNA genes have been used recently for deep phylogenetic studies in land plants [115, 116, 28, 117, 37]. We will sequence single-copy nuclear genes as well as some from small multi-gene families. Again, nuclear BAC libraries and filter arrays made from them will greatly ease the acquisition of sequence for phylogenetic analysis and provide genomic tools for other researchers. Working from BACs instead of whole genomic DNA enables PCR-based approaches to recover all copies of the genes, without interference from more readily amplified copies, a problem when using PCR on nuclear multi-gene families. We will focus on protein-coding genes that have been identified as useful for deep phylogeny in plants. The RNA polymerase II consists of several subunits each encoded by separate nuclear genes. With rare exceptions, the two largest subunits (RPB1 and RPB2) are single copy genes in all groups in which they have been studied. RNA pol II genes have been used for deep phylogenetic studies of crown eukaryotes [118], red algae [119, 120], fungi [121], and land plants (B. Hall, pers. comm.) and should help resolve our fuzzy nodes. Phytochrome genes have a good signal for seed plant and basal angiosperm phylogeny [122] where a series of duplications have yielded a clearly defined set of phytochrome genes. However, in non-seed plants [123] evidence suggests that there is a single gene with some lineages having duplications (e.g., *Selaginella*, *Psilotum*). Some of these duplications are likely to mark clades once sampling is expanded.

5. DATABASING AND ANALYSIS. Many powerful resources for genomic sequence data (e.g., Genbank and EMBL, SWISS-PROT, PIR) are archival because they do not provide integrated access to the tools needed for phylogenetic and comparative analysis. The proposed bioinformatics will provide support for an expertly curated set of data (organellar genome sequences, genomic structure, ultra-structural and morphological data), and integrate these with analytical tools needed for phylogenetic and comparative analysis. Thus we propose to create not only a database, but also a data laboratory.

Character database enhancements. Several databases, notably GOBASE ([124]; http://megasun.BCH.UMontreal.CA/gobase/gobase.html) and MitBASE (http://www3.ebi.ac.uk/Research/Mitbase/mitbase.pl), focus on organelle genomic data. These resources use a relatively simple set of tables to display published sequence, gene location, protein sequences, and genetic maps. A simple query interface allows data retrieval based on gene and protein names, exon and intron definitions, and taxonomy. GOBASE defines a standard nomenclature for mitochondrial genes, but none exists for chloroplast genes and gene products. We will extend the above database structure to include phylogenetically important structural changes such as insertion/deletion regions, inversions, and duplications. The most straightforward way to implement this is to compare each chloroplast genome to a virtual standard genome; pairwise comparisons then can be simply generated by comparing the two genomes in question to the standard genome. We will make several enhancements to GOBASE to improve its search and referencing abilities.

Annotation and alignment. The sequencing group will annotate single genomes for database deposition using the beta test versions of “Mitotater” and “Plastotater”**.** PiPmaker and MultiPipMaker will be used to identify a wide variety of structural changes that have occurred during plastid genome evolution and to generate multiple sequence alignments for downstream phylogenetic and molecular evolutionary analysis. PipMaker (http://bio.cse.psu.edu/pipmaker/) is a flexible program for visualization and evolutionary analysis of whole genome sequences, and is ideally suited to our bioinformatics needs. The PipMaker approach is based on percent identity plots (PIPs) which are linear representations of high scoring regions found with genomic-scale dot matrix analyses. This approach efficiently aligns very large sequences (≤Mbp). Alignments are fast and use as series of BLAST programs [125, 126]). PIP output is compacted to allow rapid identification of genes and other homologous sequences [127-129], repeat elements and their classification [130] and structural features, and to reveal evolutionarily conserved promoter and regulatory elements [127] regardless of their linear order in the genome [131, 132]. The website provides various tools that aid genome annotation and visual presentation of the results. MultiPiPmaker, a recent expansion of PipMaker, allows simultaneous alignment of ≤100 genomes using a new multiple alignment algorithm (Miller, unpub.). In addition to generating compact summary maps, aligned sequences can be exported for downstream phylogenetic and molecular evolutionary analysis.

6. PHYLOGENETIC ANALYSIS

Principles of OTU and character selection. Due to the integrative nature of the proposed analyses, in which data from many sources will be considered, the concepts of “OTU” and character will vary within and among datasets. Data at this level are always compiled from study of different organisms considered to represent the same OTU. Thus OTUs are always composites in practice; their composition varying depending on the scale of analysis. Likewise, what counts as a useful character changes depending on the scale of analysis. The columns in a data matrix are already refined hypotheses of phylogenetic homology. There is also a clear reciprocal relationship between OTUs and characters. An OTU can best be defined as a set of individual samples that are homogeneous for characters currently known, while a character can be defined as a potential marker for shared history of some subset of the known OTUs. This means that OTUs and characters emerge during a process of "reciprocal illumination." To a large extent their definitions are interlinked, so how do we proceed empirically in a way that avoids circularity? We will take great care to examine both concepts of character and OTU in the proposed research. We will use the relatively advanced state of knowledge of characters and phylogenetic structure in the green plants as a model system for testing alternative approaches to analysis in a systematic manner. The overarching goal is to develop ways to scale OTU composition and character definition up and down the many fractally-nested levels making up the tree of life.

Character analysis. Phylogenetic analysis can be broken down into two discrete phases: character analysis and cladistic analysis. In the former phase, a data matrix is assembled as discussed above. Potential characters are evaluated by rules of character analysis, an evaluation of evidence for: (1) homology and heritability of a character across the taxa being studied, (2) independent evolution of different characters, and (3) presence in each character of a system of at least two discrete states.

These criteria will be applied here to data matrices assembled at several scales of analysis. The deepest scale will be a matrix of the ca. 50 exemplar green plants plus outgroups (see section 4), with much of the data newly generated from this proposal. These OTUs will be thoroughly studied for the morphological characters covered above, and have completely sequenced mitochondrial and chloroplast genomes. We will also have nuclear BAC libraries constructed for most exemplars, which will facilitate discovering gene translocations between the organellar genome and the nuclear genome, and sequencing of new candidate nuclear genes. Characters will be evaluated in the following categories:

(1) *genomic characters*. Structural genomic differences resulting from inversions, translocations, gene losses, duplications, and insertion/deletion of introns will be identified within and between the three genomes and likely homologies established (e.g, examining the ends of breakpoints to see whether a single event is likely to have occurred).

(2) *morphological characters*. All features that can be compared across this deep level of analysis will be evaluated for independence and discrete states. The literature will be used, but wherever possible original material will be reexamined.

(3) *DNA sequence data*. To compare DNA sequence characters with genomic and morphological characters, we will also align all genes available in the three genomes, We will do this two ways (in order to compare results): a liberal alignment using as much sequence as possible, and a conservative alignment using only regions that are unambiguously alignable. Both amino acid and nucleotide alignments will be analyzed where appropriate for protein coding genes.

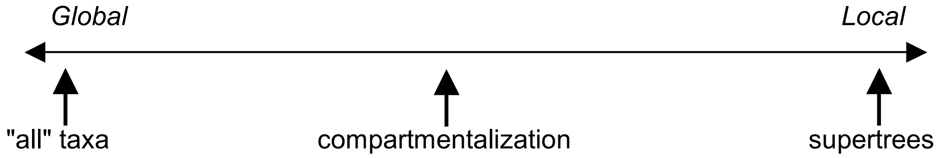
Matrices will be developed for local clades using data appropriate at that level. These data will come almost entirely from other research groups and collaborators as discussed in the management plan section.

Cladistic analysis. The second phase of phylogenetic analysis involves turning data matrices into a recontructions of a phylogenetic tree. We will explore the full spectrum of approaches to building phylogenetic trees from data matrices and how to concatenate the results from the different scales of phylogenetic analysis to be undertaken here. We will use only character-based methods of phylogenetic analysis, and mainly work within a maximum parsimony framework (given the very heterogeneous set of characters). However, we will compare and contrast equal and differentially-weighted parsimony and maximum likelihood methods as applied to DNA sequence data. The first task will be to analyse the data matrix of the ca. 50 exemplar taxa, and the mixture of genomic, morphological, and DNA sequence characters discussed above, to produce a "backbone" phylogeny of basal green plants. The next task will be to use this sparsely-sampled, but extremely character-rich, global phylogeny to connect up all the many local phylogenetic data sets available from other research groups. These local data sets sample many more taxa (thousands taken all together), but with considerably less character data available.

For this second task we will assemble all published phylogenetic trees on the relevant chlorophyte and streptophyte lineages (e.g., references above), and will closely coordinate our efforts with ongoing phylogenetic projects of direct relevance to ours (see “Management Plan”). We will insure that all relevant phylogenetic studies are entered into TreeBASE (www.treebase.org), thereby providing ready access to project members (and the entire scientific community) to phylogenetic knowledge on green plant lineages.

Concatenation analyses. This assembly of individual phylogenetic trees and data sets will be critical to the construction of large-scale concatenated trees. In collaboration with M. Sanderson (UC Davis) we will use green plant phylogenies to explore a variety of algorithms for producing supermatrices and supertrees, such as Matrix Representation Parsimony (e.g., [133-135]) and methods that can take branch lengths into account [136]. This will allow direct comparisons to be made with other approaches, such as simultaneous analysis of concatenated data matrices and compartmentalization methods (references above; also [137]. There is a full spectrum of approaches for concatenating analyses at different scales:

How to concatenate different analyses to build the tree of life? A spectrum of approaches:



At the left end of this spectrum, the approach is to include all possible OTUs and potential characters in one matrix. Generally this is not actually done, because the sheer amount of data (millions of possible OTUs) makes thorough phylogenetic analysis computationally impossible. The most common approach is to select a few representatives of a large, clearly monophyletic group (the exemplar method). Care is sometimes taken to select representatives that are "basal" OTUs within the group to be represented; however, this still does not avoid two important problems: (i) within-group variation is not fully represented in the analysis, and (ii) an increase both in terminal branch lengths and in asymmetry between lengths of different branches is introduced. These problems can lead to erroneous branch attractions in global analyses.

At the right end of the spectrum, local analyses are simply grafted together at the place where shared taxa occur, without reference back to the original data. There are many ways to do this in detail (as reviewed by Sanderson), but the important thing is that the analyses on real character data are only done locally, and the concatenation is based on the combination of local topologies rather than a combination of local data sets into a global data set.

We will explore both of these approaches even though both seem too extreme, one too global, the other too local. Thus we will also explore a promising synthetic approach called *compartmentalization* (by analogy to a water-tight compartment on a ship -- homoplasy is not allowed in or out) that represents diverse yet clearly monophyletic clades by their inferred ancestral states in larger-scale cladistic analyses. A well-supported local topology is sought first, then an inferred "archetype" or hypothetical ancestor (HTU) for the group is inserted into a more inclusive analysis. In more detail, the procedure we will use is to: (1) perform global analyses, determine the best supported clades (these become the compartments); (2) perform local analyses within compartments, including more taxa and characters (more characters can be used within compartments due to improved homology assessments among closely related organims); (3) return to a global analyses, in one of two ways, either (a) with compartments represented by single HTUs (the archetypes), or (b) with compartments constrained to the topology found in local analyses (for smaller data sets, this approach is better because it allows character optimizations within each compartment).

The compartmentalization approach differs from the exemplar approach in that the representative character-states coded for the archetype are based on all the taxa in the compartment, thus the reconstructed HTU is likely to be quite different from any real OTU. As an estimate of the states of the most recent common ancestor of all the local OTUs, the HTU is likely to have a much shorter terminal branch with respect to the global analysis, which in turn can have the beneficial global effect of reducing long-branch attraction. In addition to these advantages of compartmentalization at the global level, the local analyses will be better because one can: (1) include all local OTUs for which data are available; (2) incorporate more (and better justified) characters, by adding in those characters for which homology could not be determined (aligned) globally; (3) avoid spurious homoplasy that can change the local topology due to long-branch attractions with distant outgroups. The effects of compartmentalization are thus to cut large data sets down to manageable size, suppress the impact of spurious homoplasy, and allow the use of more information in analyses. This approach is self-reinforcing; as better understanding of phylogeny is gained, the support for compartments will be improved, leading in turn to refined understanding of appropriate characters and OTUs.

Phylogenetic database enhancements. As our data sets develop, we plan to assist larger efforts to develop a new generation of phylogenetic data-bases, including TreeBase and a pending ITR proposal for a national resource in phyloinformatics (see management plan). The next generation of data resources needs to be much more flexible than existing data bases (e.g., GenBank, which is essentially "flat" with respect to phylogeny), and sensitive to scale and the fractal nature of phylogenies (with their many hierachically nested scales).

The exploration of the basic nature of phylogenetic data described above will be applied to data-base research through modeling studies. We will address fundamental questions about the nature of data before, during, and after phylogenetic analysis. Biologists in this project will work with collaborating computer scientists to model: (1) How are elements of the data matrix (OTUs, characters, and states) defined and recognized in any particular study? (2) How can heterogeneous data types (e.g., DNA sequences, genomic rearrangements, morphology) be compared/combined? (3) How can data sets and analyses at very different scales be concatenated (e.g. supertree, compartmentalization, or global approaches as discussed above)? (4) How can data sets at these different concatenated scales, where OTUs are nested inside larger ones and character definitions (e.g., alignments) change as you move up and down the scale, be presented to the user community?

7. EVOLUTIONARY IMPLICATIONS

The efforts described above are expected to yield a well-supported phylogeny of green plants. This will be used, alone and in combination with other phylogenetic information, to make a variety of evolutionary inferences. Specifically we will focus on the evolution of a set of morphological characters, on patterns of molecular evolution, and on rates of diversification. Meaningful answers to these questions require large taxon sampling. Tree size is a critical factor with respect to the ability to distinguish between competing evolutionary models, such as symmetrical versus asymmetrical probabilities of character change (e.g., [138, 139]) and correlated character evolution (e.g., [140]). As a result of the concatenation analyses discussed in the previous section, we anticipate the assembly of phylogenetic hypotheses large and well-supported enough to ensure coverage of the groups that are necessary to address specific evolutionary hypothesis, and to allow statistically meaningful comparative analyses.

Our studies of character evolution will focus initially on issues of broad evolutionary significance from the standpoint of the entire Tree of Life. Specifically, basal green plants will allow analyses of (1) the origins of multicellularity, (2) transitions to life on land, and (3) the evolution of an exceptionally wide variety of life cycles. (1) Multicellularity appears to have evolved repeatedly within green plants. In the streptophyte clade, multicellularity preceded the evolution of embryophyte land plants. In chlorophytes there have been multiple paths to multicellularity, including colonial forms in the volvocine line, multiple origins of filamentous and more complex parenchymatous plant bodies in Chlorophyceae and Ulvophyceae, and siphonous forms culminating in coenocytic thalli within Ulvophyceae. (2) Transitions to life on land also occurred repeatedly early in green plant evolution. The embrophytes represent one such occurrence, but multiple independent events are apparent within Trebouxiophyceae (see Chapman et al., 1998), often entailing symbiotic relationships with fungi and animals. (3) Regarding life cycle evolution, great interest has focused on the transitions that preceded embryophyte evolution, and the subsequent origination of alternation of gametophyte and sporophyte generations. Of equal evolutionary interest are many apparent shifts within chlorophytes, including the evolution of animal-like life cycles (diploid dominance, products of meiosis functioning as gametes) in *Codium* and other Ulvophyceae.

In each of these cases, and others related to ultrastructural features and genome evolution, we will infer ancestral states and evolutionary sequences using parsimony and maximum likelihood approaches (see [141, 142, 138, 139, 143-145] To test for the correlated evolution of characters we will employ a battery of comparative techniques, again using parsimony, likelihood, and Bayesian methods (e.g., [146-150]), and performing relevant sensitivity analyses (e.g., [151, 152]).

Finally, we will focus on issues concerning the tempo of green plant evolution. The absolute time of key divergences will be inferred using molecular data, calibrated by the placement of fossils (see letter from A. Knoll). Methods that variously relax the molecular clock assumption will be used, including penalized likelihood and Bayesian methods, in which fossil evidence places minimum and maximum bounds on the estimates [153-160]. We will also use a variety of methods to assess if and when shifts in diversification rate occurred in basal green phylogeny (reviewed in [161-163]), and explore correlations between diversification shifts, character changes, and rates of molecular evolution (e.g., [164]).

8. BROADER IMPACTS

Support of this proposal will have an impact on the infrastructure of education within the PIs institutions, communities and beyond. We will develop a Web site dedicated to extending awareness of green plant biology, diversity and genomics to the global scientific community and the general public. On this site we will provide a synopsis of current activities and make available a comprehensive morphological database, including images.

We remain committed to providing enriching and supportive educational experiences for undergraduate and graduate students. We stress teamwork and support a mentoring network for all members of our laboratories. We will seek additional funding for student participation through the Research Experiences for Undergraduates program. To broaden the knowledge base and expertise, students and postdocs will be encouraged to rotate among research groups for extended periods of time. We are active participants in university and community programs to provide research and professional experiences for elementary, high school and undergraduate students. These activities are showcased in our biographical sketches and will be expanded upon in the proposed program. We will participate in existing summer programs for secondary school teachers and students that focus on the use of computer technology to address a range of biological problems, and that aim to increase awareness and understanding of biological issues.

The PIs have previously been involved in a number of initiatives that focus on facilitating and attracting underrepresented minorities to science disciplines, e.g., through public outreach activities. We have and will continue to actively increase the number of underrepresented minorities (African-Americans, Native Americans, Hispanics and the disabled) entering environmental biology disciplines. These include pending NSF proposals to the Undergraduate Mentoring in Environmental Biology program (enables minorities to attend Botanical Soc. America meetings) and the Persons with Disabilities Program (supports youths entering STEM-related disciplines). Moreover, the international component of our proposed program will provide resources and opportunities for students to explore collaborations with students and scientists from a diversity of cultures and backgrounds.

RESULTS FROM PRIOR SUPPORT.

Jeffrey Boore. NSF 9807100. "A Phylogeny of Major Metazoan Radiations". 1998-2001, $200,000. We determined complete mitochondrial genome sequences for 25 phylogenetically diverse invertebrates. Several contentious higher level relationships were robustly reconstructed using mitochondrial gene arrangement characters: Pogonophora are a family within Annelida; Platyhelminthes and lophophorates fall within Eutrochozoa; Sipuncula are more closely related to Annelida than to Mollusca; insects evolved within Crustacea, not from myriapods. We have also developed models for gene order rearrangements and explored many aspects of mitochondrial DNA structure as well as their impact on phylogeny reconstruction. 16 publications to date.

Michael J. Donoghue. NSF 9806937, "Duplicate genes and plant phylogeny: phytochromes and the rooting of the seed plants, angiosperms, and eudicots", 1998-2001, $180,200. A series of phytochrome gene duplications were used to root the seed plants, the angiosperms, and the eudicots. **CHR (3):** Three undergraduate women. Four publications to date.

Dina Mandoli. NSF 9630618, *"Cell biology & genetics of Acetabularia phenotypes that are arrested in development"*, 1996-2000, $200,000. We completed all 3 Aims: 1) we finished inbreeding near isogenic lines proving that we can perform genetic manipulations; 2) we demonstrated that high-throughput transformation and selection work well; and 3) we studied development, compensation and genetics of developmentally arrested phenotypes. This research makes development of insertional mutagenesis of *A. acetabulum* feasible. 16 publications to date. **CHR (29)**: 1 postdoctoral fellow (NSF), 4 graduate and 24 undergraduate students. Includes 17 women, 10 minorities, 2 learning disabled, 8 Gates or Hughes Fellows. 24 out of 29 plan or have careers in science.

Brent Mishler. USDA 94-37105-0713 (DOE/NSF/USDA Collaborative Research in Plant Biology), *"The Origins and Phylogeny of Green Plants: A Research Coordination Group";* 1994-2000, $285,459; NSF 0090227, *"Beyond 'Deep Green': Towards an Integration of Plant Phylogenetics and Plant Genomics*"; 2001-2006, $496,434. The **Green Plant Phylogeny Research Coordination Group** (**GPPRCG**; "**Deep Green**"), has been working since 1994 to facilitate the reconstruction of an ever more resolved phylogeny. See full progress report at: http://ucjeps.berkeley.edu/bryolab/greenplantpage.html. Preliminary results of the GPPRCG were presented in a series of eight symposia at the XVI International

Botanical Congress in 1999 and published in a series of papers in specialized professional journals, as well as numerous reviews. Progress of a new effort of this group, supported by an RCN grant "**Deep Gene**", can be found at: http://ucjeps.herb.berkeley.edu/bryolab/deepgene/index.html. Many professional workshops, research visits, and student training activities have already been supported.

Charles J. O'Kelly. DEB-0075601, "*Towards a Monograph of the Ulvellaceae (Ulvophyceae, Chlorophyta) and related green algae*", 2000-2004, $320,000. We are finding: (1) numerous new species (at least 15) within this assemblage; (2) significant lack of support for generic concepts based on morphology, some genera are polyphyletic at the class level (Friedl and O'Kelly 2002) while others (especially those in the Ulvellaceae sensu stricto) cannot be sustained; (3) assortment of these algae among five putative clades, including the Cladophorales, the Ulotrichales, the Ulvaceae and Kornmanniaceae of the Ulvales, and a clade, previously unrecognized at the molecular level, basal to other Ulvales and possibly identical with the "Ctenocladaceae" of some morphological classifications. One publication to date. **CHR (3)**: one postdoc and 2 research technicians.

Richard Olmstead. DEB-9727025, "*Chloroplast DNA Phylogeny of Basal Angiosperms*", 1998-2001, $205,000; DEB- 0090313; "*Chloroplast DNA Phylogeny of Seed Plants and Basal Angiosperms*", 2001-2004, $240,000. We identified 17 chloroplast genes (Graham and Olmstead 2000a) that permit resolution of ancient land plant radiations (e.g., low synonymous substitution rates and low site-to-site rate heterogeneity) to analyze ~15kb of data. Our evidence suggests that full resolution of the basal angiosperm radiation is possible with high levels of confidence. My current NSF grant expands this work to include more extensive sampling in seed plants and other major land plant lineages in collaboration with Sean Graham (U. Alberta). *Chara* and *Coleochaete* are included as outgroups. Four publications to date. **CHR (10)**: one postdoc, 2 research technicians, 3 grad students (partial support), and 6 undergrads on REU suppl., including 2 minorities.

Karen Renzaglia: DEB-9527735. °*Spermatogenesis in "pteridophytes": ultrastructure, differentiation and phylogeny.*° 1996-2001, $140,000. Our major research findings fall into two categories: (1) comparative information on cellular development and structure in land plants and (2) contributions to clarifying evolutionary trends and resolving phylogenetic relationships among basal embryophytes. We have provided detailed descriptions of sperm cell architecture and cellular development in pteridophytes, bryophytes, green algae and seed plants. Our work reveals that structural and developmental complexity in plant sperm cells are unsurpassed in any other group of organisms. We have generated new data, assembled published data and analyzed one of the most comprehensive data bases of both morphological and molecular data associated with the phylogeny of land plants. 27 publications to date, 6 with undergraduate co-authors. Nine undergraduates, one doctoral student and three master's students have worked on plant spermatogenesis since 1995.

Alan Smith: DEB-9616260 "Collaborative research: Phylogeny, character evolution, and diversification of extant ferns", 1997-2002, $25,619 (with K. M. Pryer and P. G. Wolf). DEB-9807053 "Morphological and molecular systematics of the Polypodiaceae and Grammitidaceae", 1998-2002 $55,939, (with T. A. Ranker and C. H. Haufler). The first of these awards supported phylogenetic studies on the ca. 10 basalmost families of extant ferns. By virtue of the results obtained, as well as some previous work, we expanded our study to include Psilotaceae and Equisetaceae, two groups traditionally thought to be fern allies, but now believed to be nested in the fern clade. Altogether, ca. 65 examplars (genera) from the basal clade were sequenced for four genes plus morphology. A similar approach has been applied to the study to the Polypodiaceae and Grammitidaceae, generally acknowledged to be the most recently derived of the higher leptosporangiate ferns.

Paul G. Wolf. DEB-9707087 "Collaborative Research: Phylogeny, character evolution, and diversification of extant ferns". 1997-2000, $94,990. We used data from 4 genes and morphology from over 60 taxa to resolve phylogeny of vascular plants. Our data indicate that horsetails and ferns together are the sister to seed plants. Five publications to date.

MANAGEMENT PLAN

Coordination of the project. Six main institutions and two subcontractor institutions are involved in the proposed research. In addition, 19 collaborators are expected to play a significant role in the project. The roles of each PI and collaborator are spelled out below. This team represents an amazing range of expertise, and the research is multifaceted, thus the project will require careful coordination.

Charles O'Kelly will serve as Project Coordinator for the length of the grant. He will supervise our overall progress and interface with related research groups (see below). A Steering Committee (SC) for the project will be established, consisting of all the PIs. The SC will hold a conference call each month to review progress and activities. An email list will serve for routine communication across labs.

There will be continuous interaction, data-sharing, and cross-training activities among the eight institutions and beyond. The entire group, (SC, collaborators, and students) as well as relevant invitees, will meet at least once a year in association with national meetings (usually of the Botanical Society of America). These meetings will include progress reports as well as discussions and demonstrations of new techniques and approaches. Meeting proceedings and new data availability will be shared among labs, and broadly with the general botanical public, by posting to a Web page. Inclusion in the group's activities will be open to all who are interested in its activities, as with the GPPRCG collaboration (that had well over 200 participants). To encourage broader participation, letters of invitation to meetings will be sent to key labs and broadly advertised across the community, and information on the meeting will be posted on the group's Web site and the Web sites of other relevant groups and societies.

The proposed research in relation to the GPPRCG. The Green Plant Phylogeny Research Coordination Group (GPPRCG or "Deep Green"), through a series of meetings, workshops, and collaborative analyses, was organized in 1994 to facilitate the production of a detailed phylogeny for this major branch of the tree of life. Considerable progress in understanding the phylogeny of green plants has been made, based on classical morphological characters, newly described ultrastructural features, and nucleotide sequence data from the nuclear, chloroplast, and mitochondrial genomes. Addressing a phylogenetic study of this enormous scale has necessitated improvements in data handling and analysis that have broad applicability to phylogenetic studies of other organisms. The success of this effort generated exciting new opportunities for applied and basic research and training. A full account of progress of the GPPRCG can be found at: http://ucjeps.berkeley.edu/bryolab/greenplantpage.html. The community of researchers in this area has been brought together, and a high level of communication and coordination achieved. In fact, the status of phylogenetic research on the green plants now serves as an example to all research groups interested in the other major branches of the tree of life.

Even though the original grant to fund the GPPRCG has expired, the GPPRCG Executive Committee has continued to function as an overall coordinating mechanism for several successor grants. It is composed of six regular members (three-year terms) plus one student member (one-year terms). Principal Investigators of closely related grants are also appointed by vote of the Executive Committee to serve as ex officio members while their grant is active, thus ensuring smooth cooperation across all grants. If funded, this grant would be represented by PI O'Kelly as an ex officio member (note that a number of the other PIs and collaborators on this proposal also serve on the Committee). This assures sharing of information and resources and facilitates design of co-sponsored meetings and educational outreach activities. Every effort will continue to be made to facilitate efficient and open sharing across the community and participation by all interested parties.

Current members of the GPPRCG Executive Committee: (shown with the August they rotate off)

Elected members:

Dr. Charles F. Delwiche (2002)
Dr. Pam Soltis (2002)
Dr. Richard M. McCourt (2003)
Dr. Kathleen Pryer (2003)
Dr. Louise A. Lewis (2004)
Dr. Yin-Long Qiu (2004)

Ex officio members:

Dr. Brent D. Mishler, **Chair** (rotated off as elected member 2001; PI, Deep Gene)
Dr. Pam Soltis (current elected member; PI, Biocomplexity Grant on genome evolution)
Dr. Douglas Soltis (PI, Deep Time)
Dr. Claude dePamphilis (PI, Floral Genome Project)
Dr. Robert Jansen (PI, Comparative Chloroplast Genome Project)
Dr. Dina Mandoli (PI, The Green Plant BAC Library Project)
Dr. Mark A. Buchheim (rotated off as elected member 2000; PI, Chlorophyte algae project)

Graduate Student representative:

Michael Zanis (2002)

Deep Green Research Coordination Networks. Two related Deep Green NSF RCN grants have recently begun operation to continue and extend the original Deep Green coordination -- one called "**Deep Gene**" Mishler, PI -- http://ucjeps.herb.berkeley.edu/bryolab/deepgene/index.html) to coordinate genomics and phylogenetics, the other ("**Deep Time**" D. Soltis, PI -- http://www.flmnh.ufl.edu/deeptime/) to coordinate paleontology and phylogenetics. The proposed research group will interface with both RCNs. Joint meetings of these networks will maintain communication and lead to joint sponsorships and colloquia (e.g., a joint workshop on phylogenetics for molecular biologists and paleobotanists). Through coordination among these RCNs and other groups, the network of interacting scientists will expand to include geologists, paleobotanists, morphologists, phylogeneticists, and genomic botanists.

The proposed research in relation to other research groups. The scale of the proposed research plus its relationship with other current and planned research projects dictate effective coordination and collaboration. For taxon sampling and research progress we are coordinating other groups beyond the GPPRCG including those of Mitsuyasu Hasebe (Japan), Michael Gray (Organelle Genome Mega-sequencing Project, Canada), Sean Graham (Canada). Broad collaboration among these groups minimizes redundancy in sequencing while increasing the efficiency of dissemination and analysis of results. Sequenced genomes will be of more utility than the scope of the project and it will be essential to provide detailed information to the research community, beyond deposition of data in the public domain.

Active phylogenetic projects to be coordinated with this project
In addition to a large body of published literature, there is a large number of active phylogenetic projects ongoing in the more "shallow" branches of green plants. The concatenation analyses proposed here will link together the topologies being produced by these "local" phylogenetic studies. These studies are thus complementary to the research proposed here, and will be coordinated by means of the GPPRCG and representative collaborators from these projects included in this proposal.

Investigators	group of green plants	type of data gathered	comments
Soltis et al.	embryophytes	DNA sequence data + fossils/morphology	This is a Tree of Life proposal being submitted separately by a GPPRCGd group ("Deep Time") with different interests
Qiu	embryophytes	DNA sequence data	an 8-gene data set
Jansen	seed plants	Chloroplast genome sequencing	This group is also working in the lab of Jeff Boore in the JGI, thus data gathering will be well-coordinated
Olmstead & Graham	seed plants	DNA sequence data	17 protein-coding chloroplast genes
Pryer, Wolf, Smith, et al.	ferns	DNA sequence data + morphology	This group is working on more derived clades of the ferns than in the present proposal
Shaw & Goffinet	mosses	DNA sequence data	
Delwiche & McCourt	charophyte algae	DNA sequence data + monography	A PEET grant
McCourt	zygnematalean algae	DNA sequence data	
Buchheim, Fawley, and Zechman	chlorophyte algae	DNA sequence data	This group is working on more derived clades of the Chlorophytes than in the present proposal

O'Kelly (& Wysor)	ulvophyte algae	DNA sequence data + monography	The focus is on unrecognized diversity of ulvophyte microalgae

Information & material sharing. Part of the success of Deep Green was the clear and repeated commitment of its organizers to individual ownership of data prior to publication and proper attribution of contributions by collaborators. We will continue this commitment. Therefore, on our Web site we will indicate the availability of data rather than distribute any unpublished data of individual investigators. We will indicate what resources and data are available and from whom. Our previous experience suggests that this will prevent duplication of research effort and suggest possible collaborations by allowing everyone to see who is doing what. Contribution of data to collaborative analyses will not required by participants in the group, although we anticipate that those who participate will be interested in exploring such collaborations.

The Web site will also link to related sites, including "Deep Gene" RCN, "Deep Green" and Plant and Algal BAC (planned) Web sites, as well as Web sites developed by the PIs to disseminate information on particular green plants (e.g., Land Plants Online http://www.science.siu.edu/landplants/index.html). Sharing of information reduces undue overlap of data, and provides up-to-date information on genomics and organismal resources (e.g., culture collections, data archiving, extracted DNAs, etc). Educational tools such as teaching modules for K-12 are featured on the "Deep Gene" Website. These will be further developed as results are accumulated from the Tree of Life Initiative.

Morphological Data Archive. One major problem in morphological systematics is the scattering and loss of physical materials (such as permanent slides, mounted blocks, photographs, etc.) and data as researchers retire. Some of the most important materials in green algal phylogeny have already been lost this way. Thus we will develop an archive for a wide variety of data and materials, hosted at the University and Jepson Herbaria, UC Berkeley. This will include culture collections, and morphological and ultrastructural data from deceased and retired scientists. In addition, a major effort will be made to integrate unpublished archived data made available by our collaborators into our studies (e.g., see attached letters from Brown, Duckett and Ligrone).

Training. The GPPRCG has always placed a heavy emphasis on student involvement and training. All of our workshops have included graduate students who are active in the field. The present proposal will continue that tradition, but will expand training activities from workshops and symposia into the laboratory. Where possible, students will visit among the laboratories of one of the collaborators. Additional training activities will be facilitated by the workshops sponsored jointly with the RCNs. These include the general annual workshops as well as workshops focused on specific topics and on cross-training between disciplines. Much of this cross- training will be funded through related RCN projects ("Deep Gene" & "Deep Time"), which award summer lab internships for undergraduates and laboratory exchange experiences for graduate students. Teacher workshops through "Deep Gene" are designed to disseminate information on plant genomics and to assist teachers in developing the best practices to teach this information.

Increasing diversity. The proposed RCN will welcome participation by a diverse array of scientists and will encourage participation by underrepresented groups and those individuals in diverse types of institutions. The best way to increase the participation of under-represented groups in science is through public outreach and opportunities/information for students. The proposed K-12 teacher workshops, workshops at professional meetings, as well as learning modules on the web site, will effect knowledge transfer to younger students as well as undergraduate and graduate students. These activities will increase the visibility of exciting science to all potential future scientists, including those in under-represented groups. Notices of the web site and workshops will be sent to biology and science departments at colleges and universities across the country as well as to associations such as the National Association of Science Teachers and the National Biology Teachers Association. We will actively seek out minority-serving institutions, In each competition for student awards, a portion will be reserved for deserving women and ethnic minority students. We will also encourage the participation of individuals from primarily undergraduate institutions (PUIs) by earmarking some awards for these students and faculty.

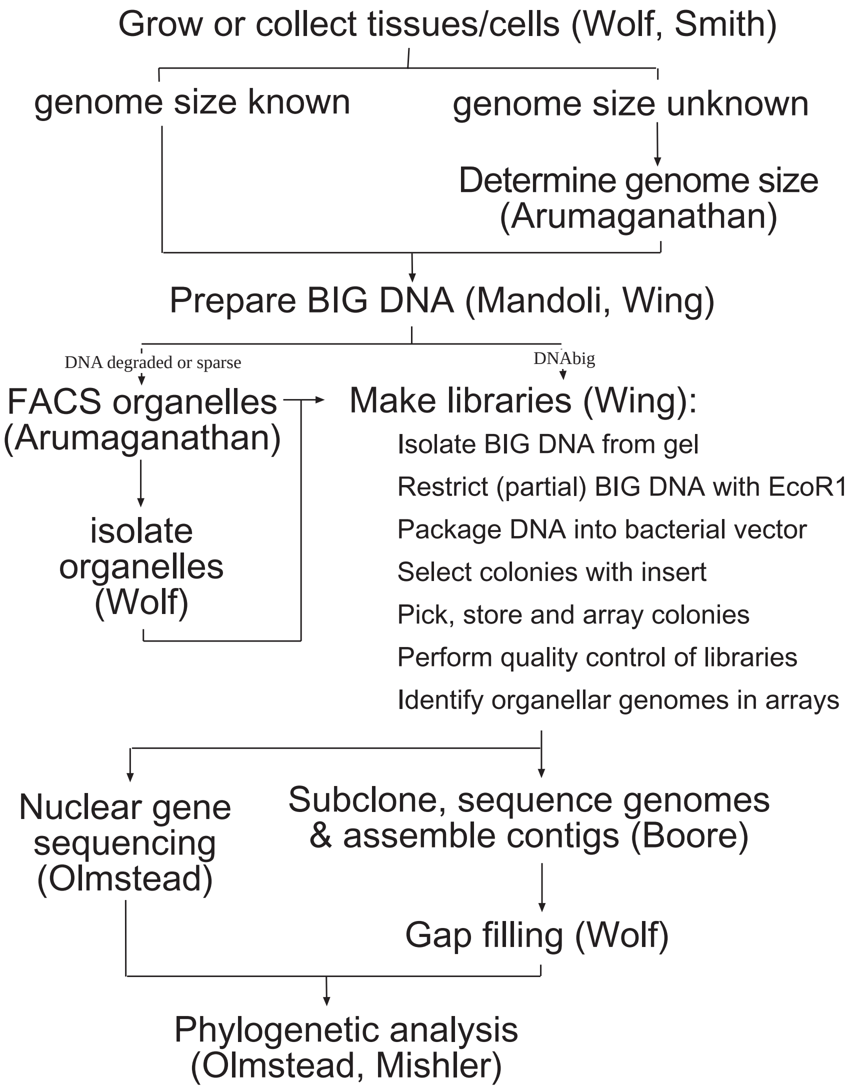
Assessment of research coordination activities. At the conclusion of each workshop, symposium, meeting, or other group event, a questionnaire will be distributed to all participants to gauge their satisfaction with the operation and productivity of the session. The SC will consider the suggestions in the survey and make appropriate changes in the operation of future meetings. Our past experience with Deep Green strongly suggests that this will be a positive and productive experience for all participants.

ROLES OF SENIOR PERSONNEL:

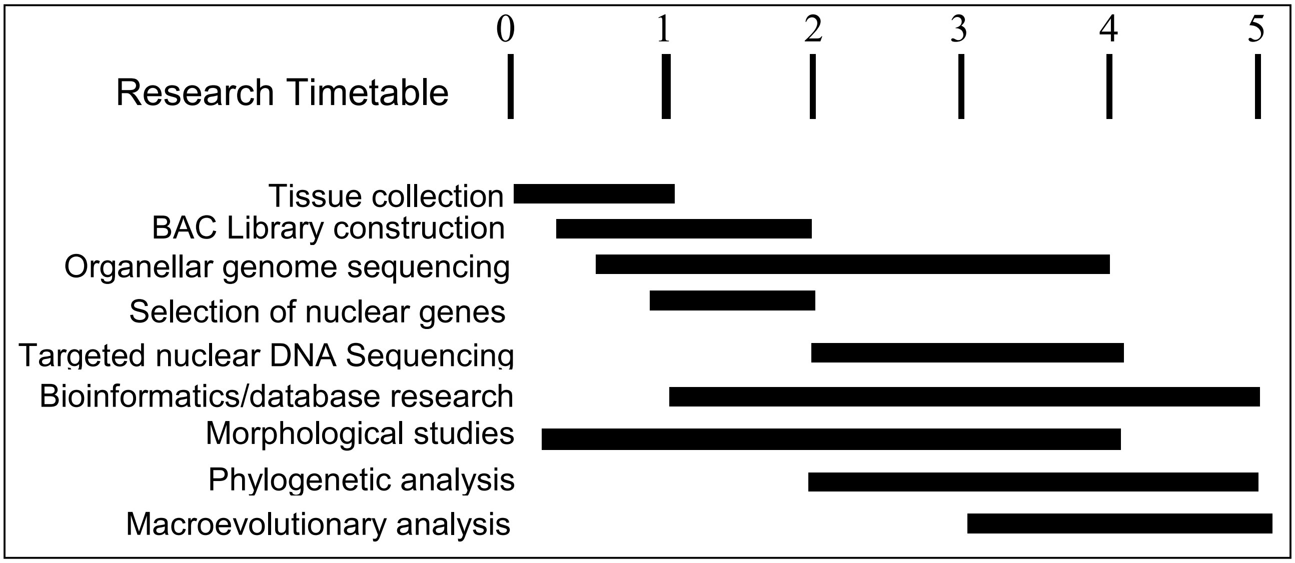
Charles O'Kelly: Project Coordinator; ultrastructural and morphological studies of "green algae" plus outgroups
Dina Mandoli: Coordination of BAC library production and interpretation; generation of BIG DNA for BAC construction; generation of data on nuclear genes
Richard Olmstead: Coordination of targeted sequencing of novel nuclear genes; phylogenetic analysis of organellar sequence data
Karen Renzaglia: Ultrastructural and morphological studies of land plants; coordination of educational activities
Paul G. Wolf: Coordinate acquisition of plant material, manage organelle extraction and DNA purification; shotgun sequence gap-filling
Brent D. Mishler: Coordinate with DEEP GENE, develop tools for phylogenetic analysis and databases, phylogenetic analysis of organellar genomic data; compartmentalization
Jeffrey L. Boore: Shotgun cloning; sequencing; assembly; annotation.
Alan R. Smith: Assist in plant material collection and identification; analysis of morphological data for land plants
Michael Donoghue: Construction of supertrees; study of character evolution; macroevolution
Rod Wing: BAC library production

***2. ROLES OF COLLABORATORS* (see also attached letters):**
Melvin Oliver: Assistance with isolating organelle genomes; proteomics of the chloroplast.
Robert Jansen: Coordinate chloroplast genome sampling with that of the seed plants; informatics and analysis of genomic characters.
Jonathan Shaw: Molecular phylogenetic studies of bryophytes
Richard McCourt: Supply charophyte material
Charles Delwiche: Supply charophyte material; study coevolution of organellar and nuclear genomes
Louise Lewis: Study other green transitions to land distinct from than embryophytes
Mark Buchheim: Phylogenetic studies of the chlorophytes
Marvin Fawley: Phylogenetic studies of the chlorophytes
Rick Zechman: Phylogenetic and evolutionary studies of the ulvophytes; assist in educational programs [letter pending]
Andy Knoll: Integration of fossils into deep green plant phylogentics
Joseph Hellerstein: Research on phylogenetic data bases for heterogeneous data
Bernard Moret: Development of novel phylogenetic algorithms; informatics, and databasing issues.
Michael Sanderson: Application of compartmentalization and supertree approaches; molecular clocks, macroevolutionary study [letter pending: Dr. Sanderson is on travel]
Roy Brown and Betty Lemmon: Ultrastructural and immunofluorescent studies of charophytes and embryophytes
Roberto Ligrone: Anatomical, immunological and ultrastructural studies of bryophytes and pteridophytes
Jeffrey Duckett: Morphological studies of streptophytes
Thomas Friedl: Phylogenetic studies of the chlorophytes [letter pending, Dr. Friedl is on travel]
B. Franz Lang: Organellar genome studies, with special reference to potential outgroups [letter pending]

Coordination of BAC and oBAC library construction and its integration with sequencing:



Grow or collect tissues/cells (Wolf, Smith)
genome size known
genome size unknown
Determine genome size (Arumaganathan)
Prepare BIG DNA (Mandoli, Wing)
DNA degraded or sparse
DNAbig
FACS organelles (Arumaganathan)
isolate organelles (Wolf)
Make libraries (Wing):
Isolate BIG DNA from gel
Restrict (partial) BIG DNA with EcoR1
Package DNA into bacterial vector
Select colonies with insert
Pick, store and array colonies
Perform quality control of libraries
Identify organellar genomes in arrays
Nuclear gene sequencing (Olmstead)
Subclone, sequence genomes & assemble contigs (Boore)
Gap filling (Wolf)
Phylogenetic analysis (Olmstead, Mishler)


Research Timetable
Tissue collection
BAC Library construction
Organellar genome sequencing
Selection of nuclear genes
Targeted nuclear DNA Sequencing
Bioinformatics/database research
Morphological studies
Phylogenetic analysis
Macroevolutionary analysis