ANNUAL REVIEWS

Annual Review of Plant Biology

Sequencing and Analyzing the Transcriptomes of a Thousand Species Across the Tree of Life for Green Plants

Gane Ka-Shu Wong,[1,2] Douglas E. Soltis,[3,4] Jim Leebens-Mack,[5] Norman J. Wickett,[6] Michael S. Barker,[7] Yves Van de Peer,[8,9] Sean W. Graham,[10] and Michael Melkonian[11]

[1]Department of Biological Sciences and Department of Medicine, University of Alberta, Edmonton, Alberta T6G 2E9, Canada; email: gane@ualberta.ca

[2]BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China

[3]Florida Museum of Natural History, Gainesville, Florida 32611, USA

[4]Department of Biology, University of Florida, Gainesville, Florida 32611, USA

[5]Department of Plant Biology, University of Georgia, Athens, Georgia 30602, USA

[6]Negaunee Institute for Plant Conservation Science and Action, Chicago Botanic Garden, Glencoe, Illinois 60022, USA

[7]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

[8]Department of Plant Biotechnology and Bioinformatics, VIB Center for Plant Systems Biology, Ghent University, 9052 Ghent, Belgium

[9]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa

[10]Department of Botany, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada

[11]Faculty of Biology, University of Duisburg-Essen, D-45141 Essen, Germany

Annu. Rev. Plant Biol. 2020. 71:741–65

First published as a Review in Advance on December 18, 2019

The *Annual Review of Plant Biology* is online at plant.annualreviews.org

https://doi.org/10.1146/annurev-arplant-042916-041040

Keywords

phylogenomics, paleopolyploidy, gene family expansions, bioprospecting, genetic networks

Abstract

The 1,000 Plants (1KP) initiative was the first large-scale effort to collect next-generation sequencing (NGS) data across a phylogenetically representative sampling of species for a major clade of life, in this case the

Viridiplantae, or green plants. As an international multidisciplinary consortium, we focused on plant evolution and its practical implications. Among the major outcomes were the inference of a reference species tree for green plants by phylotranscriptomic analysis of low-copy genes, a survey of paleopolyploidy (whole-genome duplications) across the *Viridiplantae*, the inferred evolutionary histories for many gene families and biological processes, the discovery of novel light-sensitive proteins for optogenetic studies in mammalian neuroscience, and elucidation of the genetic network for a complex trait (C_4 photosynthesis). Altogether, 1KP demonstrated how value can be extracted from a phylodiverse sequencing data set, providing a template for future projects that aim to generate even more data, including complete de novo genomes, across the tree of life.

Contents

INTRODUCTION

The 1,000 Plants (1KP) initiative generated transcriptomics data for over a thousand phylogenetically diverse species across the tree of life for green plants (*Viridiplantae*, a clade of at least 500,000 species of major ecological and economic importance). As a public-private initiative, the project's goals included basic and applied research. 1KP was launched in January 2009 at an organizational meeting hosted on the campus of the University of British Columbia in Vancouver, Canada. Illumina (Solexa) had just introduced their next-generation sequencing (NGS) platform (12, 149), and in subsequent years, many genomes were sequenced by NGS. Most of this effort was focused on economically important species. This included the publications of de novo genome assemblies for potato (107), *Brassica rapa* (150), and tomato (138), among many others. The mood of the era was succinctly captured by Jun Wang at BGI-Shenzhen, who stated, "If it tastes good, let's sequence it." This paraphrased the priorities of funding agencies worldwide, which understandably favored economically important species and placed a lower importance on the sequencing of phylogenetically diverse species. The objective of 1KP, by contrast, was to sample molecular diversity across the green plant phylogeny. We therefore employed transcriptome sequencing with de novo assembly to capture many thousands of expressed gene sequences from each selected species, at a fixed and affordable cost per sample. A reduced representation approach was necessary because, even today, with the $1,000 human genome a reality for resequencing, plant genomics can still be expensive for certain species. Genome sizes vary by 2,342-fold within land plants (74) and by 4,680-fold within green algae (65).

***Viridiplantae*:** green plants; eukaryotes with primary plastids and chlorophyll b, which include land plants (embryophytes) and green algae (nonmonophyletic; composed of chlorophyte and streptophyte algae)

Once committed to sequencing a phylogenetically representative collection of the *Viridiplantae*, many technical challenges had to be resolved. First, live tissues had to be collected at stages where homologous gene sets would be expressed in species with strikingly different life cycles. Because

major plant lineages can differ with respect to the dominance of either a diploid or a haploid generation, obtaining sufficient material from homologous phases of the life cycle was not always possible. RNA extractions were mostly carried out by laboratories close to the source so that more plant materials could be quickly provided when RNA of suitable quality and quantity was not obtained. No single RNA isolation protocol worked on all of the samples, but RNA was obtained for most of the targeted species using just a handful of protocols (63, 64). Second, de novo assembly of short NGS reads was not demonstrated for eukaryotes until 2010, upon publication of the panda genome (79). However, algorithms that were developed for genomes with uniform read coverages did not work on transcriptomes where read coverages vary with gene expression levels. The first algorithms to meet this challenge were Trans-ABySS (114) and Trinity (41). 1KP used a newer assembler, SOAPdenovo-trans (158), which performed at least as well as, and in some cases better than, Trinity (53). Likely sample misidentifications, mislabelings, and contaminations were detected by a comparison of 18S ribosomal RNA sequences with the SILVA database (159). Other problematic samples were detected by manual inspection of the computed gene trees in a comparison to the known species relationships.

A total of 1,342 transcriptomes [each with on average 2.2 gigabase pairs (Gbp) of high-quality raw sequence] were produced from 1,173 species representing all major *Viridiplantae* clades (19a). This substantially increased the number of genes for green plants, relative to what was then in the public sequence databases, especially in previously undersampled lineages (**Figure 1**). Researchers around the world were invited to analyze these data, and publications arising from these collaborations were compiled on a public website, **https://wiki.cyverse.org/wiki/display/iptol/OneKP+companion+papers**. For this review, we eschewed an exhaustive tabulation of 1KP papers to focus on a broader discussion of how one might extract value from a data set of gene sequences sampled over multiple species representing the diversity of a large taxonomic group. We begin at the highest level of abstraction, i.e., species, and move progressively to ever greater levels of detail, from genomes to gene families and from protein characterizations to complex genetic networks. The first three sections consider fundamental evolutionary questions and, in particular, discuss species tree inferences through phylotranscriptomic analyses, a survey of paleopolyploidy events across the tree of life for green plants, and assessments of gene family expansions (and contractions) in relation to the emergence of evolutionary innovations. The last two sections consider more applied issues, showing how quantitative protein characterizations facilitated optogenetic studies in mammalian neuroscience and how convergent evolution is a powerful method for investigating the evolution (or engineering) of complex traits like C_4 photosynthesis. Altogether, the multidisciplinary studies conducted for 1KP show how characterization of sequence diversity can have profound implications for human health, nutrition, and technology development.

Gene tree: phylogenetic history of an individual gene; sometimes used as a proxy for species-level phylogenetic history

Species tree: phylogenetic history of species relationships estimated from aligned gene sequences or gene trees, which may be concordant or discordant with speciation history

Phylotranscriptomics: comparative analysis of homologous RNA transcripts from multiple organisms; aims include inference of species phylogeny and related elucidation of gene, gene family, and genome evolution

Paleopolyploidy: whole-genome duplication in an ancestral lineage

C_4 photosynthesis: a supplementary method of carbon dioxide uptake that involves altered leaf anatomy and production of a 4-carbon molecule

EVOLUTIONARY HISTORIES

A key objective of 1KP was to improve our understanding of genome and gene evolution across the *Viridiplantae*. This requires a robust hypothesis of species relationships. When 1KP was launched, many of the outstanding uncertainties in the species tree were associated with poor taxon sampling and/or long branches leading to phylogenetically critical taxa. Plant molecular phylogenetics at that time was dominated largely by plastid markers and a handful of nuclear genes. The limitations of using uniparentally inherited nonrecombining loci, such as the plastome, were not being ignored; rather, large sets of nuclear encoded genes were not yet readily available, and scalable methods to account for complex gene histories (e.g., gene duplication and loss, incomplete sorting of the ancestral allele pool) were not yet developed. It was hoped that improved taxon and gene sampling would help break these long branches and provide a data set that more accurately

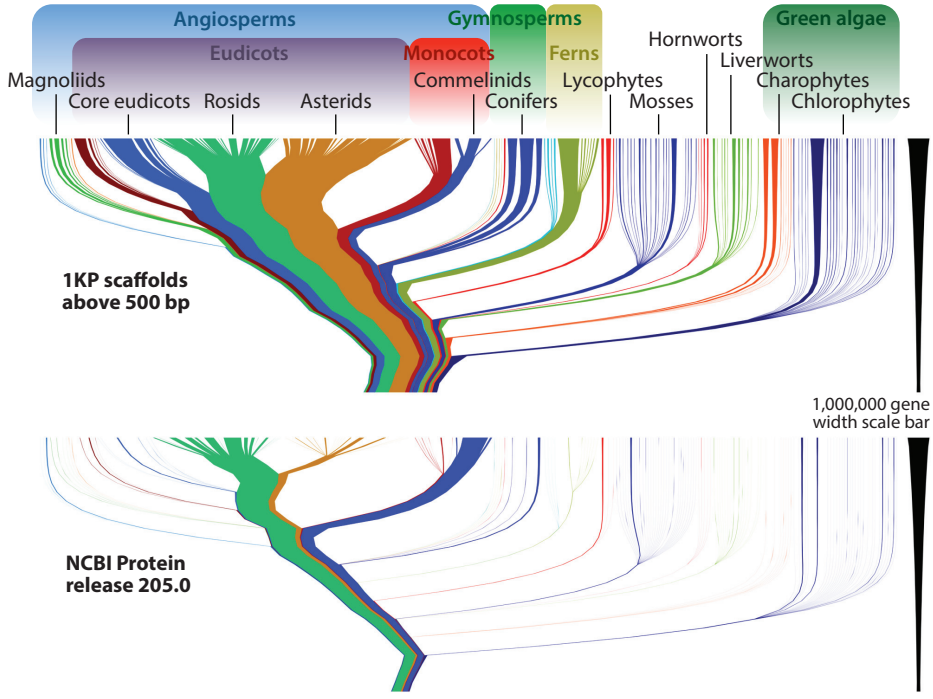


Figure 1

Number of plant genes (assemblies longer than 500 bp) sequenced by the 1,000 Plants (1KP) consortium versus the National Center for Biotechnology Information (NCBI) protein database (release 205.0). Branch widths are weighted by gene counts with a nonuniform scale along the vertical axis; black bars to the right indicate widths for one million genes [i.e., to create a tapered appearance, one million genes are depicted more narrowly as we progress from the leaf (*top*) to the root (*bottom*) of the tree]. The median ratio of 1KP to NCBI gene counts for families and orders is 66 and 46, respectively.

reflected the complex speciation history of nuclear genomes. However, we had no a priori idea of the extent to which this might ameliorate the difficulties of inferring deep phylogenies (104). Even before the full complement of 1KP analyses was complete, these data improved the understanding of diversification within and among plant lineages where few nuclear gene sequences were previously available. For example, research on fern relationships and diversification was advanced using low-copy nuclear gene sequences from 1KP (117). 1KP data also helped to facilitate the development of microsatellite primers and probes for target enrichment (49). This review focuses on a capstone analysis (73) of the full 1KP data set, with a particular emphasis on the inferences of evolutionary histories that are reflected in the gene tree and species tree estimates. As a primer for the nontaxonomist readers, **http://www.mobot.org/MOBOT/research/APweb/trees/modeltree3.gif** shows the relationship between key model organisms and major angiosperm clades.

Phylotranscriptomics-Inferred Species Trees

Even as plant phylogenetics moved to large-scale phylogenomic investigations, as opposed to inferences based on small sets of markers, reconstructions of plant species relationships focused mostly on sequencing and analyzing plastid genomes (56, 72, 99, 139). Later studies using

low-copy nuclear gene sequences extracted from genomes (126), transcriptomes (152, 160), and gene capture data (45, 46) were mostly consistent with the plastome-based phylogenies, albeit with occasional conflicts. Analyses of the full 1KP data set used both plastid-encoded genes and 410 low-copy nuclear genes that were mined from the transcriptomics data. Nuclear gene sequences yielded robust and consistent phylogenetic inferences for most *Viridiplantae* clades regardless of how the DNA or amino acid sequence alignments were analyzed, either as concatenated supermatrices or as species tree analyses that accounted for incomplete lineage sorting (ILS) (73). More importantly, species trees inferred from the nuclear gene sequences were largely concordant with species trees inferred from coding portions of the plastid genome.

Supermatrix: DNA or amino acid alignments where individual genes are concatenated into a single data matrix for phylogenetic analysis

Incomplete lineage sorting (ILS): retention of ancestral polymorphisms through successive speciations, such that individual gene tree histories do not necessarily match species history

It is, however, noteworthy that discordances among the gene histories for some important nodes in the species tree are still observed. For example, coalescence-based analyses of nuclear genes recovered a clade of extant bryophytes (mosses, liverworts, and hornworts), in contrast to a previous plastid-based inference that recovered liverworts as sister to all other extant land plants and hornworts as sister to the vascular plant clade (108). Another example concerns the relationships among core eudicot lineages, which comprise 75% of angiosperms. Plastid gene and genome analyses gave strong support for the placement of a clade incorporating the orders *Celastrales*, *Oxalidales*, and *Malpighiales* (COM) within *Fabidae*, a large subclade of rosids. In contrast, analysis of low-copy nuclear genes placed these orders within another major rosid subclade, *Malvidae*. It is unclear whether discordances between plastid and nuclear trees are a consequence of ILS (**Figure 2*b***), introgressive hybridization (**Figure 2*c***), or nonorthologous gene retention (**Figure 2*d***) following the gamma genome triplication that predated rapid core eudicot diversification (59).

Large-scale sampling of nuclear genes is necessary to characterize discordance not only between plastome and nuclear genomes but also among nuclear genes. All of the processes shown in **Figure 2** can give rise to heterogeneous gene histories and gene tree–species tree discordance. Horizontal gene transfer involving movement of genes between distantly related organisms can typically be diagnosed using gene tree comparisons (78), but ILS generates subtle gene tree–species tree discordance that can be difficult to distinguish from gene tree estimation error. In the 1KP analyses, discordance among individual gene histories was common. When most of the sampled genes share a fundamentally concordant history, concatenating individual gene alignments into a supermatrix and using an appropriate nucleotide or amino acid substitution model can mitigate many of the problems associated with the analysis of large nuclear gene sets (18, 103). For example, likelihood and Bayesian analyses can employ parameter-rich models to account for variation in substitution rates among aligned sites within and among genes sharing the same phylogenetic history (133). However, when many genes have conflicting histories, concatenation of individual gene alignments can be statistically inconsistent (70, 115) and methods are required to account for processes such as ILS.

ILS occurs when ancestral allelic variation is retained between speciation events, and random fixation of these allele lineages within the resulting species gives rise to gene tree–species tree discordance (**Figure 2*b***). The nonparametric summary methods implemented in the species tree estimation program ASTRAL II (87, 121) have been shown to accurately reconstruct species relationships by assessing the frequencies of species quartets observed within and across gene tree estimates (86). Although parametric methods can recover species trees with greater statistical efficiency when genes evolved under a pure coalescence model (110), the nonparametric species quartet summary method implemented in ASTRAL II is robust in the presence of low-to-moderate rates of horizontal gene transfer (23). Moreover, summary methods that aim to identify a species tree exhibiting maximum quartet support across estimated gene trees may also be robust to gene tree discordance due to interspecific gene flow (**Figure 2*c***) or gene duplication/loss (**Figure 2*d***), if the probability of hybridization or retention of paralogous gene copies decreases with increasing

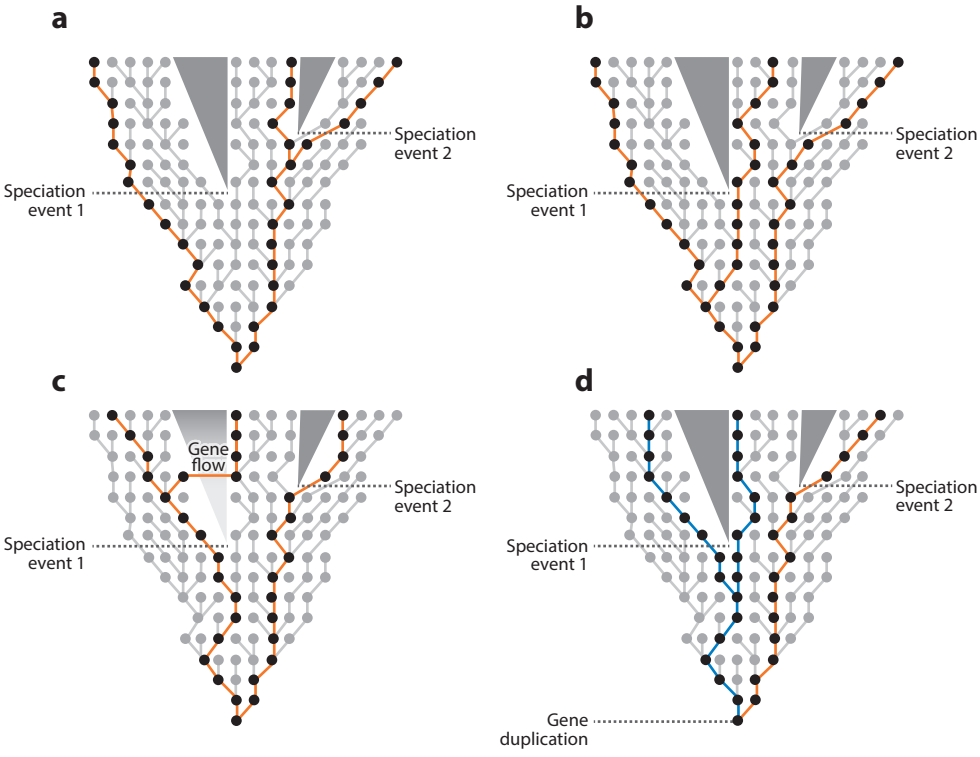


Figure 2

Gene genealogies within species trees, adapted from References 75 and 81, illustrating how (*a*,*b*) coalescence, (*c*) gene flow, and (*d*) gene duplication can produce gene trees that are (*a*) concordant or (*b–d*) discordant with species trees. The species trees all depict the same two speciation events, but the ancestries of the sampled genes differ, exhibiting (*a*) shallow coalescence, (*b*) deep coalescence due to incomplete lineage sorting between speciation events, (*c*) gene flow, and (*d*) retention of paralogous genes (*blue* and *orange*) following an ancient gene duplication. Black/gray dots represent sampled genes and their ancestors with lines tracing the ancestry of sampled genes. Gray wedges represent barriers to gene flow.

Homoplasy: trait similarity due to evolutionary convergence, parallelism, or reversal (i.e., similarity not due to common ancestry)

Hemiplasy: homologous (i.e., sharing a common ancestry) character states that, due to incomplete lineage sorting, appear homoplasious when mapped onto a species tree

time between speciation, similar to the probability of ILS. Many critical relationships throughout the diversification of plants are characterized by short branches that represent brief times between successive speciations, so the use of summary methods that account for processes disproportionately affecting these nodes was essential. Critically for 1KP analyses, ASTRAL II can rapidly estimate species trees from large data sets with thousands of genes and taxa. Gene tree summary methods are also robust to missing data (88, 157) when the most fragmentary sequences (<33% complete) are removed from the gene sequence alignments (122). This is important because incomplete sampling of expressed genes and fragmentary gene assemblies are common issues for transcriptomics data.

Comparative analyses typically aim to map ecological or phenotypic trait changes onto an estimated species tree to determine if those traits arose once or multiple times and whether trait changes are associated with species diversification. Traits with simple genetic architectures may appear to arise multiple times as a result of homoplasy (i.e., parallel or convergent evolution), but this could also reflect discordance between the species phylogeny and the histories of the genes encoding a particular trait. The latter is referred to as hemiplasy (3) and is a clear possibility for those nodes in a species tree with extensive discordance among the gene trees due to ILS, introgressive

hybridization, gene duplication/loss, or rapid radiations in which more than two daughter species are derived from a single ancestral species (i.e., hard polytomies) (98). Nodes exhibiting extensive gene tree discordance and increased opportunity for hemiplasy include those representing the last common ancestors of the core eudicots, land plants, and green algal class Ulvophyceae. In the face of true gene tree–species tree discordance, no single bifurcating species tree adequately describes the history of all the genes. Importantly, many of the historically problematic relationships in the plant tree of life have been characterized by varying levels of gene tree discordance. This includes the monophyly of bryophytes, the monophyly of conifers, the placement of *Amborella* as the sister group of all other living flowering plants, and the placement of *Equisetales* as sister to the remaining extant monilophytes (ferns). All of these inferred relationships exhibited robust support in species tree estimates despite moderate levels of gene tree discordance (73). Conversely, positioning of the *Gnetales* within conifers, positioning of the *Bryopsidales* among chlorophyte algae, and resolution of the early radiation of core eudicots all exhibited stronger support for alternative hypotheses in analyses of concatenated gene alignments (73). For the latter cases, well-supported conflict among gene trees suggests that the genes do not share the same evolutionary history and that concatenation of gene alignments in a supermatrix analysis may be problematic.

Ancient Whole-Genome Duplications

Renewed interest in polyploidy dates back to the sequencing of the *Arabidopsis* genome, which, despite having only five chromosomes and being approximately 150 megabase pairs (Mbp) in length, experienced at least three rounds of whole-genome duplications (WGDs) in its evolutionary history (16, 147). There is now evidence for many tens, perhaps even hundreds, of ancient plant WGDs (8, 129, 143, 151). The data suggest that one-third of all plant species are paleopolyploids (6, 84, 156). There is also a correlation between recent polyploidy and domesticated plants (119). What is still being debated is whether recently formed polyploid species diversify at higher rates than their diploid relatives (32, 66, 83, 84, 123, 128). Another issue is the link between ancient WGDs and the evolution of plant diversity (30, 135). Despite this flurry of activity, many important questions remain unanswered due to poor or phylogenetically unbalanced sampling of plant genomics data (**Table 1**).

Analysis of the full 1KP data set produced the first comprehensive survey of paleopolyploidy across the green plant phylogeny (73). Using a combination of gene age distributions (7, 9) and multispecies phylogenomic methods (80), evidence was found for 244 putative WGDs, of which 138 were not previously reported. These include putative WGDs in all extant monilophytes (ferns and their close relatives), all extant mosses, and WGDs deep in hornwort and liverwort history. These investigations also recovered support for the placements of many previously reported WGDs in plant phylogeny, including one in seed plants (60, 80), an ancestral angiosperm case (1), and hexaploidization in the core eudicots (59, 146). Consistent with previous work (5), the only major land plant clade with no evidence of paleopolyploidy was the lycophyte *Selaginella*. This may be why they have some of the smallest genomes and the slowest rates of genome-size evolution among vascular plants (4). Given that the genomes of many other land plants have been duplicated at least once, *Selaginella* is an intriguing outlier and an exemplar for evolution of land plant genomes in the absence of polyploidy.

A major insight from 1KP was that the incidence of paleopolyploidy differs substantially across lineages of green plants. WGDs were relatively rare among the green algae (both chlorophytes and streptophytes). In contrast, species within most clades of land plants experienced one to two rounds of WGDs, with many angiosperms demonstrating evidence for least five rounds of WGDs (**Figure 3**). Intriguingly, the green algal clade with a frequency of paleopolyploidy most similar

Table 1 Omic resources by species for land plants and relatives[a]

Higher-level taxonomy	Major clade	Number of families	1KP transcriptomes	NCBI [or Phytozome] genomes
Red algae			28	4 [1]
Glaucophytes			4	0 [0]
Green plants/chlorophyte algae			112	31 [8]
Green plants/streptophyte algae			46	1 [0]
Streptophytes/embryophytes/bryophytes	Mosses	111	41	1 [2]
	Liverworts	87	22	1 [1]
	Hornworts	5	8	0 [0]
Streptophytes/embryophytes/vascular plants	Lycophytes	3	21	2 [1]
	Monilophytes (ferns)	48	77	0 [0]
Streptophytes/embryophytes/vascular plants/seed plants	Gymnosperms	12	82	5 [0]
Vascular plants/seed plants/angiosperms/ANA-grade angiosperms	*Amborellales*	1	1	1 [1]
	Nymphaeales	3	2	0 [0]
	Austrobaileyales	3	4	0 [0]
Angiosperms/magnoliids	*Canellales*	2	2	0 [0]
	Laurales	7	11	0 [0]
	Magnoliales	6	6	0 [0]
	Piperales	3	6	0 [0]
Angiosperms/unplaced	*Ceratophyllales*	1	1	0 [0]
	Chloranthales	1	2	0 [0]
Angiosperms/monocots	*Acorales*	1	1	0 [0]
	Alismatales	14	4	2 [2]
	Asparagales	14	45	4 [0]
	Dioscoreales	3	1	1 [0]
	Liliales	10	6	0 [0]
	Pandanales	5	5	1 [0]
	Petrosaviales	1	0	0 [0]
Angiosperms/monocots/commelinids	*Arecales*	2	5	3 [0]
	Commelinales	5	0	1 [0]
	Poales	14	27	38 [12]
	Zingiberales	8	7	3 [2]
Angiosperms/eudicots	*Buxales*	1	1	0 [0]
	Proteales	4	6	1 [0]
	Ranunculales	7	22	1 [1]
	Trochodendrales	1	1	0 [0]
Angiosperms/eudicots/core eudicots	*Gunnerales*	2	1	0 [0]
	Dilleniales	1	2	0 [0]
Angiosperms/eudicots/core eudicots/superrosids	*Saxifragales*	15	24	0 [2]
…Superrosids/rosids	*Vitales*	1	3	3 [1]
…Superrosids/rosids/fabids	*Celastrales*	2	2	0 [0]
	Cucurbitales	8	4	5 [1]
	Fabales	4	27	16 [5]
	Fagales	7	11	4 [0]
	Malpighiales	36	43	8 [6]
	Oxalidales	7	4	1 [0]
	Rosales	9	25	17 [3]
	Zygophyllales	2	3	0 [0]

(*Continued*)

Table 1 (*Continued*)

Higher-level taxonomy	Major clade	Number of families	1KP transcriptomes	NCBI [or Phytozome] genomes
…Superrosids/rosids/malvids	*Brassicales*	17	23	28 [10]
	Crossosomatales	7	2	0 [0]
	Geraniales	2	4	0 [0]
	Huerteales	4	1	0 [0]
	Malvales	10	10	10 [3]
	Myrtales	9	30	4 [1]
	Picramniales	1	0	0 [0]
	Sapindales	9	14	9 [3]
Angiosperms/eudicots/core eudicots/superasterids	*Berberidopsidales*	2	2	0 [0]
	Caryophyllales	38	66	11 [2]
	Santalales	7	7	0 [0]
…Superasterids/asterids	*Cornales*	7	9	0 [0]
	Ericales	22	25	6 [0]
…Superasterids/asterids/campanulids	*Apiales*	7	12	1 [1]
	Aquifoliales	5	4	0 [0]
	Asterales	11	39	5 [1]
	Bruniales	2	0	0 [0]
	Dipsacales	2	6	0 [0]
	Escalloniales	1	2	0 [0]
	Paracryphiales	1	0	0 [0]
…Superasterids/asterids/lamiids	*Boraginales*	1	15	0 [0]
	Garryales	2	2	0 [0]
	Gentianales	5	17	3 [1]
	Icacinales	2	2	1 [0]
	Lamiales	24	64	15 [1]
	Metteniusales	1	0	0 [0]
	Solanales	5	24	19 [2]
	Vahliales	1	0	0 [0]

[a]Contrasting taxonomic diversity of transcriptomes generated by the 1,000 Plants (1KP) consortium with public genomes from the National Center for Biotechnology Information (NCBI) Genome database (**https://www.ncbi.nlm.nih.gov/genome**) and the Phytozome v12.1 database (**https://phytozome.jgi.doe.gov/pz/portal.html**), both as of August 2017.

to land plants was *Zygnematophyceae*, although those WGDs were concentrated in one clade of this class. *Zygnematophyceae* have recently been recognized as the most likely sister lineage to land plants (73, 136, 152). This finding suggests that the proclivity for polyploidy in land plants has very deep evolutionary roots.

Clearly, paleopolyploidy has been rampant in *Viridiplantae* evolution, but a mechanistic understanding of the consequences of WGDs for plant evolution is still the subject of vigorous debate (34, 129, 143). For example, there are claims that they explain major evolutionary transitions like the sudden appearance and radiation of flowering plants (Darwin's abominable mystery) (60). They may also be a mechanism for increasing diversification rates (6, 135). Although mathematical models (95) and a recent yeast experiment (124) suggest that polyploid genomes can generate and harbor more genetic variation than diploid genomes, much remains to be done to connect microevolutionary processes with macroevolutionary patterns of evolution. For instance, notwithstanding the many WGDs inferred for some plant lineages, there is a limited range of

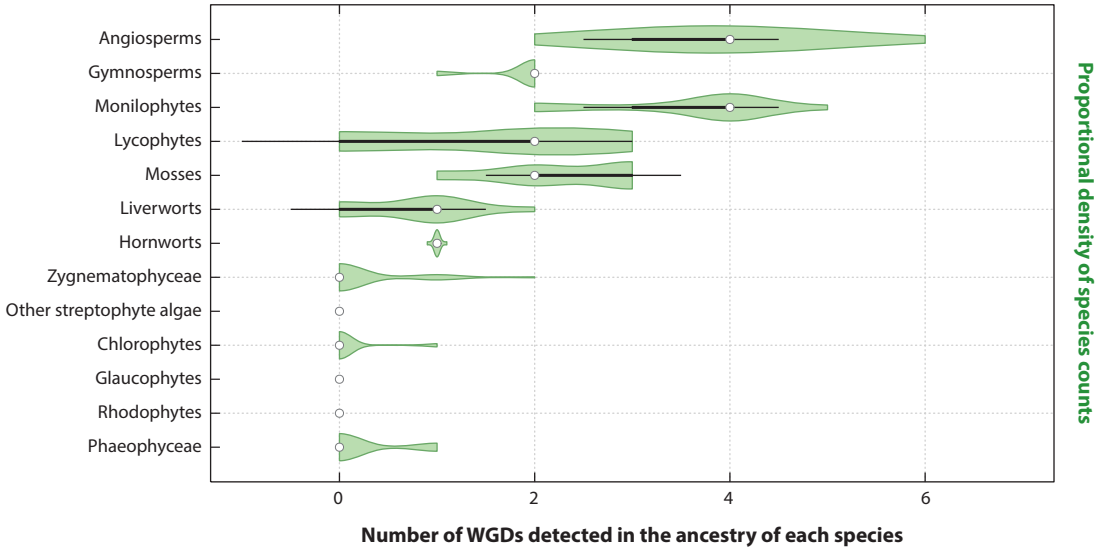


Figure 3

Violin plots showing the number of rounds of whole-genome duplications (WGDs) inferred in the ancestry of each species within each clade. We counted WGDs from root to tip in the 1,000 Plants (1KP) phylogeny. Results varied from zero in many algal lineages to as many as six for some angiosperms. The white dot is the median, the thick black bars show the interquartile range, the thin black lines define the 95% confidence interval, and the green shading represents the density of data points. Thicker areas of shading represent relatively higher concentrations of species with a particular number of WGDs, whereas thinner areas indicate relatively fewer species.

variation in the number of polyploid lineages that survive long term within these clades. One explanation for the restricted number of WGDs in the ancestry of each lineage is that the probability of polyploid survival may increase due to the reduced competitive pressure that occurs during periods of environmental upheaval (143). If so, the number of rounds of WGDs should reflect the number of climatic perturbations, and there would be a similar number of rounds of WGD among lineages. Alternatively, the observed incidences may simply reflect differences in the average rates of diversification of polyploid and diploid species (84, 123).

The relative scarcity of WGDs in algae studied by 1KP, as compared to land plants, raises additional questions. Although it may be that polyploid speciation is more frequent or has more advantages on land, it is not clear why the incidence of paleopolyploidy would have shifted with the colonization of land. The transition to land is associated with other genome-level changes, such as an increase in the number of genes (54, 137). This increase may be related to shifts in gene and genome duplication rates among different lineages of green plants. Whatever the resolution of these questions might be, 1KP data will invigorate research into the century-old question of why polyploidy is so common in many plant groups (89, 94).

Evolution of Known Genes and Processes

1KP collaborators performed many studies of genes and processes of wide interest. A partial listing of the studied genes (or gene families) includes *LEAFY* transcription factors (120), the horizontally acquired neochrome photoreceptor (78), auxin efflux carriers (11), phytochromes (76), phototropins (77), nitric oxide synthase-like genes (57), *NAC* transcription factors (82), microbial-type terpene synthase genes (58), histone deacetylases (15), hydroxyproline-rich glycoproteins (62), and strigolactone receptor genes (19). Typically, expert consultants were needed to define the bioinformatic algorithms for identifying genes, and in one instance, a novel algorithm was developed

to detect intrinsically disordered proteins (61). Other investigators studied multiple (different) genes associated with specific biological processes, such as fruit development (96), crassulacean acid metabolism (CAM) and C_4 photosynthesis (21), arbuscular mycorrhizal symbiosis (24), and leaf evolution (145). The capstone analysis of the full 1KP data set (73) also considered *N*-fold (or multifold) changes in gene family sizes that are only discernable within larger gene families, given the limitations of transcriptomics data. *Arabidopsis*, for example, has 26 gene families with at least 53 members (91). The majority of these are transcription factors; many others are genes involved in signaling pathways with regulatory functions. Among the 23 large gene families that were analyzed, there were 6 births, 41 expansions, and 9 contractions across *Archaeplastida* (*Plantae*) history. Not including births, the most prominent events were expansions of 39.5- and 13.5-fold in size. All expansions occurred early in streptophyte evolution, and only one coincided with the 244 ancient WGDs identified across the green plants.

On the early appearance of land plant genes. Recent transcriptomic and genomic studies of the closest algal relatives of land plants (streptophyte algae) have documented that the molecular tool kit for life in a terrestrial environment was already present in streptophyte algae (10, 26), before the origin of land plants. Genes once thought to have been restricted to land plants are now being detected in streptophyte algae. Examples include those involved in symbiotic interactions with soil microbes (24), secondary metabolism and phytohormonal signaling (27, 50, 54, 144), desiccation/stress responses (51, 52), plastid/nucleus communication (28), and cell-wall biosynthesis (85, 130). Critically, many transcription factors thought to be specific to land plants not only originated in streptophyte algae but also expanded significantly there. *NAC* transcription factors, one of the largest families of plant-specific transcription factors, are involved in many biological processes including organ development and response to biotic and abiotic stresses (90). They have been detected in subgroups of streptophyte algae using data from 1KP (82). Similar results have been reported in other transcription factors such as *LEAFY* (120) and *GRAS* (73). Of the large gene families analyzed for the 1KP capstone, seven revealed a minimum of threefold expansions in streptophyte algae, compared with three in other green algae and two in vascular plants. Of the six gene births, two were found in streptophyte algae, four in other green algae, and none in embryophytes. These findings raise many new questions about the functional role of land plant genes that appeared substantially earlier in green plant phylogeny and highlight the need to develop a genetically tractable model system for streptophyte algae (29, 111).

It is a longstanding misconception that terrestrial habitats supported no photosynthetic organisms until the origin of land plants. In reality, many photoautotrophic microorganisms can inhabit terrestrial habitats, growing when water is available and persisting in a dry state when it is not. Microbial crusts, containing mostly cyanobacteria, have covered terrestrial surfaces for billions of years. It is highly likely that the origin of plastids and therefore *Plantae* (*Archaeplastida*) occurred in a subaerial/terrestrial environment (14, 106). Two of the three *Plantae* clades are either exclusively seen in freshwater (glaucophytes) or likely had a subaerial/terrestrial origin (red algae). It has even been argued that streptophyte algae were terrestrial from the beginning (44). Subaerial/terrestrial versus aquatic habitats often coexist or are identical and only temporally separated (e.g., when a puddle dries out, or when a depression fills with water after rain), so streptophyte algae have experienced both submerged and dry environments since their origin. Even for subgroups of streptophyte algae that now exclusively thrive in aquatic habitats (*Coleochaetophyceae*, *Charophyceae*), persistence in a subaerial/terrestrial environment is possible (42, 116).

Recent phylogenomic analyses have concluded that the *Zygnematophyceae* represent the most likely sister group of land plants (73, 136, 152, 155). When compared to the more deeply diverging *Charophyceae* and *Coleochaetophyceae*, *Zygnematophyceae* are structurally simple (unicells or

***Archaeplastida* (*Plantae*):** a eukaryotic supergroup with primary plastids consisting of *Viridiplantae* (green plants), *Glaucophyta*, and *Rhodophyta* (red algae)

Streptophyte algae: a grade of green algal lineages that share more recent common ancestry with embryophytes (land plants) rather than chlorophytic green algae

Subaerial habitats: terrestrial environments that are not underwater or underground

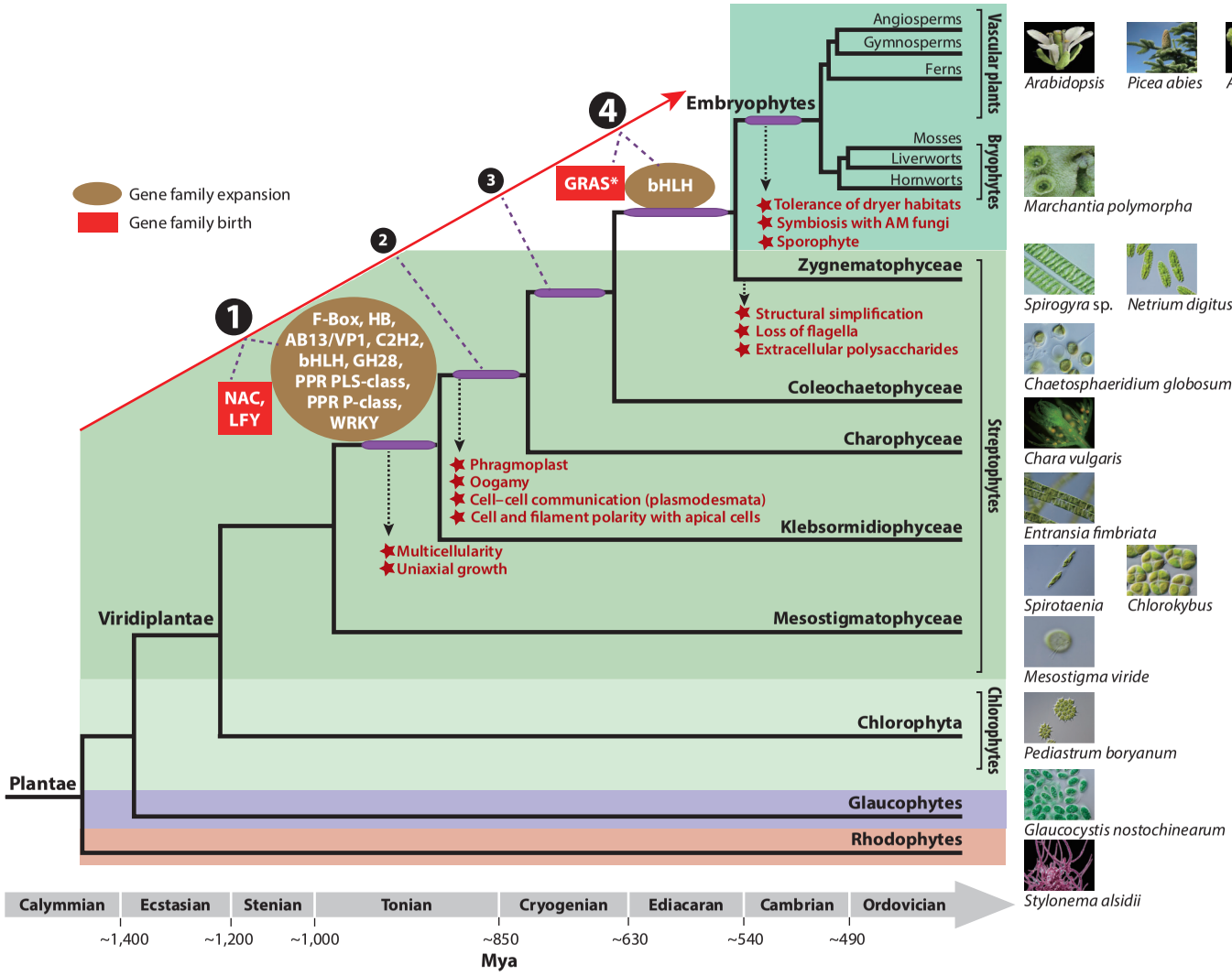
unbranched filaments). This is now interpreted as a secondary simplification, related to adaptation to drier habitats (26, 152). Consistent with this finding, lineages that span the earliest splits in the *Zygnematophyceae* phylogeny tend to thrive in subaerial/terrestrial environments. Like many other organisms with assumed terrestrial ancestry and external fertilization, such as fungi, red algae, and pennate diatoms, *Zygnematophyceae* apparently lost all flagellate stages early in their evolutionary history. Copious extracellular polysaccharides are produced by these organisms, not only to aid in sexual reproduction but also to slow down desiccation. Later in their evolution, many species and lineages of *Zygnematophyceae* adapted to a specific aquatic habitat (nutrient-poor bogs) in which production of their anionic extracellular polysaccharides would have conferred a selective advantage by trapping and thus concentrating scarce cations (nutrients).

There is now a compelling argument that adaptation to subaerial/terrestrial habitats is a feature of streptophyte algae, arising from their dual existence in aquatic and subaerial/terrestrial environments throughout their evolutionary history. Genes/proteins involved in adaptation to subaerial/terrestrial habitats were gradually acquired in streptophyte algae, ultimately culminating in the common ancestor of *Zygnematophyceae* and embryophytes. Mapping traits, gene families, and their expansions on the streptophyte phylogeny (**Figure 4**), in conjunction with functional studies in model streptophytes, may provide clues as to the sequence of evolutionary steps that were important to set the stage for the emergence of land plants, which forever changed the surface of our planet.

The importance of a more phylodiverse sampling. Continuing growth in the number of plant genomes available in databases like Phytozome (36), which had 77 (mostly angiosperm) genomes as of August 2017, is a positive for comparative plant genomics, but the collections of species in these databases remain an enormously biased subset of *Viridiplantae* lineages (**Table 1**). This reflects historical funding priorities, which focused mostly on crops, model plants, and their close relatives. Although the associated research programs have been spectacularly successful in expanding our understanding of plant functional, molecular, and developmental biology, initiatives such as 1KP and the *Amborella* Genome Project (1) point to a growing appreciation of the need for more broadly based comparative genomic tools in plant biology.

Broader sampling of green plant diversity is required to discern those gene families or gene regulatory networks that are localized to specific subclades of the *Viridiplantae* versus those that had a more ancient origin. For example, some gene regulatory networks related to the angiosperm carpel, the edible angiosperm reproductive structure in all grain and fruit crops, utilize homologs that originated in recently evolved clades of angiosperms (such as *Brassicales*). The availability of nonangiosperm genomes has allowed us to infer that significant portions of the carpel gene regulatory networks originated much earlier in land plant evolution, before the origin of the carpel or even flowering plants. However, significant phylogenetic gaps in genome availability prevent us from understanding exactly where and how they originated. For example, carpel-related genes like *SPATULA* (*SPT*) and *ALCATRAZ* (*ALC*) are present in the common ancestor of seed plants but not in mosses or *Selaginella* (101, 102) and may be shared with ferns. To examine this possibility, future analyses will need to include a fern genome.

Better-populated studies will allow us to capture hidden gene duplication events, which if missed can lead to inaccurate orthology assessments, inaccurate predictions of gene function, and underestimates of the age and diversity of gene families (97). Correct orthology assessment is also an implicit requirement for understanding neo- and subfunctionalization following gene duplication (55). This is best performed with a phylogenetic approach, as was done for a study of leaf development regulators, which identified a probable neofunctionalization in the ancestor of euphyllophytes (145). Understanding homology in a phylogenetic context also allows us to identify



(Caption appears on following page)

Figure 4 (*Figure appears on preceding page*)

Mapping the origins (births) and expansions of large gene families (known from *Arabidopsis*) onto streptophyte branches of a simplified 1,000 Plants (1KP) phylogeny for the *Plantae* (*Archaeplastida*). The purple colors link the step numbers (1 through 4) to the corresponding branches of the species tree. Transcription factors involved in stress responses and interactions of plants with soil microbes, traits known to be important for the process of terrestrialization, originated/expanded throughout the evolution of streptophyte green algae, but mostly in step 1 (early) and step 4 (late). In the case of *GRAS*, the asterisk (*) indicates that this gene family was born and expanded on this branch. Notice that evolution of structural complexity in streptophyte algae (step 2) did not coincide with births or expansions of these large gene families. Algae photographs are courtesy of the Culture Collection of Algae at the University of Cologne (**http://www.ccac.uni-koeln.de**), Gerd Günther (**http://www.mikroskopia.de/index.html**), and Ingo Botho Reize (**http://www.fotosdernatur.de**). Photos of *Arabidopsis*, *Picea*, *Azolla*, and *Marchantia* are adapted from Wikimedia Commons, public domain. Abbreviations: AM, arbuscular mycorrhizal; bHLH, basic helix-loop-helix; Mya, million years ago.

horizontal gene transfers. 1KP examples include horizontal gene transfers of light-sensing (78) and terpene synthase (58) genes.

The paucity of phylogenetically representative large-scale omic resources has long stymied studies of gene functional diversity outside of model-plant systems (69, 112, 125). 1KP was but a start in the effort to correct this shortfall. One might also hope that access to phylodiverse sequence data would allow researchers to explore previously untapped pools of biological innovation and facilitate the improvement of crop plants (13). The practical implications, however, extend beyond crop plants. In the final two sections of this review, we describe the less conventional approaches used by 1KP investigators to successfully solve engineering challenges that were otherwise intractable to more traditional methods.

DIVERSITY APPLICATIONS

The genome sequences of all extant species constitute a record of the outcome of billions of years of evolutionary experiments, with the life of each individual in each species representing a separate experiment. As scientists, we cannot hope to replicate the number of experiments that were performed over the eons, and the challenge for us is learn how to read from evolution's laboratory books. This can be done at two levels of complexity. The simplest approach is to search for useful biomolecules, e.g., secondary metabolites or proteins. However, many of the traits that characterize species are complex, necessitating the solution of genetic networks for a hundred or more genes. This is obviously a much more difficult problem.

Quantitative Protein Characterizations

Perhaps the most oft-cited application of biodiversity is the discovery of a novel bioactive or medicinal compound. There is no question that many approved therapeutic agents are either natural products or directly derived from them (92). However, to the extent that the bioactive compound is a secondary metabolite, it is difficult to predict what metabolites are produced by a plant species, given just its genome or transcriptomes. Algorithms to do so are in their infancy at best, and 1KP was not funded to do metabolomics. In addition, even if we were given a species of known medicinal value, it is difficult to identify what the bioactive compound might be. We therefore took a different approach to bioprospecting, focusing instead on proteins and searching for quantitative differences in categories of proteins with known uses. The importance of quantitative differences is not well appreciated in biology, but it is conventional wisdom in engineering (71). People who create tools for biological research understand this—particularly those in optogenetics (17), the Method of the Year 2010 for *Nature Methods*.

Optogenetics: techniques using light to control living cells that have been genetically modified to express light-sensitive proteins

In its most common incarnation, optogenetics is the use of light to activate genetically and spatially defined cells in intact tissues of awake and behaving animals. For example, one experiment created false memories of an active fear response in mice, by manipulating memory engram-bearing cells in their hippocampus (109). The most prevalent optogenetics tool is a blue-light-activated cation channelrhodopsin, ChR2, originally found in *Chlamydomonas reinhardtii*. This is put into a gene therapy vector and injected into a live animal. Cell-type-specific gene expression is achieved through cell-specific promoters. Implanting a fiber-optic cable in the brain provides spatial control of neuronal activation, with millisecond temporal resolutions that are compatible with neuronal action potentials. Prior to 1KP, researchers had spent nearly a decade trying to improve the ChR2 protein with the goal of making it more sensitive, more efficacious (larger photocurrent), faster, or red-shifted, which is important because red light penetrates deeper into the brain. However, the latter spectral-shifting efforts were largely in vain. By contrast, we put our faith in biodiversity and searched 1KP for novel channelrhodopsin sequences. Ed Boyden's laboratory at MIT heterologously expressed these genes in mouse neurons and quantified their functional parameters using an automated microscope customized for whole-cell electrophysiology. The results were reported in a pair of papers (48, 68). In *Chlamydomonas noctigama*, they discovered Chrimson, which is still the most red-shifted channelrhodopsin known. Chronos, an unusually fast blue-channelrhodopsin, was found in *Stigeoclonium helveticum*. Other discoveries included a very strong channelrhodopsin, CoChR, and a blue-shifted channelrhodopsin, Cheriff. These new proteins are now among the fastest spreading molecular tools in neuroscience. Most of the initial publications arising from these discoveries have been on *Drosophila*, simply because flies grow faster than mice. For example, light-sensing proteins encoded by these new genes have been used to deliver wake-promoting signals (105) and to trigger neural circuits in a demonstration of a novel method that labels active neural circuits in vivo (33).

Another example of the power of searching biodiversity was the discovery of channelrhodopsins to inactivate neurons in response to light. The original channelrhodopsins admitted positively charged ions and as such were called cation channelrhodopsins. Anion channelrhodopsins (ACRs) were not previously known to exist. By mutating a single amino acid, E90, it is possible to convert a cation channelrhodopsin into an ACR (153), but the response to light is not very strong. When the first naturally evolved ACR was finally discovered, in *Guillardia theta* (39), it proved to be a hundred times stronger. Upon further examination of other naturally evolved ACRs in marine cryptophytes, it became apparent that none changed E90, and many other amino acids were changed instead (40). In other words, evolution found a better solution, but it is more complicated. ACRs have already been used to inhibit cardiomyocyte electrical activity (38), and the search for even better variants continues, again using the biodiversity approach.

Optogenetics is a widely used tool in biological research because much of what we as biologists do is done under a light microscope. Channelrhodopsins allow us to couple light to the electrical signals of the brain, but cells can also communicate by chemical signals. Proteins that couple light to the myriad of chemical signals used by cells have the potential to greatly expand the applicability of optogenetics. LOV proteins combine a light actuator LOV domain with one or more effector domains and are among the most commonly used optogenetic tools for perturbing intracellular signaling; for example, controlling organelle transport and positioning (142). A bioinformatics survey of the public and 1KP databases (35) identified putative effector domains that had not been previously observed, e.g., GTP cyclohydrolase type II, lipase, and glutamine aminotransferase. Effectors thought to be rare turned out to be common, e.g., regulators of G protein signaling. The most important finding of general relevance to bioprospecting, which is reproduced in **Figure 5**, showed that architectural complexity from this survey of LOV proteins scaled better with evolutionary diversity (as measured by the number of phyla) than with other

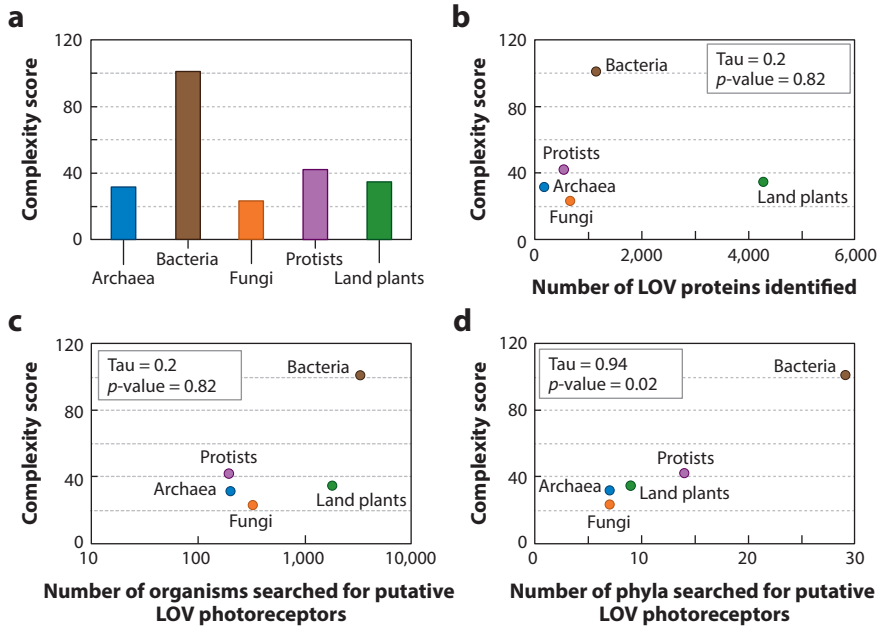


Figure 5

Architectural complexity among LOV domain photoreceptor genes correlates with evolutionary diversity. Figure modified from Reference 35 with permission from the authors. (*a*) Computed complexity quotient for each kingdom quantifies domain architectural complexity as the product of the average number of effector domains per LOV photoreceptor and the total number of different effector types observed. (*b–d*) Complexity quotients for each kingdom plotted against (*b*) the total number of putative LOV sequences identified, (*c*) the total number of organisms searched for LOV sequences, and (*d*) the total number of phyla searched for LOV sequences. Kendall's rank correlation tau coefficients and their accompanying *p*-values are shown on each scatterplot. A strong correlation between the number of phyla searched and the complexity of the resulting LOV photoreceptors suggests that evolutionary diversity is a greater predictor of complexity than sample size. Abbreviation: LOV, light-oxygen-voltage sensing.

metrics of sample size. Not coincidentally, most of the novel architectures were found in algae (and in bacteria). This is consistent with the well-known result from population genetics that microbial species with large population sizes and short generation times tend to be more genetically diverse (43) and is another reason why 1KP tried to maximize the number of phyla sequenced.

Solving Genetic Networks for Complex Traits

One of the landmarks of synthetic biology was the Keasling laboratory's production, in yeast, of a precursor to the antimalarial drug artemisinin (113). Artemisinin was originally isolated from the sweet wormwood, *Artemisia annua*, which had long been used in traditional Chinese medicine. Youyou Tu was awarded the 2015 Nobel Prize in Physiology or Medicine for extracting and characterizing this medicinal compound (132). Inspiring as this story is, it remains a formidable challenge to solve a biosynthetic pathway to the level of detail that is necessary to reconstitute it in a synthetic organism, regardless of how much omics data is generated. For example, over two years of work was required to identify four enzymes that catalyze the first six steps in the biosynthesis of cyclopamine, a promising anticancer medicinal compound derived from the monocot *Veratrum californicum* (2). Most of that effort was devoted to validating leads generated from transcriptomic

and metabolomic data. The process is sequential, i.e., not amenable to parallelization, and scales nonlinearly with the number of steps. It speaks to the inherent difficulty of solving genetic networks for complex traits. More generally, imagine if there were a hundred genes and no simple procedure, e.g., correlations between transcriptomic and metabolomic data, to generate leads for validation. This is the reality for many complex traits, and although there is no general-purpose solution, it is sometimes possible to reduce the noise in comparative genomics data when the trait of interest is convergently evolved. Here, we present an example from C_4 photosynthesis, which was initiated as a subproject of 1KP.

In tropical and subtropical regions, the C_4 pathway represents a more efficient form of photosynthesis than the C_3 pathway from which it emerged. This transition occurred independently at least 65 times within the angiosperms, and it is one of the most remarkable instances of convergent evolution known (118). The differences between the two pathways include alterations to leaf development, cell biology, and biochemistry. More than a hundred genes are likely to be involved. Some of the earlier 1KP studies looked at key enzymes, including *PEPC* (21) and *FtsZ1-FtsZ2-RP5B-PARC6* (131). However, these investigators were members of an international consortium dedicated to creating a C_4 rice (47, 148). A more comprehensive understanding of the pathway was required, and it was unclear if traditional methods, such as gene knockouts, would identify all of the requisite components. Hence, they devised a complementary approach using comparative genomics, generating expression data for 30 C_4 and 17 C_3 species representing 18 independent origins of C_4. This included all seven orders within the eudicots that are known to perform C_4 photosynthesis. The comparisons identified 149 genes with consistently altered abundances, 113 (or 36) of which were more (or less) abundant in all C_4 species when compared to related C_3 species. The list included many genes already known to be differentially expressed in C_4 versus C_3 species. But it also identified four novel metabolic pathways not previously linked to C_4 photosynthesis, including a novel mechanism for concentrating CO_2 (67). Several algorithms were developed to address certain analytic challenges, such as the assessment of transcriptome assemblies in the absence of reference genomes (127) and inference of orthogroups from transcriptome assemblies (31). In the following two paragraphs, we discuss why convergent evolution is a powerful tool for solving complex traits, the implications for creating a C_4 rice, and the underlying nature of convergent evolution for this and possibly other traits.

It is always a statistical challenge to reduce the many tens of thousands of genes that are encoded in a plant genome to the hundred or so genes (or whatever that number might be) that are relevant to a trait of interest. One can always exchange sensitivity for specificity but it is difficult to have both, yet both were achieved using convergent evolution to elucidate the genes for C_4 photosynthesis. Good sensitivity was reflected in the fact that all previously known pathways (and some new ones) were detected through comparative genomics. Good specificity was reflected in the fact that most of the identified genes were plausibly related to C_4 photosynthesis. This is because the species analyzed were evolutionarily diverse, and hence, their gene expression patterns often share little in common other than C_3 versus C_4 status. Importantly, there was no a priori reason to expect that evolutionarily diverse species would have independently selected the same 149 genes some 18 times. In fact, the chances of this happening are astronomically small. More likely in our conception is that a few master regulators were selected. These regulators could take any form, including transcription factor genes, noncoding binding motifs, microRNAs, and so on. A famous example of how just a few regulators can trigger a complex process was the discovery by Yamanaka and colleagues (93, 134) that four transcription factors induce pluripotent stem cells from adult mouse fibroblasts. The more direct evidence for this hypothesis is the observation that the monocot *Eleocharis vivipara* switches between C_3 and C_4 photosynthesis in response to environmental stimuli (140, 141), implying that all necessary genes for both forms of photosynthesis

are present in this one (and presumably other) species, and that a few simple regulatory changes may be all that is needed to create a C_4 rice.

Exaptation: a trait used for a novel function that is distinct from the one for which it originally evolved

Among the 149 genes identified by the 1KP study, four transcription factors were more abundant in all C_4 species (67). While intriguing, direct experimentation is needed to prove that transitions between C_3 and C_4 photosynthesis can be induced by just a few master regulators. If so, this would be another instance of preadaptation, or in less teleologically loaded terms, exaptation (37). Consistent with this hypothesis, molecular alterations defined by 1KP support phenotypic studies indicating that early modifications to leaf anatomy in C_4 plants are associated with processes unlikely to impact photosynthesis and more likely to affect water storage (22, 154). As a mechanism for convergent evolution of complex traits, exaptation is appealingly simple, but we do not yet know how often this might be true. We may get an answer to this question in the near future, as convergent evolution has been reported in other important biological processes, for example, nitrogen fixation by root nodule and plant-cyanobacteria symbioses (25). Convergent evolution studies are already under way at BGI-Shenzhen for both forms of nitrogen fixation by bacterial symbiosis. It would be ironic if, in the process of studying important agricultural traits, we discover something even more profound about the evolution of complex traits.

FUTURE DIRECTIONS

As the first next-generation sequencing project focused on the diversity of a major clade of life, 1KP inspired similar projects for other taxa, e.g., insects, **https://1kite.cngb.org/home**; fish, **https://db.cngb.org/fisht1k/**; and birds, **https://b10k.genomics.cn**. Given the incomplete nature of transcriptomics data, it is unavoidable that some 1KP results, notably the WGD and gene family analyses, await complete genomes to resolve finer points that cannot be resolved by transcriptomics alone. With the continually decreasing cost of sequencing (**https://www.genome.gov/sequencingcosts**), there are now plans to study every known form of life (100), starting with 10,000 plant genomes (20). We hope that such initiatives will fill the many gaps in the currently available plant genomes (**Table 1**). Of note, the most significant gaps include streptophyte algae (particularly *Zygnematophyceae*), almost all of the nonflowering plant lineages, angiosperm lineages outside of the core eudicot and grass clades, and many other lineages within these clades.

Evolutionary insights are the most immediate outcome from the sequencing of phylogenetically diverse species. Any benefits to industry, although not unrealistic in the long term, must navigate a gauntlet of unlikely interpersonal ties. For example, the 1KP optogenetics work was led by protein engineers with no interest in or ability to generate sequence data, let alone do the database searches. They also had no knowledge of or access to many of the phylodiverse species that were sequenced. Conversely, few of the 1KP contributors had ever heard of optogenetics when this project began. Serendipity often plays a major role in such discoveries. Our success was, in large part, a consequence of being given the freedom to develop collaborations that were not even remotely hinted at in our initial funding application. We are forever grateful that we were not subjected to a project management fixed-timetable deliverables approach to grant oversight.

SUMMARY POINTS

1. Sequencing of phylogenetically diverse species has greatly expanded the number of gene sequences for green plants and generated the most comprehensive phylogenomic analyses to date for *Archaeplastida*.

2. Single-copy gene phylogenies are largely concordant with each other and with previous species tree estimates, but elevated gene tree discordance at some nodes in the species tree alludes to periods of rapid radiation.
3. Circumscription and analyses of multicopy gene families have identified 138 whole-genome duplications that were not previously described.
4. The birth and diversification of many gene families contributing to the developmental complexity of land plants occurred in their algal ancestors, substantially earlier than the origin of land plants.
5. Over a billion years of green plant evolution have produced a cornucopia of protein diversity that is proving invaluable for a wide range of pharmaceutical, industrial, and biotech applications.
6. Sequencing of phylogenetically diverse species exhibiting a convergently evolved trait is a statistically powerful method to solve gene regulatory networks.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review article.

ACKNOWLEDGMENTS

1KP was initiated with funding to G.K.S.W. from the Alberta Ministry of Advanced Education, Alberta Innovates (iCORE Strategic Chair), Musea Ventures, and China National GeneBank (CNGB). CNGB founding director Yong Zhang was coleader of 1KP, as was Michael Deyholos at the University of Alberta. We thank the many individuals who provided plant materials. All of the transcriptome sequencing was done by BGI-Shenzhen. Shifeng Cheng created **Figure 4**, in the context of an ongoing collaboration with M.M. and G.K.S.W. This manuscript was greatly improved with the advice of the following people: Martin Porsch, Kristian Ullrich, Ram Samudrala, Brian Y. Chow, Toni Kutchan, Steven Kelly, Matt Stata, and Julian Hibberd.

LITERATURE CITED

1. *Amborella* Genome Proj. 2013. The *Amborella* genome and the evolution of flowering plants. *Science* 342(6165):1241089
2. Augustin MM, Ruzicka DR, Shukla AK, Augustin JM, Starks CM, et al. 2015. Elucidating steroid alkaloid biosynthesis in *Veratrum californicum*: production of verazine in Sf9 cells. *Plant J.* 82(6):991–1003
3. Avise JC, Robinson TJ. 2008. Hemiplasy: a new term in the lexicon of phylogenetics. *Syst. Biol.* 57(3):503–7
4. Baniaga AE, Arrigo N, Barker MS. 2016. The small nuclear genomes of *Selaginella* are associated with a low rate of genome size evolution. *Genome Biol. Evol.* 8:1516–25
5. Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, et al. 2011. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* 332(6032):960–63
6. Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA. 2016. On the relative abundance of autopolyploids and allopolyploids. *New Phytol.* 210(2):391–98
7. Barker MS, Dlugosch KM, Dinh L, Challa RS, Kane NC, et al. 2010. EvoPipes.net: Bioinformatic tools for ecological and evolutionary genomics. *Evol. Bioinform. Online* 6:143–49

8. Barker MS, Husband BC, Pires JC. 2016. Spreading Winge and flying high: the evolutionary importance of polyploidy after a century of study. *Am. J. Bot.* 103(7):1139–45
9. Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, et al. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* 25(11):2445–55
10. Becker B, Marin B. 2009. Streptophyte algae and the origin of embryophytes. *Ann. Bot.* 103(7):999–1004
11. Bennett T, Brockington SF, Rothfels C, Graham SW, Stevenson D, et al. 2014. Paralogous radiations of PIN proteins with multiple origins of noncanonical PIN structure. *Mol. Biol. Evol.* 31(8):2042–60
12. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456(7218):53–59
13. Bevan MW, Uauy C, Wulff BBH, Zhou J, Krasileva K, Clark MD. 2017. Genomic innovation for crop improvement. *Nature* 543(7645):346–54
14. Blank CE. 2013. Origin and early evolution of photosynthetic eukaryotes in freshwater environments: reinterpreting Proterozoic paleobiology and biogeochemical processes in light of trait evolution. *J. Phycol.* 49(6):1040–55
15. Bourque S, Jeandroz S, Grandperret V, Lehotai N, Aimé S, et al. 2016. The evolution of HD2 proteins in green plants. *Trends Plant Sci.* 21(12):1008–16
16. Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422(6930):433–38
17. Boyden ES. 2011. A history of optogenetics: the development of tools for controlling brain circuits with light. *F1000 Biol. Rep.* 3:11
18. Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst. Biol.* 54(5):743–57
19. Bythell-Douglas R, Rothfels CJ, Stevenson D, Graham SW, Wong G-S, et al. 2017. Evolution of strigolactone receptors by gradual neo-functionalization of KAI2 paralogues. *BMC Biol.* 15:52
19a. Carpenter EJ, Matasci N, Ayyampalayam S, Wu S, Sun J, et al. 2019. Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). *Gigascience* 8(10):giz126
20. Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, et al. 2018. 10KP: A phylodiverse genome sequencing plan. *Gigascience* 7:giy013
21. Christin P-A, Arakaki M, Osborne CP, Bräutigam A, Sage RF, et al. 2014. Shared origins of a key enzyme during the evolution of C_4 and CAM metabolism. *J. Exp. Bot.* 65(13):3609–21
22. Christin P-A, Osborne CP, Chatelet DS, Columbus JT, Besnard G, et al. 2013. Anatomical enablers and the evolution of C_4 photosynthesis in grasses. *PNAS* 110(4):1381–86
23. Davidson R, Vachaspati P, Mirarab S, Warnow T. 2015. Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. *BMC Genomics* 16(Suppl. 10):S1
24. Delaux P-M, Radhakrishnan GV, Jayaraman D, Cheema J, Malbreil M, et al. 2015. Algal ancestor of land plants was preadapted for symbiosis. *PNAS* 112(43):13390–95
25. Delaux P-M, Radhakrishnan G, Oldroyd G. 2015. Tracing the evolutionary path to nitrogen-fixing crops. *Curr. Opin. Plant Biol.* 26:95–99
26. Delwiche CF, Cooper ED. 2015. The evolutionary origin of a terrestrial flora. *Curr. Biol.* 25(19):R899–910
27. de Vries J, de Vries S, Slamovits CH, Rose LE, Archibald JM. 2017. How embryophytic is the biosynthesis of phenylpropanoids and their derivatives in streptophyte algae? *Plant Cell Physiol.* 58(5):934–45
28. de Vries J, Stanton A, Archibald JM, Gould SB. 2016. Streptophyte terrestrialization in light of plastid evolution. *Trends Plant Sci.* 21(6):467–76
29. Domozych DS, Popper ZA, Sørensen I. 2016. Charophytes: evolutionary giants and emerging model organisms. *Front. Plant Sci.* 7:1470
30. Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, et al. 2015. The butterfly plant arms-race escalated by gene and genome duplications. *PNAS* 112(27):8362–66
31. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157

32. Estep MC, McKain MR, Vela Diaz D, Zhong J, Hodge JG, et al. 2014. Allopolyploidy, diversification, and the Miocene grassland expansion. *PNAS* 111(42):15149–54
33. Fosque BF, Sun Y, Dana H, Yang C-T, Ohyama T, et al. 2015. Neural circuits. Labeling of active neural circuits in vivo with designed calcium integrators. *Science* 347(6223):755–60
34. Freeling M. 2017. Picking up the ball at the K/Pg boundary: the distribution of ancient polyploidies in the plant phylogenetic tree is a spandrel of asexuality and occasional sex. *Plant Cell* 29(9):202–6
35. Glantz ST, Carpenter EJ, Melkonian M, Gardner KH, Boyden ES, et al. 2016. Functional and topological diversity of LOV domain photoreceptors. *PNAS* 113(11):E1442–51
36. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40(Database Issue):D1178–86
37. Gould SJ, Vrba ES. 1982. Exaptation—a missing term in the science of form. *Paleobiology* 8(1):4–15
38. Govorunova EG, Cunha SR, Sineshchekov OA, Spudich JL. 2016. Anion channelrhodopsins for inhibitory cardiac optogenetics. *Sci. Rep.* 6:33530
39. Govorunova EG, Sineshchekov OA, Janz R, Liu X, Spudich JL. 2015. Natural light-gated anion channels: a family of microbial rhodopsins for advanced optogenetics. *Science* 349(6248):647–50
40. Govorunova EG, Sineshchekov OA, Rodarte EM, Janz R, Morelle O, et al. 2017. The expanding family of natural anion channelrhodopsins reveals large variations in kinetics, conductance, and spectral sensitivity. *Sci. Rep.* 7:43358
41. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29(7):644–52
42. Graham LE, Arancibia-Avila P, Taylor WA, Strother PK, Cook ME. 2012. Aeroterrestrial *Coleochaete* (Streptophyta, Coleochaetales) models early plant adaptation to land. *Am. J. Bot.* 99(1):130–44
43. Graur D, Li WH. 2000. *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer. 2nd ed.
44. Harholt J, Moestrup Ø, Ulvskov P. 2016. Why plants were terrestrial from the beginning. *Trends Plant Sci.* 21(2):96–101
45. Heyduk K, McKain MR, Lalani F, Leebens-Mack J. 2016. Evolution of a CAM anatomy predates the origins of Crassulacean acid metabolism in the Agavoideae (Asparagaceae). *Mol. Phylogenet. Evol.* 105:102–13
46. Heyduk K, Trapnell DW, Barrett CF, Leebens-Mack J. 2015. Estimating relationships within Sabal (Arecaceae) through multilocus analyses of sequence capture data. *Biol. J. Linn. Soc. Lond.* 117:106–20
47. Hibberd JM, Sheehy JE, Langdale JA. 2008. Using C_4 photosynthesis to increase the yield of rice—rationale and feasibility. *Curr. Opin. Plant Biol.* 11(2):228–31
48. Hochbaum DR, Zhao Y, Farhi SL, Klapoetke N, Werley CA, et al. 2014. All-optical electrophysiology in mammalian neurons using engineered microbial rhodopsins. *Nat. Methods* 11(8):825–33
49. Hodel RGJ, Gitzendanner MA, Germain-Aubrey CC, Liu X, Crowl AA, et al. 2016. A new resource for the development of SSR markers: millions of loci from a thousand plant transcriptomes. *Appl. Plant Sci.* 4(6):1600024
50. Holzinger A, Becker B. 2015. Desiccation tolerance in the streptophyte green alga *Klebsormidium*: the role of phytohormones. *Commun. Integr. Biol.* 8(4):e1059978
51. Holzinger A, Kaplan F, Blaas K, Zechmann B, Komsic-Buchmann K, Becker B. 2014. Transcriptomics of desiccation tolerance in the streptophyte green alga *Klebsormidium* reveal a land plant-like defense reaction. *PLOS ONE* 9(10):e110630
52. Holzinger A, Pichrtová M. 2016. Abiotic stress tolerance of charophyte green algae: new challenges for omics techniques. *Front. Plant Sci.* 7:678
53. Honaas LA, Wafula EK, Wickett NJ, Der JP, Zhang Y, et al. 2016. Selecting superior *de novo* transcriptome assemblies: lessons learned by leveraging the best plant genome. *PLOS ONE* 11(1):e0146062
54. Hori K, Maruyama F, Fujisawa T, Togashi T, Yamamoto N, et al. 2014. *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nat. Commun.* 5:3978
55. Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11(2):97–108
56. Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *PNAS* 104(49):19369–74

57. Jeandroz S, Wipf D, Stuehr DJ, Lamattina L, Melkonian M, et al. 2016. Occurrence, structure, and evolution of nitric oxide synthase–like proteins in the plant kingdom. *Sci. Signal.* 9(417):re2
58. Jia Q, Li G, Köllner TG, Fu J, Chen X, et al. 2016. Microbial-type terpene synthase genes occur widely in nonseed land plants, but not in seed plants. *PNAS* 113(43):12328–33
59. Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* 13(1):R3
60. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473(7345):97–100
61. Johnson KL, Cassin AM, Lonsdale A, Bacic A, Doblin MS, Schultz CJ. 2017. Pipeline to identify hydroxyproline-rich glycoproteins. *Plant Physiol.* 175(2):886–903
62. Johnson KL, Cassin AM, Lonsdale A, Wong GK-S, Soltis DE, et al. 2017. Insights into the evolution of hydroxyproline-rich glycoproteins from 1000 plant transcriptomes. *Plant Physiol.* 174(2):904–21
63. Johnson MTJ, Carpenter EJ, Tian Z, Bruskiewich R, Burris JN, et al. 2012. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PLOS ONE* 7(11):e50226
64. Jordon-Thaden IE, Chanderbali AS, Gitzendanner MA, Soltis DE. 2015. Modified CTAB and TRIzol protocols improve RNA extraction from chemically complex Embryophyta. *Appl. Plant Sci.* 3(5):1400105
65. Kapraun DF. 2007. Nuclear DNA content estimates in green algal lineages: Chlorophyta and Streptophyta. *Ann. Bot.* 99(4):677–701
66. Kellogg EA. 2016. Has the connection between polyploidy and diversification actually been tested? *Curr. Opin. Plant Biol.* 30:25–32
67. Kelly S, Covshoff S, Wanchana S, Thakur V, Quick WP, et al. 2017. Wide sampling of natural diversity identifies novel molecular signatures of C_4 photosynthesis. bioRxiv. **https://doi.org/10.1101/163097**
68. Klapoetke NC, Murata Y, Kim SS, Pulver SR, Birdsey-Benson A, et al. 2014. Independent optical excitation of distinct neural populations. *Nat. Methods* 11(3):338–46
69. Koenig D, Weigel D. 2015. Beyond the thale: comparative genomics and genetics of *Arabidopsis* relatives. *Nat. Rev. Genet.* 16(5):285–98
70. Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56(1):17–24
71. Lazebnik Y. 2002. Can a biologist fix a radio?—Or, what I learned while studying apoptosis. *Cancer Cell* 2(3):179–82
72. Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, et al. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. *Mol. Biol. Evol.* 22(10):1948–63
73. Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, et al. (One Thousand Plant Transcriptomes Initiat.). 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574:679–85
74. Leitch IJ, Leitch AR. 2013. Genome size diversity and evolution in land plants. In *Plant Genome Diversity*, Vol. 2, ed. IJ Leitch, J Greilhuber, J Dolezel, JF Wendel, pp. 307–22. Vienna: Springer
75. Leliaert F, Verbruggen H, Vanormelingen P, Steen F, López-Bautista JM, et al. 2014. DNA-based species delimitation in algae. *Eur. J. Phycol.* 49(2):179–96
76. Li F-W, Melkonian M, Rothfels CJ, Villarreal JC, Stevenson DW, et al. 2015. Phytochrome diversity in green plants and the origin of canonical plant phytochromes. *Nat. Commun.* 6:7852
77. Li F-W, Rothfels CJ, Melkonian M, Villarreal JC, Stevenson DW, et al. 2015. The origin and evolution of phototropins. *Front. Plant Sci.* 6:637
78. Li F-W, Villarreal JC, Kelly S, Rothfels CJ, Melkonian M, et al. 2014. Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *PNAS* 111(18):6672–77
79. Li R, Fan W, Tian G, Zhu H, He L, et al. 2010. The sequence and *de novo* assembly of the giant panda genome. *Nature* 463(7279):311–17
80. Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, et al. 2015. Early genome duplications in conifers and other seed plants. *Sci. Adv.* 1(10):e1501084
81. Maddison WP. 1997. Gene trees in species trees. *Syst. Biol.* 46(3):523

82. Maugarny-Calès A, Gonçalves B, Jouannic S, Melkonian M, Wong GK-S, Laufs P. 2016. Apparition of the NAC transcription factors predates the emergence of land plants. *Mol. Plant* 9(9):1345–48
83. Mayrose I, Zhan SH, Rothfels CJ, Arrigo N, Barker MS, et al. 2015. Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al. (2014). *New Phytol.* 206(1):27–35
84. Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, et al. 2011. Recently formed polyploid plants diversify at lower rates. *Science* 333(6047):1257
85. Mikkelsen MD, Harholt J, Ulvskov P, Johansen IE, Fangel JU, et al. 2014. Evidence for land plant cell wall biosynthetic mechanisms in charophyte green algae. *Ann. Bot.* 114(6):1217–36
86. Mirarab S. 2017. Phylogenomics: constrained gene tree inference. *Nat. Ecol. Evol.* 1:56
87. Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31(12):i44–52
88. Molloy EK, Warnow T. 2018. To include or not to include: the impact of gene filtering on species tree estimation. *Methods Syst. Biol.* 67(2):285–303
89. Muller HJ. 1925. Why polyploidy is rarer in animals than in plants. *Am. Nat.* 59:346–53
90. Nagata T, Hosaka-Sasaki A, Kikuchi S. 2015. The evolutionary diversification of genes that encode transcription factor proteins in plants. In *Plant Transcription Factors*, ed. D Gonzalez, pp. 73–97. Boston: Academic
91. Nelson D, Werck-Reichhart D. 2011. A P450-centric view of plant evolution. *Plant J.* 66(1):194–211
92. Newman DJ, Cragg GM. 2012. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* 75(3):311–35
93. Okita K, Ichisaka T, Yamanaka S. 2007. Generation of germline-competent induced pluripotent stem cells. *Nature* 448(7151):313–17
94. Orr HA. 1990. "Why polyploidy is rarer in animals than in plants" revisited. *Am. Nat.* 136(6):759–70
95. Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu. Rev. Genet.* 34(1):401–37
96. Pabón-Mora N, Wong GK-S, Ambrose BA. 2014. Evolution of fruit development genes in flowering plants. *Front. Plant Sci.* 5:300
97. Page RD, Charleston MA. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.* 7(2):231–40
98. Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5(5):568–83
99. Parks M, Cronn R, Liston A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* 7:84
100. Pennisi E. 2017. Biologists propose to sequence the DNA of all life on Earth. *Science*, Feb. 24. **http://www.sciencemag.org/news/2017/02/biologists-propose-sequence-dna-all-life-earth**
101. Pfannebecker KC, Lange M, Rupp O, Becker A. 2017. An evolutionary framework for carpel developmental control genes. *Mol. Biol. Evol.* 34(2):330–48
102. Pfannebecker KC, Lange M, Rupp O, Becker A. 2017. Seed plant-specific gene lineages involved in carpel development. *Mol. Biol. Evol.* 34(4):925–42
103. Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* 19(8):706–12
104. Philippe H, Laurent J. 1998. How good are deep phylogenetic trees? *Curr. Opin. Genet. Dev.* 8(6):616–23
105. Pimentel D, Donlea JM, Talbot CB, Song SM, Thurston AJF, Miesenböck G. 2016. Operation of a homeostatic sleep switch. *Nature* 536(7616):333–37
106. Ponce-Toledo RI, Deschamps P, López-García P, Zivanovic Y, Benzerara K, Moreira D. 2017. An early-branching freshwater cyanobacterium at the origin of plastids. *Curr. Biol.* 27(3):386–91
107. Potato Genome Seq. Consort., Xu X, Pan S, Cheng S, Zhang B, et al. 2011. Genome sequence and analysis of the tuber crop potato. *Nature* 475(7355):189–95
108. Qiu Y-L, Li L, Wang B, Chen Z, Knoop V, et al. 2006. The deepest divergences in land plants inferred from phylogenomic evidence. *PNAS* 103(42):15511–16
109. Ramirez S, Liu X, Lin P-A, Suh J, Pignatelli M, et al. 2013. Creating a false memory in the hippocampus. *Science* 341(6144):387–91
110. Rannala B, Yang Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. *Syst. Biol.* 66(5):823–42

111. Rensing SA. 2017. Why we need more non-seed plant models. *New Phytol*. 216(2):355–60
112. Richards CL, Hanzawa Y, Katari MS, Ehrenreich IM, Engelmann KE, Purugganan MD. 2009. Perspectives on ecological and evolutionary systems biology. In *Annual Plant Reviews, Volume 35: Plant Systems Biology*, ed. G Coruzzi, R Gutiérrez, pp. 331–49. Oxford, UK: Wiley-Blackwell
113. Ro D-K, Paradise EM, Ouellet M, Fisher KJ, Newman KL, et al. 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440(7086):940–43
114. Robertson G, Schein J, Chiu R, Corbett R, Field M, et al. 2010. *De novo* assembly and analysis of RNA-seq data. *Nat. Methods* 7(11):909–12
115. Roch S, Steel M. 2014. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor. Popul. Biol.* 100C:56–62
116. Romanov RE, Bulionkova TS. 2016. Research note: observations on a terrestrial charophyte in a temperate environment. *Phycol. Res.* 64(2):118–20
117. Rothfels CJ, Li F-W, Sigel EM, Huiet L, Larsson A, et al. 2015. The evolutionary history of ferns inferred from 25 low-copy nuclear genes. *Am. J. Bot.* 102(7):1089–107
118. Sage RF. 2017. A portrait of the C_4 photosynthetic family on the 50th anniversary of its discovery: species number, evolutionary lineages, and Hall of Fame. *J. Exp. Bot.* 68(2):4039–56
119. Salman-Minkov A, Sabath N, Mayrose I. 2016. Whole-genome duplication as a key factor in crop domestication. *Nat. Plants* 2:16115
120. Sayou C, Monniaux M, Nanao MH, Moyroud E, Brockington SF, et al. 2014. A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. *Science* 343(6171):645–48
121. Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* 33(7):1654–68
122. Sayyari E, Whitfield JB, Mirarab S. 2017. Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. *Mol. Biol. Evol.* 34(12):3279–91
123. Scarpino SV, Levin DA, Meyers LA. 2014. Polyploid formation shapes flowering plant diversity. *Am. Nat.* 184(4):456–65
124. Selmecki AM, Maruvka YE, Richmond PA, Guillet M, Shoresh N, et al. 2015. Polyploidy can drive rapid adaptation in yeast. *Nature* 519:349–52
125. Sessa EB, Banks JA, Barker MS, Der JP, Duffy AM, et al. 2014. Between two fern genomes. *GigaScience* 3:15
126. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, et al. 2011. The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* 43(2):109–16
127. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* 26(8):1134–44
128. Soltis DE, Segovia-Salcedo MC, Jordon-Thaden I, Majure L, Miles NM, et al. 2014. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). *New Phytol.* 202(4):1105–17
129. Soltis DE, Visger CJ, Marchant DB, Soltis PS. 2016. Polyploidy: pitfalls and paths to a paradigm. *Am. J. Bot.* 103:1146–66
130. Sørensen I, Pettolino FA, Bacic A, Ralph J, Lu F, et al. 2011. The charophycean green algae provide insights into the early origins of plant cell walls. *Plant J.* 68(2):201–11
131. Stata M, Sage TL, Hoffmann N, Covshoff S, Wong GK-S, Sage RF. 2016. Mesophyll chloroplast investment in C_3, C_4 and C_2 species of the genus *Flaveria*. *Plant Cell Physiol.* 57(5):904–18
132. Su X-Z, Miller LH. 2015. The discovery of artemisinin and the Nobel Prize in Physiology or Medicine. *Sci. China Life Sci.* 58(11):1175–79
133. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In *Molecular Systematics*, ed. DM Hillis, BK Mable, C Moritz, pp. 407–514. Sunderland, MA: Sinauer. 2nd ed.
134. Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126(4):663–76
135. Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, et al. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol.* 207(2):454–67

136. Timme RE, Bachvaroff TR, Delwiche CF. 2012. Broad phylogenomic sampling and the sister lineage of land plants. *PLOS ONE* 7(1):e29696
137. Tirichine L, Bowler C. 2011. Decoding algal genomes: tracing back the history of photosynthetic life on Earth. *Plant J.* 66(1):45–57
138. Tomato Genome Consort., Sato S, Tabata S, Hirakawa H, Asamizu E, et al. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485(7400):635–41
139. Turmel M, Otis C, Lemieux C. 2006. The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. *Mol. Biol. Evol.* 23(6):1324–38
140. Ueno O. 1998. Induction of Kranz anatomy and C_4-like biochemical characteristics in a submerged amphibious plant by abscisic acid. *Plant Cell* 10(4):571–84
141. Ueno O, Samejima M, Muto S, Miyachi S. 1988. Photosynthetic characteristics of an amphibious plant, *Eleocharis vivipara*: expression of C_4 and C_3 modes in contrasting environments. *PNAS* 85(18):6733–37
142. van Bergeijk P, Adrian M, Hoogenraad CC, Kapitein LC. 2015. Optogenetic control of organelle transport and positioning. *Nature* 518(7537):111–14
143. Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18:411–24
144. Van de Poel B, Cooper ED, Van Der Straeten D, Chang C, Delwiche CF. 2016. Transcriptome profiling of the green alga *Spirogyra pratensis* (Charophyta) suggests an ancestral role for ethylene in cell wall metabolism, photosynthesis, and abiotic stress responses. *Plant Physiol.* 172(1):533–45
145. Vasco A, Smalls TL, Graham SW, Cooper ED, Wong GK-S, et al. 2016. Challenging the paradigms of leaf evolution: Class III HD-Zips in ferns and lycophytes. *New Phytol.* 212(3):745–58
146. Vekemans D, Proost S, Vanneste K, Coenen H, Viaene T, et al. 2012. Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. *Mol. Biol. Evol.* 29(12):3793–806
147. Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* 290(5499):2114–17
148. von Caemmerer S, Paul Quick W, Furbank RT. 2012. The development of C_4 rice: current progress and future challenges. *Science* 336(6089):1671–72
149. Wang J, Wang W, Li R, Li Y, Tian G, et al. 2008. The diploid genome sequence of an Asian individual. *Nature* 456(7218):60–65
150. Wang X, Wang H, Wang J, Sun R, Wu J, et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43(10):1035–39
151. Wendel JF. 2015. The wondrous cycles of polyploidy in plants. *Am. J. Bot.* 102(11):1753–56
152. Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *PNAS* 111(45):E4859–68
153. Wietek J, Wiegert JS, Adeishvili N, Schneider F, Watanabe H, et al. 2014. Conversion of channelrhodopsin into a light-gated chloride channel. *Science* 344(6182):409–12
154. Williams BP, Johnston IG, Covshoff S, Hibberd JM. 2013. Phenotypic landscape inference reveals multiple evolutionary paths to C_4 photosynthesis. *eLife* 2:e00961
155. Wodniok S, Brinkmann H, Glöckner G, Heidel AJ, Philippe H, et al. 2011. Origin of land plants: Do conjugating green algae hold the key? *BMC Evol. Biol.* 11:104
156. Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *PNAS* 106(33):13875–79
157. Xi Z, Liu L, Davis CC. 2016. The impact of missing data on species tree estimation. *Mol. Biol. Evol.* 33(3):838–60
158. Xie Y, Wu G, Tang J, Luo R, Patterson J, et al. 2014. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30(12):1660–66
159. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, et al. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* 42(Database Issue):D643–48
160. Zeng L, Zhang N, Zhang Q, Endress PK, Huang J, Ma H. 2017. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytol.* 214(3):1338–54