Hindawi Publishing Corporation
Journal of Botany
Volume 2011, Article ID 646198, 10 pages
doi:10.1155/2011/646198

Review Article

The Genomes of All Angiosperms: A Call for a Coordinated Global Census

David W. Galbraith,[1] Jeffrey L. Bennetzen,[2] Elizabeth A. Kellogg,[3] J. Chris Pires,[4] and Pamela S. Soltis[5]

[1] *School of Plant Sciences and Bio5 Institute, The University of Arizona, Tucson, AZ 85721, USA*
[2] *Department of Genetics, University of Georgia, Athens, GA 30602, USA*
[3] *Department of Biology, University of Missouri-St Louis, St Louis, MO 63121-4400, USA*
[4] *Division of Biological Sciences, University of Missouri, Columbia, MO 65211-7400, USA*
[5] *Florida Museum of Natural History and The Genetics Institute, University of Florida, Gainesville, FL 32611, USA*

Correspondence should be addressed to David W. Galbraith, galbraith@arizona.edu

Received 1 June 2011; Accepted 22 August 2011

Academic Editor: Andrew Wood

Recent advances in biological instrumentation and associated experimental technologies now permit an unprecedented efficiency and scale for the acquisition of genomic data, at ever-decreasing costs. Further advances, with accompanying decreases in cost, are expected in the very near term. It now becomes appropriate to discuss the best uses of these technologies in the context of the angiosperms. This white paper proposes a complete genomic census of the approximately 500,000 species of flowering plants, outlines the goals of this census and their value, and provides a road map towards achieving these goals in a timely manner.

1. Introduction

The angiosperms (flowering plants) are believed to comprise somewhere around 360,000 species [1, 2] in 415 families [3]. An exact enumeration of the total number of existing species is impossible, because this includes a guess at the number of species that have not yet been discovered. If these were all found, a final total of somewhere between 400,000 and 500,000 species seems likely [4]. The enumeration and description of all angiosperm species takes on particular urgency because thousands of species become extinct every year, largely due to the impact of human activities. The risks accompanying the depletion of angiosperm biodiversity resources relate to ecosystem maintenance, to food production, for example, by the loss of wild relatives of crops that may have desirable alleles, and to drug discovery, given the observation that plant secondary products represent not simply the primary source of drugs and drug leads, but also the chemical inspiration for synthetic organic syntheses. These observations lead us to propose the timely establishment of a global network of coordinated research activities across the angiosperms with the objectives listed below.

2. Objectives

The main objective is to establish vibrant and efficient communication among scientific experts in the disparate disciplines of taxonomy, systematics, cytometry, genomics, and bioinformatics, with the aim of planning a comprehensive molecular census of the angiosperms. This census is envisaged to start with measurement of nuclear genome sizes (*C*-value) for all angiosperms and would progress, over a five-year period, through sample survey sequencing to complete genome sequencing for a subset of these species. The census would start with planning activities preparative to coordinated research. Such planning would involve face-to-face meetings, the establishment of virtual communications, training courses and workshops, support for student internships, and support for postdoctoral fellows, graduate students, and junior faculty for cross-disciplinary visits to other participating laboratories and institutions. It would also include the development and publication of white papers covering technical aspects of the planned research as well as identifying Grand Challenge questions that require coordinated funding. The planning process needs to be

sensitive to the international aspects necessarily associated with a global census, including issues of ownership, of specimen import/export, of ethics, and of cultural differences. The planning process should provide intelligent and workable solutions to issues that are identified, and distribute these freely to the general scientific community for adoption. The census would then proceed to implementation, in which funding would be obtained for the three data-intensive stages of the project.

3. Rationale and Justification

3.1. What Is the Value of a Complete Molecular Census of the Angiosperms? The rationale underlying a complete global molecular census is primarily and simply that it would provide an accounting of existing plant biodiversity generated by millions of years of evolution, using the most advanced and cost-effective means available. The angiosperms, or flowering plants, emerged approximately 140 million years ago, their sudden appearance and rapid diversification being described by Darwin as "an abominable mystery" [5]. The ease with which angiosperms became the dominant group, displacing ferns, cycadophytes, and conifers, excites considerable interest, as does the dramatic evolution of angiosperm-derived characteristics critical for animal evolution and diversification.

As the primary autotrophs on land, angiosperms represent an invaluable resource of genetic and genomic information. It is therefore surprising, and somewhat alarming, how little genetic and genomic information has been collected for the angiosperms. In part, this has been due to cost. Historically, information of the type that we propose be gathered—cytological data and DNA sequences—has been obtained using expensive and delicate equipment, with high costs of operation.

At present, biases in genomic information for angiosperms reflect the interests of society: Dirzo and Raven [6] noted that the ~2,500 species of crop plants (0.5% of all angiosperms, based on Bramwell's estimate [4]) are distributed in only about a third of the 415 families of flowering plants. The grass and legume families are most represented, making up about one-third of all crops. Ten additional families (Amaranthaceae, Apiaceae, Arecaceae, Asteraceae, Brassicaceae, Lamiaceae, Rosaceae, Rutaceae, Solanaceae, and Zingiberaceae) are represented by a few to many dozens of crop species whereas the numerous other families contribute only one or a few crop species. Of the 2,500 crop species, 103 supply over 90% of the calories for human consumption, with rice, wheat, and maize (all Poaceae) making up over 60% of this total. For this reason alone, collection of cytological, molecular, and genomic data has naturally focused on crops. It should be noted that additional species are cultivated as sources of fiber, as ornamentals, or as sources of medicines. Dirzo and Raven [6] estimate that a further 25,000 plant species, are, or have been, used as sources of medicines, but most of these are not cultivated, with materials being collected directly from nature. Therefore, this group of plants, despite their importance, is not well described at the molecular or cytological level.

The justification for a comprehensive molecular census of the angiosperms is easy to articulate: firstly, because we lack most molecular information from the great majority of plant species, any attempt to acquire this information will provide value. Secondly, it is clear that increasing pressures on the environment because of the largely uncontrolled growth of human activity is increasingly resulting in loss of biodiversity [7]. The rate of extinction of species has been likened in severity to the five mass extinction events over the last 600 million years in which 65–95% of all species disappeared [8]. In the present case, this is due to a combination of anthropogenic climate change, superimposed on the clearing of lands for agricultural and industrial purposes. The rates of species extinction are debatable, given caveats, amongst other things, as to the definition of a species, and as to the method of estimating extinction. Figures of 17,500 species extinctions per year are accepted by some, as are estimates for the time to extinction of 50% of birds (1,500 years) or mammals (6,500 years; all data from Stork, 2010). Barnosky et al. [8] compared extinction rates of mammals over the last 500 years to those of mammals in the fossil record. They found the current rate exceeds that associated with all five previous mass extinctions, and coupled to data from other orders of life forms, predicts extinction of 75% of amphibian, mammalian, and avian species within ~200–500 years. Hence, there is broad consensus amongst scientists that an extinction comparable in magnitude to the previous five mass extinctions is indeed currently underway, in agreement with the prediction of a 1998 survey by the American Museum of Natural History [9].

Putting these two observations together, it is obvious we risk extinction of many (perhaps most) plant species *before they are described in any detail*. At the molecular and genomic level, this means we risk the nondiscovery of DNA sequences that encode gene products having useful properties, ranging, for example, from genes that might encode resistance to biotic or abiotic stresses, to metabolic pathways that produce compounds of potential pharmacological value. At the level of basic biology, we risk missing the observation of interesting and novel forms of eukaryotic plant life styles, given that our current understanding of plants is predominantly shaped by study of a very limited number of agricultural crops and model species.

If we accept that it would be valuable to derive a comprehensive molecular census of the angiosperms, it is reasonable to move on to examine the value of each of the census components. Within the context of this article and this volume, these components are the *C*-value, sample survey sequencing, and complete genome sequencing.

3.2. What is the Value of the Information Represented by the Individual Components of the Census?

3.2.1. C-Value

Rationale. The *C*-value of eukaryotic species is a fundamental parameter that can be measured, given the presence of

a nucleus as a defining characteristic of eukaryotic cells. Differences in *C*-values have impacts at multiple biological levels within eukaryotes, with effects on genome replication time and mechanism, effects on scaling associated with nuclear volume and surface area relative to other cellular components and to overall cell size, effects on numbers of chromosomes and the occurrence and regulation of endo- and polyploidy [10, 11], effects on the copy numbers of genes, transposable elements, and on other structural components of chromosomes, and effects relating to genetic and epigenetic transmission of information. Direct measurement of genome size by flow cytometry provides a useful reality check on genome sizes estimated from full sequence assemblies, and, as a general rule, due to misassembly or absence from these assemblies of repetitive elements, the true genome size is larger than these assemblies [12].

Many research groups have an ongoing interest in studying the processes responsible for genome structure variation in multicellular plants, and the influence of this variation on gene and chromosome function. It is clear that nuclear genomes in angiosperms are highly dynamic, particularly as reflected in the amplification of transposable elements (TEs), in the removal of selectively neutral DNA sequences, in the occurrence of polyploidy, in recombinational rearrangement by ectopic events, and the resolution of double-strand DNA breaks [13–20]. Dramatic differences have been observed in the rates of these processes in different lineages [21], but too few lineages have been analyzed to define when these changes occur, and to associate these changes with alterations in the environment or in the evolved fitness for that environment. Other workers have examined correlations between nuclear DNA content and histone H3 methylation, between rDNA copy number and genome size, and between genome size and recombination rate (reviewed in [12]).

At the cellular and organismal levels, analysis of *C*-values has been used for a number of large comparative studies [12]. These recently have included the relationship between genome size and B-chromosomes, duration of the cell division cycle, seed size and mass, photosynthetic rate, plant growth form and distribution, leaf cell size and stomatal density, and patterns of genome size evolution (see [12, 22, 23] for a full discussion).

Differences in *C*-value have practical implications as well, notably for identification of species that can be most efficiently sequenced (i.e., have the smallest genomes), identification of unsuspected polyploid series, identification of rare hybrid species, and identification of the best plant lineages to investigate recent changes in DNA removal rate, TE amplification, or ploidy. Genome size determinations also indicate the best lineages for sequence analysis that can provide the broadest range of information about plant gene diversity (content and function), by allowing choices of appropriate size genomes from a phylogenetically informative set of species. In agriculture, nuclear DNA content measurements allow rapid identification of novel hybrids, detection of desired or undesired ploidy classes within certified seed, efficient screening for haploids and doubled haploids, and rogueing of aneuploids emerging from genetic engineering and tissue culture [24].

Justification. As of now, genome sizes have been measured and compiled for 8,959 species [12] of which 6,287 have been awarded "prime" status, meaning that their values are strongly supported by curated experimental results. The latter number represents 1-2% of angiosperms, depending on whether one includes as yet undiscovered species. Leadership in creating a repository of this information has been taken by Drs. Bennett and Leitch at the Royal Botanic Gardens, Kew, in the form of the Plant DNA *C*-values Database (release 5.0, December 2010; [25]). This database provides a number of searchable functions and modes, returning output fields chosen from Family, Genus, Species, Authority, Chromosome number ($2n$), Ploidy level (x), Estimation method, Voucher, *C*-value, and Original Reference. One can also search for species having genome sizes falling within specific defined ranges.

Most *C*-value measurements included in the database, and essentially all that are done now, employ flow cytometry to estimate *C*-value, following methods pioneered some years ago [26]. This involves simple chopping of plant samples, filtering the released nuclei to remove large debris, and adding a DNA-specific fluorochrome. The fluorescence emission of the nuclei is then measured by flow cytometry, and compared to that of standards [26–28]. Although flow cytometry is convenient and rapid (up to 10 samples per hour can be readily processed on one instrument), and the required chemicals are inexpensive, the low level of overall sampling of angiosperm *C*-values achieved to date is, in part, due to the historically high cost of the flow cytometers themselves. The instruments have also typically been physically large, difficult and expensive to maintain, and have required highly-trained technicians as operators. Some examples of low-cost cytometers have been available over the last 25 years (Partec) having a small laboratory footprint, and minimal energy and fluidic requirements. Other manufacturers have recently developed low-cost instruments, including Accuri, Millipore, and Life Technologies [29]. Three further features of these instruments are relevant as follows.

Dynamic Range and Linearity. The Accuri C6, costing around $35,000, has an extraordinarily large dynamic range of measurement. This dynamic range easily accommodates DNA content values spanning 0.32–80.00 pg in a single instrument run, and analysis of noise and upper limits indicates the C6 should be able to handle the largest and smallest angiosperm values yet reported [30]. For all samples, optical linearity can also be validated from the accumulated data, such as appropriate positioning of 4*C versus* 2*C* nuclei at both ends of the scale. Further, working with fixed settings offers significant gains in robust processing of data between days, and across users and laboratories.

Automated Throughput. Low-cost plate samplers are now available. A general criticism of much current *C*-value data from botanical studies is their inadequate sampling: few replicates, and few populations, in particular. The published value needs to convey a context, implying more numbers and more discriminate sampling, which is evidently facilitated by automation.

Portability. A follow-on is that part of these analyses would be more effectively done in the field, such that immediate results guide subsequent sampling. The Partec CyFlow miniPOC instrument, designed for CD4 and CD4% testing in developing countries but, based on its technical specifications, clearly adaptable for genome size measurements, is particularly notable, since it weighs less than 5 kg, operates on a 12-volt battery supply, costs less than $20,000, and, being equipped with USB ports, can be readily interfaced with wireless communication devices.

An additional impediment to large-scale *C*-value measurement has been the typical requirement of fresh tissue for flow cytometric analysis, as originally described [26]. However, modifications to standard techniques now allow the use of dried leaf tissue [31, 32]—an innovation that may dramatically increase the rate of *C*-value measurement, particularly if the approach can be altered to accommodate the vast botanical resources housed in the world's herbaria.

Based on these advances, it is now conceivable, in terms of affordability, methodology, and practicality, to consider extending *C*-value measurements to all angiosperms. The general concept would be to distribute significant numbers (20–30) of these instruments across the globe, placing them in appropriate locations, such as botanical gardens, seed repositories, and academic institutions, and then operate them continuously to analyze samples collected at those particular locations. Our aim, therefore, is to develop an optimal design to achieve this goal; this will require identifying and addressing issues associated with sample collection and identification, setting up standard practices for sample processing, and devising schemas for solving any technical problems that might be encountered. An ancillary, yet crucial, aim is to facilitate the process of obtaining funds to support data collection. For both aims, a key is the establishment of collaborations, as has been done for animal genomics [33]. In recent years, the number of groups undertaking *C*-value measurements around the world has steadily increased, and scientists are beginning to recognize the potential of medium- and large-scale surveys. A number of reports are emerging that address *C*-values of species within specific geographical areas (see [12] for a discussion), and participants in these projects are already highly collaborative.

It is, finally, worth emphasizing quality control of *C*-value estimates. The concept of a "gold standard" has already been mentioned in terms of the values provided as prime estimates in the Kew *C*-value database. However, in the scientific community beyond cytologists, evolutionary biologists, population biologists, and systematists, there has been a naïve acceptance of very narrow data generation and analysis in the *C*-value field. For instance, analyses that look at only one individual within a species and with only one *C*-value determination for that sample/species are often judged acceptable. Analysis of multiple individuals and information on ploidy, populations, vouchers, and cytogenetics should be provided and required for definitive assessments. In considering a census of largely unexplored species, it must be underlined that a 2*C*-value gains its pertinence from its context and representativity. While technical quality control is simple to develop and ensure, the conceptual setting requires scientific maturation. We need to avoid the situation in which we are attempting to integrate often dubious lists of *C*-values (albeit easy to use and therefore readily appreciated) with laborious geobotanical and evolutionary studies. We also need to develop means to continue to identify and, to the greatest extent possible, eliminate inaccurate or incorrect values from the *C*-value literature.

3.2.2. Sample Survey Sequencing

Rationale. The angiosperms are unusual, compared to most eukaryotic clades, in that their genome sizes span an extraordinary range. The smallest reported angiosperm genome size, that of *Genlisea margaretae* (Lentibulariaceae), corresponds to a 2*C*-value of 0.129 pg [34]. The largest currently measured is that of *Paris japonica* (Melanthiaceae) with a 2*C*-value of 304.46 pg [12], representing a genome of ~150 billion base pairs. Because angiosperms contain somewhere between 25,000 and 75,000 genes (see, e.g., http://chibba.agtec.uga.edu/duplication/), the immediate question is why some angiosperms have 2,400-fold more nuclear DNA than others in a life style that is evolutionarily viable. Further questions concern the relationship between chromosome number and *C*-value; an extreme example is that of *Sedum suaveolens* with $2n = \sim 640$ and yet only 18.3 pg of DNA per nucleus [35].

In part, the first question is answered by the established occurrence and dynamic amplification and loss of repetitive DNA sequences within the genome (see, for example, [15, 36, 37]). However, insufficient samples exist to draw meaningful conclusions. Certainly the same is true for investigating the second question. This aspect of the census, therefore, aims to provide an in-depth analysis of the correlates of *C*-value diversity, and an understanding of its biological consequences. This would be achieved by obtaining more information, across carefully selected species, of the patterns of repetitive genomic sequences within the genomes. For this type of survey, very low sequence coverage is needed, but it nevertheless can be guided by *C*-value information obtained as previously described.

Justification. Dramatic decreases in costs of sequencing have accompanied the emergence of new sequencing technologies [38]. Colloquially termed next-generation (NextGen) sequencing, these technologies rely on highly parallel implementation of sequencing reactions, at specific locations on slides or in microwells, which greatly reduces the cost per reaction as compared to capillary-based (first-generation) sequencing. A penalty is paid in terms of read length, although incremental improvements in this parameter continue to be achieved. However, the disadvantages of short read lengths are greatly offset by dramatic increases in sequence output. For example, a single sequencing center, BGI Shenzhen, now has the capacity to produce about 75 terabytes of sequence data per day, approximately equal to 1,500 human genomes. Indeed, the extraordinary increase in global sequencing capacity,

coupled with plunging costs, has spawned a number of large-scale collaborative projects, including the 1,000 human genomes project (http://www.1000genomes.org/), the 10,000 vertebrate genomes project (http://www.genome10k.org/), the 1,001 *Arabidopsis* project (http://1001genomes.org/), the 1,000 plant transcriptomes project (http://www.onekp.com/ http://www.onekp.com/), and various metagenomics surveys (see, for example, The International Human Microbiome Consortium: http://www.human-microbiome.org/).

Massively parallel sequencing has revolutionized molecular biology by making genomic sequencing possible for many more organisms than previously attainable. Low redundancy and shallow coverage genome survey sequences from this type of sequencing have the potential to rapidly provide large, cost-effective datasets for phylogenetic inference, to replace single gene or spacer regions as DNA barcodes, and to provide a plethora of data for other comparative molecular evolution studies [39–41].

Coupling low coverage (1–5X) sequencing to methods of DNA indexing, to allow sample multiplexing, provides low-cost surveys of genomes assuming one has an idea of the genome size. Repetitive sequences predominate in the results, and tools have been provided for their specific analysis in this type of data [39–42], so this provides a rapid means to chart the occurrence, amplification and contraction, and phylogenetic distributions of these interesting genomic elements. The role of coordination in these activities is to provide an international framework that can most efficiently define those angiosperm species to be sequenced.

3.2.3. Complete Genome Sequencing

Rationale. The angiosperms are also unusual in terms of the widespread occurrence of ancestral genome duplication events [43]. For example, even a genome as small as that of *Arabidopsis thaliana* contains clear evidence of several large-scale duplication events, although their precise timing remains contentious [44]. Angiosperms readily accommodate polyploidy, and this mechanism provides the greatest proportion of duplicated genes [45]. At the same time, tandem duplications are an important contributor to the evolution of agriculturally and biomedically important traits, such as disease resistance and the biosynthesis of secondary products. Full genomic sequencing, hand-in-hand with transcriptome analysis, would greatly augment the molecular analyses at the heart of this census, since it would enumerate all sequences that might have evolutionary relevance to plant performance and productivity. Clearly, achieving such a goal cannot be done relying solely on NextGen sequencing, due to cost and throughput considerations at this large scale. However, radically different sequencing (Gen3) technologies are near commercialization and promise quantum leaps in throughputs and correspondingly decreased costs (see, e.g., Pacific Biosciences http://www.pacificbiosciences.com/ [46]). Further, no limitations on continued decreases in costs of sequencing, and associated increases in sequence output, are envisaged over the next decade or so [38]. Planning for the rational use of Gen3 (and GenN and Gen N+1) sequencers therefore represents an important component of coordination activities, given that they should be fully functional in the research setting in the upcoming year.

Justification. As indicated above, rapid advances in technology will ensure that our capabilities to acquire sequence data run the danger of outstripping our abilities to select appropriate questions to be addressed. In effect, we will be in the position of the amateur photographer a few years back, encountering the replacement of film by digital technologies. Without the limitation imposed by the time and cost of buying and developing film, the art of photography has been infinitely changed. Care in selection of subject, exposure, and so on, has been replaced by taking many more pictures in the expectation that at least some will be outstanding. The danger is the inevitable deluge of data, and the lack of both software to adequately curate these data, and storage to archive them. It will be critical to design appropriate experiments and strategies, to assemble teams, and to generate funding resources, such that a complete molecular census of the angiosperms can be achieved.

3.3. What Is the Value of Coordinated Planning in This Area?

Performing a complete molecular census of the angiosperms will require coordinated planning. First and foremost, we need to establish a network of scientists who display the ability to work interactively and without egotism, who also are aware of the technical, technological, and infrastructural issues that relate to the collection of census data, and who are sensitive to the cultural, social, and geographical issues that necessarily accompany collection activities. Coordination will also lead to the identification of taxonomic and collection deficiencies required for census implementation. This will impact the training programs recognized as important in the plant sciences, and is certainly anticipated to put much more emphasis on the "classical" training of plant systematists, morphologists, and taxonomists.

The scientists wishing to participate in coordination activities should have a common interest in genome structure, function, and evolution, and, since they use many different genetic, molecular, and computational strategies to investigate related questions, a greater level of interaction would promote synergistic activities. Analyzing genome sizes across all angiosperms would provide a solid foundation for initiating collaborations. For instance, if some lineage of plants were found to have unusually rapid rates of genome size change, this would provide an impetus to those investigating transposable elements to look for *de novo* element activity, DNA repair experts to see how this process can change over short evolutionary timeframes, and ecologists/evolutionary biologists to investigate the possible mechanistic origins and functional outcomes of these changes.

As in all comprehensive research projects, the great majority of the samples will be simple to collect and analyze, but the last few samples are likely to be recalcitrant to analysis or (more likely) collection. One of the first tasks of the coordinating group will be to set initial priorities

for sample analysis. Although these priorities cannot be established in advance of assembly of a coordinating group with broad disciplinary and international representation, it is likely that the first studies would be focused on full sampling across the angiosperm phylogeny. One can imagine that the secondary priorities will be for deeper sampling of specific families that have interesting priorities discovered in the first broad sampling and/or families that have active research communities.

A first glance at genome size data can also identify lineages that have recently changed their ploidy, where 1*X*, 2*X*, 4*X* and so forth, genome size variation is observed among close relatives, since measurement of *C*-values is less cumbersome than measurement of chromosome numbers. Also, analyzing a full range of angiosperms for genome size would indicate patterns in genome change that may have been missed as a consequence of observing across an unevenly sampled and tiny subset of plant genomes. It could identify, for example, which TEs are the most likely to amplify dramatically in one lineage compared to another. The data so far show no patterns beyond shared events in closely related species. However, with each genome comprising only a single data point, we only have a couple dozen (mostly wildly scattered) data points. More lineages need to be investigated, and genome size determinations would be a first step toward rationally selecting sample sequence analyses that can determine genome composition with a small amount of shotgun sequence data [42].

Finally, research coordination activities including participants who have been conducting field work in many countries over the years would make use of current best practices and the contacts for collaboration. The characterization, preservation, and utilization of angiosperm genetic diversity is a vital issue worldwide, so a full buy-in to this proposed approach is expected from the full international community of plant scientists.

In summary, the proposed coordination activities would aim to bring together a range of scientists with similar interests who have not previously worked together, but whose collective interests span multiple relevant, interconnected disciplines and technologies. We expect that the proposed collaboration will generate genome size data leading to myriad and diverse spinoffs, as the data will implicate particularly interesting lineages for further investigation, including cases where genome change appears to be rapid or in an unexpected direction. Once genome sizes are known broadly, future choices of genomes that deserve full shotgun (or complete, pending technological advances and funding) sequence analysis can be based on a much more complete knowledge foundation.

3.4. How Does This Census Relate to Crops and Food Production? The census relates to crops at multiple direct levels. First, the data generated by grants that will be written or otherwise facilitated by the coordination and planning activities should first uncover unusual and unexpected properties of angiosperm genome size, leading to discoveries regarding repeat structure and, finally, genome sequence. The identification of wild diploid relatives of major and minor crops will provide a road map for complete sequencing, and will focus the development of other molecular tools for those crops. The relevance of orphan crops, particularly in developing countries, should be mentioned. In general, these have not benefited from the application of modern and intense breeding programs, and therefore dramatic yield improvements may be feasible. Finally, through the many investigations stimulated by the interactions enabled through the census, we fully anticipate the discovery of novel genomic features and sequence types, whose evolutionary properties and mechanistic underpinnings presently are unknown but which might have direct relevance to crop improvement.

4. Implementation of Research Coordination Activities

What type of research coordination might be considered? We propose the following.

4.1. An Annual Workshop. This workshop would be designed to provide theoretical underpinning and practical training in census activities. Ideally, workshop participants would include graduate students and postdoctoral research associates, in order to provide them direct experience in technical methods, which are envisaged as being either based on wet laboratory or *in silico* approaches. An embedded theme in our call is the establishment of distributed technologies of reasonable cost to address the acquisition of a complete angiosperm census. Workshop topics could include (i) measurement of plant genome *C*-values using small footprint (essentially portable) flow cytometric instrumentation, (ii) the use of small footprint sequencers in a distributed environment (e.g., the Life Technologies Ion Torrent, and the Illumina MySeq), (iii) data storage and manipulation (particularly using cloud-based solutions), (iv) GPS-based methods for plant imaging, taxonomic assignment, and archiving, and (v) current issues relating to plant classification. Issues of resource protection, sovereignty and ownership, and ethics will also be integrated into these workshops. Finally, outreach activities to schools would be designed and implemented to extend awareness of these emerging issues.

4.2. An Annual Conference. An annual conference centered around census goals and activities would provide a natural environment to establish collaborative activities.

4.3. A Website. This would be set up to allow registration of network participants, to promote interactions among these participants, to archive methods and technologies, and to provide hyperlinks to resources, for example, to tools for resolving conflicts in the application of scientific names.

4.4. A Program of Exchange Visits. These would be between participating laboratories, to allow efficient transfer of information, particularly for cross-training students, postdoctoral research associates, and beginning faculty.

4.5. A Program of White Papers. These would outline future research needs and directions, Grand Challenge questions, and would provide written support and resources for proposals to acquire the specific data types, starting with *C*-value measurements, and moving on to sample sequencing, and then to complete genome sequencing. Specific topics might include (i) establishing a *C*-value analysis pipeline for seed storage repositories, including those at botanical gardens, government facilities, and international agencies (e.g., CGIAR centers), (ii) issues and problems of sample collection and taxonomic identification, (iii) integration of disparate information types, and (iv) ongoing maintenance and curation of databases.

4.6. A Research Coordination Committee. The plant *C*-value community would be polled to elect a coordination committee that would discuss issues related to implementation of the proposed research program. The election, tenure, and activities of this committee would be comparable to that of the Maize Genetics Executive Committee (http://www.maizegdb.org/mgec.php) and other organism-specific research organization groups, like those developed for *Arabidopsis*, sorghum, soybean, and wheat. The primary purpose of such a group would be to foster communication within the plant *C*-value research community and between the researchers and external interested parties, like governments, funding agencies, industry, and NGOs.

4.7. Coordinated Searches for Funding. This will be a critical part of any proposed census, and the search for funds for implementation of the collection, measurement, and archiving activities will be challenging. Review panels in national funding agencies tend not to favor projects that propose surveys without also addressing an underlying biological question. Part of this bias comes from the predominance of hypothesis testing as the core of the scientific method within biological research, and it contrasts dramatically with the situation in other observational sciences, such as astronomy, where large-scale collection of data in the absence of hypotheses is the norm [38]. Nonetheless, construction of biological databases (e.g., GenBank and before it the Atlas of Protein Sequence and Structure) has been integral to the development of modern biology; surveys and their resulting data are incorporated into more experimental realms of biology almost automatically when they are found to be useful [47]. It is likely that progressive recognition, by society, of anthropogenic change to the environment will alter the perception of how biological research should be done, and will lend urgency to completing the proposed census. If so, funding will naturally follow. Nonetheless, coordination of funding activities on an international scale will present challenges of its own

5. Available Resources

Notable resources in terms both of infrastructure and research personnel already exist, and these would provide an excellent foundation for the proposed coordination activities.

5.1. Natural Repositories. Botanical gardens represent the historical locations for plant collections, both living and in archival vouchered forms. Some important examples are in the developed countries, including the Royal Botanic Gardens, Kew (RBG Kew), the Arnold Arboretum, the New York Botanical Garden, and the Missouri Botanical Garden. Others, in developing countries, access notable biodiversity (for example, Xishuangbanna, China, and Bogor, Indonesia). RBG Kew has established a program of measurement of *C*-values and maintains important web resources relating to plant biodiversity. All of these gardens have implemented, or are implementing, programs for extraction and archiving of genomic DNA. Some have active federal support for assessing biodiversity, for example, the Center for Tropical Forest Science, a joint venture of the Arnold Arboretum and the Smithsonian Tropical Research Institute, has funding that supports Biodiversity Workshops in collaboration with scientists in China and the developing world.

A second type of collection is exemplified by seed storage programs at the Svalbard Global Seed Vault, at the USDA facility in Fort Collins, Colorado, and by RBG Kew at Wakehurst Place. The first and second are repositories of seed accessions primarily of crop species of importance to the world and to US agriculture, respectively. The third has a comprehensive mandate, having already banked dry seed of 10% of the world's seed plant diversity since 2000, and aiming to reach 25% by 2020. For the two latter locations, a regular cycle of seed germination, to verify viability in storage over time, is an integral part of the program. The seedling samples generated in this way represent an obvious and available resource for genome size measurements and for genomic DNA extraction.

Global resources in terms of research personnel include cytologists active in genome size measurements, scientists involved in large-scale genome sequencing efforts, taxonomists, systematists, evolutionary biologists, and bioinformaticians. These interested and trained scientists are the most vital resource of all, and the proposed pan-angiosperm *C*-value census will add tremendous impetus for growth of this community.

5.2. Cytologists. A number of cytologists are active in flow cytometric genome measurement, particularly in Europe. Work by cytologists and collaborators has defined the means to efficiently obtain *C*-values from plants (e.g., [26, 28, 48], to correlate *C*-values with environmental population distributions within species [49], to examine genome size within the context of angiosperm evolution [22, 50], to employ unexpected sources of plant samples for *C*-value measurement [51], to sample geographic locations [52–54], and to organize and provide this information in searchable form to the community ([25] and references therein). Researchers in Europe are particularly active in medium-scale projects to define angiosperm *C*-values within specific geographic regions. Coordinating activities within

this community would serve primarily to prioritize activities to be covered by grant applications, to avoid duplication, and to efficiently translate research findings into practical advances across the network.

5.3. Large-Scale Genomics Analysis. The rate of production of genomic sequences is increasing nearly exponentially, driven by considerable innovation in instrumentation and sequencing technologies [38]. Coordination of sequencing activities is largely driven by funding concerns: many crop species that have been sequenced (e.g., maize), and that are being sequenced (wheat), have such large genomes that whole genome sequencing can only be achieved by consortia. Other consortia are sampling many individuals within single species, or are sampling representative individuals within a specific genus. Many of these consortia span the globe, and some employ combinations of expertise in cytology and sequencing to address previously intractable issues (e.g., flow sorting of chromosome arms of wheat prior to the use of NextGen sequencing). Some consortia center around a single, large-scale sequencing facility, the Beijing Genomics Institute in Shenzhen and the JGI in the US being prime examples. One advantage of large sequencing centers is the speed with which they can incorporate advanced sequencing technologies into their pipelines, and the consequent commoditization of sequencing costs.

5.4. Taxonomists, Systematists, and Evolutionary Biologists. A number of collaborative mechanisms currently exist in the US that link personnel and activities within this group. For example, an NSF Research Coordination Network already exists on the topic of Microevolutionary Molecular and Organismic Research in Plant History (microMorph; http://www.colorado.edu/eeb/microMORPH/index.html), as well as two NSF Centers: the National Evolutionary Synthesis Center (NESCent) in Durham, North Carolina, and the National Center for Ecological Analysis and Synthesis (NCEAS) at UC Santa Barbara. Efficient collaboration with these centers, and others, such as the NSF iPlant Collaborative, which is charged with development of cyberinfrastructure to address, for example, issues of evolutionary patterns and mechanisms and appropriate phylogenetic sampling, should facilitate attaining the goals of the angiosperm census.

5.5. Bioinformaticians. A key program currently driving collaboration in plant bioinformatics in the US is the iPlant Collaborative. Close interaction with this program and its scientists will be important for developing and using informatics tools to link international efforts in angiosperm genomics. iPlant has links with the 1KP project, and via that to NESCent, further extending the potential for collaborative links.

Overall, the proposed pan-angiosperm census should develop and strengthen communication, knowledge, and scientific training between individuals, and their laboratory members, that are acknowledged experts in one, or at most two, of the five research focus areas identified above. Understanding the gaps in information between experts in these areas will allow us to design workshops that effectively identify these gaps, and fill them. Through serving as a bridge between different international groups that are addressing similar census goals, we should be able to facilitate the set-up and funding of formal activities that efficiently divide census tasks across geographical locations, and that successfully engage the support of the various governments. We envisage the process of writing white papers as an important aspect of our activities, since this should provide local funding applications with an international imprimatur signifying approval from the global scientific community. We emphasize that although a number of local activities are in place to address the goal of a molecular census, none are comprehensive and all would benefit from a coordination mechanism such as that suggested here.

6. Summary and Conclusions

We are currently at an unprecedented moment for the investigation of life on Earth. The scale of generation of DNA sequence information dwarfs any previous type of biological information gathering. Decreasing costs of these analyses, and increasing power in tools for the extraction of biologically meaningful insights from these data, continue to advance at extraordinary rates. All plants harbor genetic novelty with potential agricultural, biomedical, environmental, and industrial value, and the small number of species with any molecular analysis at all indicates that only a tiny portion of this value has been identified. Still, with 500,000 species, full genome sequence analysis of all angiosperms is not on the near-term horizon. Priorities need to be set, with genome size and ploidy as key criteria. Moreover, genome size is itself an important biological feature that can help tell us how genomes evolve and function. A broad characterization of *C*-values across the angiosperms would identify biological and/or environmental correlates with nuclear genome size, would uncover the general rules of genome size variation, would discover lineages where unusual genomic alterations occurred or are occurring, and would indicate which species are most appropriate for sample sequence or full genome sequence analysis. The technologies for angiosperm *C*-value analysis have themselves dramatically decreased in cost and increased in robustness/availability of late. Hence, the time is right to undertake study of the nuclear genome DNA content for the entire range of angiosperms.

We propose a coordinated process to set priorities and foster communications between laboratories worldwide that will pursue *C*-value analysis of the angiosperms. This coordination would include an annual workshop and an annual meeting, web-based tools, an elected coordination committee, exchange visits, white papers, and organized searches for funding. The community to undertake the research exists, although we expect it to grow dramatically over the next few years, but it is currently composed of a large number of scattered and uncoordinated research laboratories. Our proposal seeks to maintain the creative independence of these groups and new entrants to the field, but to provide them with the communication, coordination, and data analysis tools that will make their work most productive.

References

[1] R. F. Thorne, "How many species of seed plants are there?" *Taxon*, vol. 51, no. 3, pp. 511–512, 2002.

[2] A. J. Paton, N. Brummitt, R. Govaerts et al., "Towards target 1 of the global strategy for plant conservation: a working list of all known plant species—progress and prospects," *Taxon*, vol. 57, no. 2, pp. 602–611, 2008.

[3] B. Bremer, K. Bremer, M. W. Chase et al., "An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III," *Botanical Journal of the Linnean Society*, vol. 161, no. 2, pp. 105–121, 2009.

[4] D. Bramwell, "How many plant species are there?" *Plant Talk*, vol. 28, pp. 32–34, 2002.

[5] T. J. Davies, T. G. Barraclough, M. W. Chase, P. S. Soltis, D. E. Soltis, and V. Savolainen, "Darwin's abominable mystery: insights from a supertree of the angiosperms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 7, pp. 1904–1909, 2004.

[6] R. Dirzo and P. H. Raven, "Global state of biodiversity and loss," *Annual Review of Environment and Resources*, vol. 28, pp. 137–167, 2003.

[7] N. E. Stork, "Re-assessing current extinction rates," *Biodiversity and Conservation*, vol. 19, no. 2, pp. 357–371, 2010.

[8] A. D. Barnosky, N. Matzke, S. Tomiya et al., "Has the Earth's sixth mass extinction already arrived?" *Nature*, vol. 471, no. 7336, pp. 51–57, 2011.

[9] American Museum of Natural History, "National survey reveals biodiversity crisis—scientific experts believe we are in midst of fastest mass extinction in Earth's history," 1998, http://www.amnh.org/museum/ press/feature/biofact.html/.

[10] D. W. Galbraith, K. R. Harkins, and S. Knapp, "Systemic endopolyploidy in *Arabidopsis thaliana*," *Plant Physiology*, vol. 96, no. 3, pp. 985–989, 1991.

[11] M. Barow and A. Meister, "Endopolyploidy in seed plants is differently correlated to systematics, organ, life strategy and genome size," *Plant, Cell and Environment*, vol. 26, no. 4, pp. 571–584, 2003.

[12] M. D. Bennett and I. J. Leitch, "Nuclear DNA amounts in angiosperms: targets, trends and tomorrow," *Annals of Botany*, vol. 107, no. 3, pp. 467–590, 2011.

[13] J. L. Bennetzen, J. Ma, and K. M. Devos, "Mechanisms of recent genome size variation in flowering plants," *Annals of Botany*, vol. 95, no. 1, pp. 127–132, 2005.

[14] N. Chantret, J. Salse, F. Sabot et al., "Molecular basis of evolutionary events that shaped the *Hardness locus* in diploid and polyploid wheat species (Triticum and Aegilops)," *Plant Cell*, vol. 17, no. 4, pp. 1033–1045, 2005.

[15] K. M. Devos, J. K. M. Brown, and J. L. Bennetzen, "Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis," *Genome Research*, vol. 12, no. 7, pp. 1075–1079, 2002.

[16] K. Ilic, P. J. SanMiguel, and J. L. Bennetzen, "A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 21, pp. 12265–12270, 2003.

[17] E. Isidore, B. Scherrer, B. Chalhoub, C. Feuillet, and B. Keller, "Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels," *Genome Research*, vol. 15, no. 4, pp. 526–536, 2005.

[18] J. Ma and J. L. Bennetzen, "Rapid recent growth and divergence of rice nuclear genomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 34, pp. 12404–12410, 2004.

[19] J. Ma and J. L. Bennetzen, "Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 2, pp. 383–388, 2006.

[20] R. T. Gaeta and J. C. Pires, "Homoeologous recombination in allopolyploids: the polyploid ratchet," *New Phytologist*, vol. 186, no. 1, pp. 18–28, 2010.

[21] C. Vitte and J. L. Bennetzen, "Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 47, pp. 17638–17643, 2006.

[22] I. J. Leitch, D. E. Soltis, P. S. Soltis, and M. D. Bennett, "Evolution of genome size in the angiosperms," *American Journal of Botany*, vol. 90, no. 11, pp. 1596–1603, 2003.

[23] I. J. Leitch, D. E. Soltis, P. S. Soltis, and M. D. Bennett, "Evolution of DNA amounts across land plants (embryophyta)," *Annals of Botany*, vol. 95, no. 1, pp. 207–217, 2005.

[24] D. W. Galbraith, J. Bartos, and J. Dolezel, "Flow cytometry and cell sorting in plant biotechnology," in *Flow Cytometry in Biotechnology*, L. A. Sklar, Ed., pp. 291–322, Oxford University Press, New York, NY, USA, 2005.

[25] M. D. Bennett and I. J. Leitch, "Plant DNA C-values database," 2010, http://data.kew.org/cvalues/.

[26] D. W. Galbraith, K. R. Harkins, J. M. Maddox et al., "Rapid flow cytometric analysis of the cell cycle in intact plant tissues," *Science*, vol. 220, no. 4601, pp. 1049–1051, 1983.

[27] J. S. Johnston, M. D. Bennett, A. L. Rayburn, D. W. Galbraith, and H. J. Price, "Reference standards for determination of DNA content of plant nuclei," *American Journal of Botany*, vol. 86, no. 5, pp. 609–613, 1999.

[28] J. Doležel, J. Greilhuber, and J. Suda, "Estimation of nuclear DNA content in plants using flow cytometry," *Nature Protocols*, vol. 2, no. 9, pp. 2233–2244, 2007.

[29] K. R. Chi, "Going with the flow," *The Scientist*, vol. 25, no. 5, p. 57, 2011.

[30] D. W. Galbraith, "Simultaneous flow cytometric quantification of plant nuclear DNA contents over the full range of described angiosperm 2C values," *Cytometry A*, vol. 75, no. 8, pp. 692–698, 2009.

[31] J. Suda and P. Travnicek, "Reliable DNA ploidy determination in dehydrated tissues of vascular plants by DAPI flow cytometry—new prospects for plant research," *Cytometry A*, vol. 69, no. 4, pp. 273–280, 2006.

[32] A. V. Roberts, "The use of bead beating to prepare suspensions of nuclei for flow cytometry from fresh leaves, herbarium leaves, petals and pollen," *Cytometry A*, vol. 71, no. 12, pp. 1039–1044, 2007.

[33] D. Haussler, S. J. O'Brien, O. A. Ryder et al., "Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species," *Journal of Heredity*, vol. 6, pp. 659–674, 100.

[34] J. Greilhuber, T. Borsch, K. Muller, A. Worberg, S. Porembski, and W. Barthlott, "Smallest angiosperm genomes found in *Lentibulariaceae*, with chromosomes of bacterial size," *Plant Biology*, vol. 8, no. 6, pp. 770–777, 2006.

[35] B. J. M. Zonneveld, "New record holders for maximum genome size in eudicots and monocots," *Journal of Botany*, Article ID 527357, 4 pages, 2010.

[36] P. SanMiguel, A. Tikhonov, Y.-K. Jin et al., "Nested retrotransposons in the intergenic regions of the maize genome," *Science*, vol. 274, no. 5288, pp. 765–768, 1996.

[37] C. Vitte, O. Panaud, and H. Quesneville, "LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss," *BMC Genomics*, vol. 8, article 218, 2007.

[38] D. W. Galbraith, "The grand challenges in enabling data-intensive biological research," *Frontiers in Genomic Assay Technology*, vol. 2, no. 26, 2011.

[39] P. Green, "2x genomes—does depth matter?" *Genome Research*, vol. 17, no. 11, pp. 1547–1549, 2007.

[40] D. A. Rasmussen and M. A. F. Noor, "What can you do with 0.1× genome coverage? A case study based on a genome survey of the scuttle fly *Megaselia scalaris* (Phoridae)," *BMC Genomics*, vol. 10, article 382, 2009.

[41] P. R. Steele and J. C. Pires, "Biodiversity assessment: state-of-the-art techniques in phylogenomics and species identification," *American Journal of Botany*, vol. 98, no. 3, pp. 415–425, 2011.

[42] J. DeBarry, R. Liu, and J. L. Bennetzen, "Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the assisted automated assembler of repeat families (AAARF) algorithm," *BMC Bioinformatics*, vol. 9, article 235, 2008.

[43] Y. Jiao, N. J. Wickett, S. Ayyampalayam et al., "Ancestral polyploidy in seed plants and angiosperms," *Nature*, vol. 473, no. 7345, pp. 97–100, 2011.

[44] M. D. Ermolaeva, M. Wu, J. A. Eisen, and S. L. Salzberg, "The age of the *Arabidopsis thaliana* genome duplication," *Plant Molecular Biology*, vol. 51, no. 6, pp. 859–866, 2003.

[45] L. E. Flagel and J. F. Wendel, "Gene duplication and evolutionary novelty in plants," *New Phytologist*, vol. 183, no. 3, pp. 557–564, 2009.

[46] J. Eid, A. Fehr, J. Gray et al., "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.

[47] B. J. Strasser, "Collecting, comparing, and computing sequences: the making of margaret O. Dayhoff's *Atlas of Protein Sequence and Structure*, 1954–1965," *Journal of the History of Biology*, vol. 43, no. 4, pp. 623–660, 2010.

[48] J. Loureiro, E. Rodriguez, J. Doležel, and C. Santos, "Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species," *Annals of Botany*, vol. 100, no. 4, pp. 875–888, 2007.

[49] K. H. Keeler, B. Kwankin, P. W. Barnes, and D. W. Galbraith, "Polyploid polymorphism in big bluestem (*Andropogon gerardii* Vitman)," *Genome*, vol. 29, no. 2, pp. 374–379, 1987.

[50] G. Bharathan, G. M. Lambert, and D. W. Galbraith, "Nuclear DNA content of monocotyledons and related taxa," *American Journal of Botany*, vol. 81, no. 3, pp. 381–386, 1994.

[51] E. Sliwinska, I. Pisarczyk, A. Pawlik, and D. W. Galbraith, "Measuring genome size of desert plants using dry seeds," *Botany*, vol. 87, no. 2, pp. 127–135, 2009.

[52] J. Suda, J. Loureiro, P. Travnicek et al., "Flow cytometry and its applications in plant population biology, ecology and biosystematics: new prospects for the cape flora," *South African Journal of Botany*, vol. 75, no. 2, pp. 389–389, 2009.

[53] J. Loureiro, D. Kopecky, S. Castro, C. Santos, and P. Silveira, "Flow cytometric and cytogenetic analyses of Iberian Peninsula Festuca spp," *Plant Systematics and Evolution*, vol. 269, no. 1-2, pp. 89–105, 2007.

[54] S. Siljak-Yakovlev, F. Pustahija, E. M. Solic et al., "Towards a genome size and chromosome number database of balkan flora: C-values in 343 taxa with novel values for 242," *Advanced Science Letters*, vol. 3, no. 2, pp. 190–213, 2010.