Azolla – A Model Organism for Plant Genomic Studies

Yin-Long Qiu[1]*, Jun Yu[2, 3]
[1] *Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109-1048, USA;* [2] *Beijing Genomics Institute, Chinese Academy of Scicences, Beijing 101300, China;* [3] *University of Washington Genome Center, Seattle, WA 98195, USA*

The aquatic ferns of the genus *Azolla* are nitrogen-fixing plants that have great potentials in agricultural production and environmental conservation. *Azolla* in many aspects is qualified to serve as a model organism for genomic studies because of its importance in agriculture, its unique position in plant evolution, its symbiotic relationship with the N_2-fixing cyanobacterium, *Anabaena azollae*, and its moderate-sized genome. The goals of this genome project are not only to understand the biology of the *Azolla* genome to promote its applications in biological research and agriculture practice but also to gain critical insights about evolution of plant genomes. Together with the strategic and technical improvement as well as cost reduction of DNA sequencing, the deciphering of their genetic code is imminent.

Key words: *Azolla*, nitrogen fixation, genome, model organism

Introduction

Azolla, a genus of aquatic ferns (Fig. 1), consists of seven species that all form symbiotic relationships with N_2-fixing cyanobacterium *Anabaena azollae* (Fig. 1; *1*), and is used as a biofertilizer in agriculture worldwide (*2*, *3*). Its most common utilization is the co-cultivation with rice, as water-filled rice paddies provide a perfect habitat for the water fern to propagate. The *Azolla-Anabaena* symbiosis is the only other major natural N_2-fixing process besides the legumes-rhizobia symbiosis that is being utilized in large-scale for agricultural purposes. By a rough estimate, it supplies 40-60 kg nitrogen per hectare to the crop field (*3*). Given the acreage of rice cultivated worldwide, the potential to exploit this low cost, self-renewable, and environmentally sustainable N_2-fixing symbiosis remains enormous.

Plant genomics has been advancing at a breathtaking pace over the past few years. With the sequencing of *Arabidopsis* (*4*) and rice genomes (*5 – 8*) completed and *Lotus japonicus* (*9*), *Medicago truncatula* (http://medicago.toulouse.inra.fr/), and poplar (http://bahama.jgi-psf.org/prod/bin/populus/home.populus.cgi) genomes now underway, we will soon be able to learn more about the organization, diversity, and evolution of the nuclear genome of flowering plants. The primary goal of sequencing these genomes is to identify genes that encode economically important traits and cellular processes in crops and forest trees so that we can further improve them to meet the challenge of providing sufficient and high quality food, fiber, medicine, and shelter to the world's growing population. A comparative approach has proven effective and efficient in genome sequence annotation and gene identification or prediction aside from libraries of experimental evidence, which are known to be both laborious and time-consuming. Moreover, to gain a full understanding of these genes and their functions as well as the genomic environment within which these genes operate, evolutionary analyses complement and enhance the power of experimental assays. From this perspective, it will be of tremendous value to have information from non-flowering plant genomes to realize the full potential of the sequenced plant genomes. The green alga *Chlamydomonas reinhardtii* genome has just been sequenced (http://www.biology.duke.edu/chlamy_genome/) can to some extent fill this information gap, but its distant relationship to flowering plants naturally undermines its utility. Hence, *Azolla*, being a fern and lying at the mid-point between algae and flowering plants in the phylogeny, will be an excellent choice for a future genome sequencing project to gather information for

* **Corresponding author.**
E-mail: yqiu@bio.umass.edu

understanding the working and evolution of the nuclear genome in green plants.

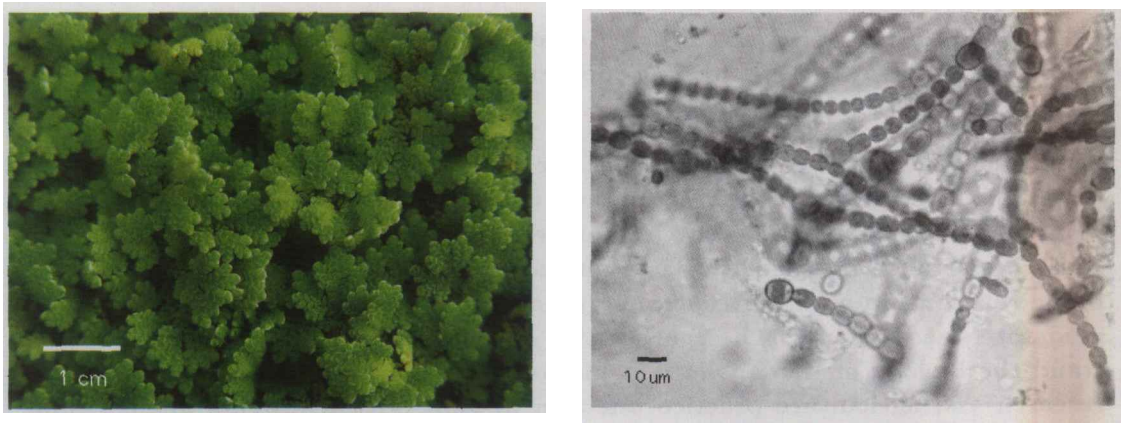

Fig. 1. The growing *Azolla* plant, an aquatic water fern (left, scale bar = 1 cm), and *Anabaena azollae*, the filamentous cyanobacteria from cavities within the leaves of *Azolla* (right, magnified filaments, scale bar = 10 μm).

Biology and Utilization of *Azolla*

Azolla is a water fern that floats on the surface of fresh water ponds, rivers, and flooded fields in the tropics and subtropics of both the New and Old Worlds. Its size is typically 1-2.5 cm in diameter, but some species can achieve the size of 15 cm or more (*1*, *2*, *10*). It belongs to a derived group of ferns Azollaceae and is closely related to several other aquatic taxa such as *Salvinia* and *Marsilea* (*11*). Its fossil history extends back to at least the Upper Cretaceous (*1*, *12*). It is one of the few pteridophyte lineages that have evolved heterospory (with male and female spores in different sizes), a condition that precedes evolution of the seed. Traditionally, seven species are recognized and they are divided into three series, with *A. filiculoides*, *A. rubra*, *A. caroliniana*, *A. microphylla*, and *A. mexicana* in section *Azolla* (found in tropical areas of the Americas, New Zealand, and Australia), *A. pinnata* in section *Rhizosperma* (occurring in Africa, Australia, and Asia), and *A. nilotica* in section *Tetrasporocarpia* (distributed only in Africa) (*1*). A molecular phylogenetic study has now more or less confirmed this taxonomic scheme (*13*), with the only change that *A. nilotica* and *A. pinnata* form a monophyletic instead of a paraphyletic group as in the morphological cladistic analysis (*1*). The cyanobacteria are housed in a cavity of dorsal lobe of the leaf. All of the strains isolated from these seven *Azolla* species belong to *Anabaena azollae*. According to two molecular phylogenetic studies, the bacterial phylogeny is in parallel to that of their host plants (*14*, *15*). Hence, a symbiont-host coevolution has obviously occurred since establishment of the symbiotic relationship in the common ancestor of *Azolla*. (*Salvinia*, the sister group of *Azolla*, does not harbor any of the cyanobacterial symbiont.) *Azolla* has both sexual and asexual reproduction, but the latter is more common. During sexual reproduction, each (female) megaspore contains a small colony of *Anabaena* filaments bearing akinetes (spores of the cyanobacterium), and so the bacterium is passed down to the next generation of the water fern. Its asexual reproduction is achieved simply by multiplication of fragmented fronds (fern leaves). Under ideal growth conditions, the water fern grows and reproduces (asexually) extremely fast, often able to double its weight in 2-5 d (*2*), making it a popular green manure species in tropical and subtropical agriculture.

The *Azolla-Anabaena* symbiosis has been exploited by man for agricultural production for many centuries. Its earliest written record appeared in the ancient Chinese book *Er-Ya*, which was published about 2000 years ago. Another Chinese book, *Qi-Min-Yao-Shu*, on agricultural techniques published by Sixue Jia in 540 A.D., mentioned *Azolla* with respect to applied plant cultivation (*16*). Today, *Azolla* is used extensively in China, India, Vietnam, Thailand, the Philippines, Indonesia, Italy, Senegal, West Africa, Brazil, and other tropical countries (*2*, *3*). Its

primary use is cocultivation with rice to provide natural manure. The most common practice is that before or right after rice planting, *Azolla,* grown from a nursery or a pond, is dispersed into the field. In the first few weeks, the thin canopy of rice plants provides a partial shade and plenty of sunlight required for the water fern to grow. Application of some phosphorus and a small amount of nitrogen fertilizers during this period stimulates growth of both *Azolla* and rice. Under such condition, *Azolla* is able to double its mass between 2-5 d. Later in the rice-growing season, overshading by the rice canopy and draining of water from the field cause *Azolla* to die. Subsequently, the nitrogen fixed by *Anabaena* in *Azolla* as well as the phosphorous, potassium, and other nutrients assimilated by the water fern are released into the environment and used by rice plants for grain formation as well as by the crop planted in the next season. Besides nitrogen, *Azolla* also fixes a substantial amount of carbon through photosynthesis so that organic matters in a steady amount in each rice-growing season are incorporated into the soil to constantly improve the soil composition. Therefore, the use of *Azolla* has a long-term positive effect on agricultural production. This long-term benefit of preserving and maintaining the soil fertility and sustainability of the agricultural ecosystem indeed becomes the single most important advantage that *Azolla* has over chemical fertilizers.

Other than cocultivation with rice, *Azolla* has been used as green manure also for other crops, including wheat, taro, banana, and vegetables. For any crop that requires a waterlogged growth condition, *Azolla* can simply be used through intercropping. In most other cases, *Azolla* grown in fields as a monocrop is to be harvested and ploughed into the soil like any fertilizer before crop planting. These applications have been used in China, India, Southeast Asia, and tropical Africa (*2*). *Azolla* can also be used as a food supplement for domesticated animals such as pigs, cows, ducks, chickens, and fish (*2*). In fact, the rice-*Azolla*-fish and rice-*Azolla*-duck agricultural ecosystems have been practiced in certain parts of China for a long time (*17*). Finally, *Azolla* has been used increasingly for bioremediation. The water fern is found to be able to absorb and break down antibiotics and pesticides used for insect and pathogen control in agricultural fields (*18*, *19*). It is also used to clean up heavy metal pollution in water (*20*, *21*). Thus, there are great potentials for the *Azolla-Anabaena* symbiosis to be exploited both as a biofertilizer and a bioremediation agent.

Why *Azolla* as a Model Organism?

The extensive exploitation of *Azolla* as a biofertilizer for rice cultivation as well as for other purposes such as animal feed and bioremediation provides a strong incentive to study many aspects of its biology.

Several reasons justify selection of *Azolla* as a model organism for functional, comparative, and evolutionary genomic studies of plants. First, it is an economically and environmentally important plant. Model organisms are usually chosen based on either species with convenient biological traits for laboratory studies, such as short life cycle, small dimension, and feasibility for genetic manipulation, or species with economic significance, such as major species of domesticated animals and plants. For genomic studies, because it is relatively expensive to sequence the entire genome of an organism, economically important species are usually given a priority for consideration, e.g., rice, poplar, and *Medicago truncatula*. *Azolla* as a major biofertilizer that can provide an environmentally sustainable and long term self-renewable nitrogen source to agricultural production, despite being only moderately well-studied thus far (*2*, *3*), deserves the level of research investment that has been accorded to other economically important plants like cereals and legumes because of the ultimate enormous payback. An estimate based on the numbers from the year 2000 shows that 88 million metric tons (Tg) of nitrogen fertilizers are needed for agricultural production in the world every year, and 45-50 Tg actually comes from symbiotic N_2 fixation. By 2030-2040, 120 Tg nitrogen fertilizers will then be needed in agriculture to produce food to feed the world's population of 8-10 billions (*22*, *23*). If the contribution of symbiotic N_2 fixation does not increase proportionally, addition of another 32 Tg chemical nitrogen fertilizers each year to the arable soil alone would have an unthinkable and probably uncontrollable impact on the earth's aquatic, terrestrial, and atmospheric ecosystems. Thus, it is imperative that we employ all means to increase the use of symbiotic N_2 fixation systems in order to reduce the use of inorganic nitrogen fertilizers. Among all natural N_2 fixation symbioses, *Azolla-Anabaena* system is the only major one that is applicable to aquatic ecosystems (most legumes-rhizobia systems are terrestrial), especially in rice cultivation in developing countries where much of the future pop-

ulation growth is believed to take place. Hence, its improvement through intensive studies, ranging from acquisition of basic genomic information (such as the genome sequences and gene maps of the relevant organisms) to field applications of our understanding on its basic biological processes, will contribute enormously to our well-being in the future in terms of both food production and environmental protection.

Second, its status as a fern provides another good reason for *Azolla* to be chosen as a model organism for plant genomic studies. Phylogenetic position is being increasingly used to evaluate whether a species makes a good model because the knowledge gained from model organisms will eventually be applied to other species to gain a full understanding of the life on the earth, and a limited number of species that represent the entire spectrum of life's diversity are needed to achieve this goal. This is especially true in the era of genomics as demonstrated by the recent genomic studies from their comparative analyses (*5*, *24*, *25*). At present, most plants that have been or are being sequenced are flowering plants, with the sole exception of *Chlamydomonas reinhardtii* (Table 1), but the latter is a unicellular green alga. Hence, a multicellular non-flowering vascular plant like *Azolla* has extremely good properties to fill the vast gap between green algae and angiosperms to understand evolution of the nuclear genome in plants (*26*). The fact that *Azolla* is the only economically important non-seed plant provides a doubly strong argument for it to be the next target for genome sequencing.

Third, *Azolla* has a moderate-sized genome of 720 Mb. It is more practical to sequence its genome than the 4,000-Mb-sized one of *Ceratopteris richardi* (Table 1), the only other moderately well studied fern. Given the current sequencing technology, in which a large proportion of the sequencing cost is attributed to the sequencing reagents, the genome size is a critical factor in determining whether a species is a good choice. Another strong argument in favor of sequencing a smaller genome is that the larger the genome is, the higher fraction it contains of repetitive sequences that evoke many technical challenges for completing the project.

Fourth, the symbiotic relationship of *Azolla* with N_2-fixing *Anabaena* is another aspect of the water fern that is likely to generate a wealth of information from genomic studies that may enhance our understanding of the well-studied legumes-rhizobia symbiosis through comparative analyses. N_2-fixing cyanobacteria-plant symbiosis evolved only a few times in the 500 million years of land plant evolution: in the liverworts *Blasia* and *Cavivularia*, the moss *Pleurozium schreberi*, the hornworts *Anthoceros*, *Dendroceros*, *Notothylas*, and *Phaeoceros*, the fern *Azolla*, all cycad genera (gymnosperms), and the angiosperm *Gunnera* (*27* – *29*). Among these, only the *Azolla-Anabaena* symbiosis is likely to be intensively studied because of its wide agricultural utilization. The fact that this symbiotic relationship involves a non-flowering plant and a cyanobacterium is going to provide a new angle for understanding many aspects of the rhizobia-legumes symbiosis as well as plant-microbe interaction in general. The symbiont-host recognition, suppression of plant defense reactions, nutrient transport between the two partners, and genetic adaptation and coevolution of both the bacterium and the plant are all important topics to be addressed scientifically. The information gathered on the bacterium-plant symbiosis will also help us to understand another agriculturally and environmentally important plant-fungus symbiotic system— mycorrhizae, as the two systems seem to have involved the same set of plant genes (*30*).

Fifth, *Azolla* is one of the few heterosporous pteridophyte lineages, the others being *Isoetes-Selaginella*, Marsileaceae, and Salviniaceae (*31*). Heterospory is an intermediate condition between homospory and seed-pollen in the evolution of reproductive dispersal units in land plants. The seed is an extremely important plant structure that provides nutrition for human and animals. As of yet, there have been only a few studies initiated to investigate evolution of the seed from a genomic perspective (*32*, *33*). Because none of the heterosporous pteridophyte taxa except *Azolla* has any economic application, they are unlikely to be subjected to intensive studies like most model organisms. *Azolla* would be the only plant, if its genome is sequenced, that can provide a comparative perspective for understanding evolution of heterospory and seed.

Finally, because *Azolla* has been used in agricultural production and environmental pollution control for quite some time, a relatively large body of literature is already in existence, particularly on its physiology and ecology (*2*, *3*, *16*). Its small size, fast growth, and easy culturing should make it fairly easy to grow in laboratories. The International Rice Research Institute in the Philippines (http://www.irri.org/hot1.htm) also has an extensive germplasm collection of all seven species. All these factors should make it easy to popularize *Azolla* as a

model organism.

The Goals

What should be the goals for extensive investigations of *Azolla*, especially the large-scale genomic studies? First, a number of constraints that currently limit the use of *Azolla* as a biofertilizer can be relaxed or removed, so that it can be used on a larger scale in a wide variety of environmental conditions. The wild species of *Azolla* grow optimally at the temperature of 20-30 °C. In the summer of tropics the water fern cannot propagate fast enough to keep in pace with the growth of rice plants. Likewise, in temperate regions

Table 1 The Genome Size of Selected Green Plants*

Plant	Genome Size in Mb	Genome Size in pg (1C)	Reference
green algae			
Chlamydomonas reinhardtii	100 Mb		see text
Caulerpa mexicana	100 Mb		Mandoli
Mesostigma viride	100 Mb		Mandoli
Coleochaete orbicularis	94 Mb		Mandoli
Chara aspera	7200 Mb		Mandoli
bryophytes			
Marchantia polymorpha	300 Mb		Mandoli
Andreaea rupestris	203 Mb	0.21	Kew
Physcomitrella patens	475 Mb		(*53*)
Tortula ruralis	377 Mb	0.39	Kew
Anthoceros sp.	382 Mb		Mandoli
pteridophytes			
Lycopodium clavatum	931 Mb		Mandoli
Selaginella kraussiana	58 Mb	0.06	Kew
Equisetum hyemale	11368 Mb	11.75	Kew
Angiopteris evecta	388 Mb		Mandoli
***Azolla* sp.**	716 Mb	0.74	Arumuganathan/ Qiu/Mandoli unpubl.
Marsilea quadrifolia	426 Mb	0.44	Kew
Ceratopteris richardii	4000 Mb		Mandoli
gymnosperms			
Cycas revoluta	12336 Mb	12.75	Kew
Ginkgo biloba	9627 Mb	9.95	Kew
Gnetum ula	2177 Mb	2.25	Kew
Welwitschia mirabilis	6966 Mb	7.20	Kew
Ephedra tweediana	8611 Mb	8.90	Kew
Pinus caribaea	5564 Mb	5.75	Kew
Abies balsamea	12722 Mb	13.15	Kew
Podocarpus acutifolius	7934 Mb	8.20	Kew
Taxus baccata	10691 Mb	11.05	Kew

Table 1 (*Continued*)

Plant	Genome Size in Mb	Genome Size in pg (1C)	Reference
basal angiosperms			
Amborella trichopoda	900 Mb		Mandoli
Nuphar adventa	3234 Mb		Mandoli
Nymphaea caerula	532 Mb	0.55	Kew
Illicium anisatum	3241 Mb	3.35	Kew
Liriodendron tulipifera	790 Mb		Mandoli
Cinnamomum camphora	580 Mb	0.60	Kew
Aristolochia fimbriata	435 Mb	0.45	Kew
monocots			
Acorus gramineus	400 Mb		Mandoli
Pistia stratiotes	319 Mb	0.33	Kew
Dioscorea togoensis	474 Mb	0.48	Kew
Asparagus officinmalis	1306 Mb	1.35	Kew
Ananas bracteatus	435 Mb	0.45	Kew
Oryza sativa	400 Mb		(*5*)
Zea mays	2641 Mb	2.73	Kew
Triticum aestivum	16767 Mb	17.33	Kew
basal eudicots			
Papaver nudicaule	1693 Mb	1.75	Kew
Amaranthus hypochondriacus	464 Mb	0.48	Kew
Spinacia oleracea	997 Mb	1.03	Kew
Beta vulgaris	1209 Mb	1.25	Kew
rosids			
Arabidopsis thaliana	125 Mb		(*4*)
Citrus aurantium	368 Mb	0.38	Kew
Gossypium klotzschianum	1161 Mb	1.20	Kew
Carya illinoensis	803 Mb	0.83	Kew
Morus alba	822 Mb	0.85	Kew
Cucurbita moschata	416 Mb	0.43	Kew
Rosa wichuraiana	126 Mb	0.13	Kew
Pyrus communis	532 Mb	0.55	Kew
Prunus persica	271 Mb	0.28	Kew
Malus communis	2177 Mb	2.25	Kew
Manihot esculenta	803 Mb	0.83	Kew
Populus balsamifera	550 Mb		see text
Lotus japonicus	472 Mb		(*54*)
Medicago truncatula	464 Mb	0.48	Kew
Glycine max	1093 Mb	1.13	Kew
asterids			
Helianthus annuus	2351 Mb	2.43	Kew
Daucus carota	967 Mb	1.00	Kew
Ipomoea triloba	726 Mb	0.75	Kew
Solanum tuberosum	851 Mb	0.88	Kew

* The species are selected based on their economic or biological importance, arranged according to recent phylogenetic studies (*55 – 59*). Most of the genome size information was obtained from the websites at Royal Botanic Gardens, Kew (http://www.rbgkew.org.uk/cval/homepage.html) and Dina Mandoli's lab at University of Washington, Seattle (http://faculty.washington.edu/mandoli/). For the taxa listed at the Kew site, we converted the genome size measurement in picograms (pg) of DNA to million base pairs (Mb). The species in bold (or their congeners) have been or are being sequenced, or are being characterized for some aspects of their genomes. See the corresponding websites below besides those listed in the text:

Moss genomics project - http://www.plant-biotech.net/pb/pb.html

Cycas genomics project - http://www.nybg.org/pr/PPCycads3.htm

Pine genomics projects - http://pinetree.ccgb.umn.edu/, http://cc.oulu.fi/~genetwww/plants/pinegen.htm

Potato functional genomics project - http://www.tigr.org/tdb/potato/

application of *Azolla* is hampered due to the low temperature. Another problem often encountered in *Azolla* application is its phosphor (P) requirement, as the water fern can derive nitrogen from the cyanobacterial symbiont. A fairly minimal work on mutagenesis, selection, and breeding of *Azolla* and their symbionts has resulted in significant improvement of both these traits. Mutant strains of *Azolla* are able to grow well up to a temperature of 40 °C and requiring as low as 50% P (*3*). The third area where improvement can be made to increase the use of *Azolla* in rice cultivation is its propagation. The current practice is that *Azolla* is raised in a nursery field or a pond, where the water fern mostly reproduces by asexual reproduction, and is then transported to rice fields (*2*). There is a fair amount of labor involved in harvesting, transporting, and dispersing the water fern in this practice. If sexual reproduction can be used to propagate *Azolla*, only a few handfuls of spores need to be thrown into the field and the resulting labor cost can be greatly reduced (*3*). Yet, this method is still not available for large-scale agricultural production.

For these three major bottleneck factors, a completely sequenced *Azolla* genome and a large number of ensuing studies can surely improve our understanding of physiology (particularly temperature sensitivity and phosphor requirement) and reproductive biology of the water fern, and thus increase our ability to genetically engineer strains that can grow in different environmental conditions and be propagated by spores. Hence, it can be predicted that *Azolla* will be used more commonly in agricultural production and bioremediation.

The kind of extensive and intensive studies as usually applied to model organisms on *Azolla* is also likely to greatly enhance our understanding on the symbiotic relationship of the water fern with N_2-fixing cyanobacteria. Given the paucity of N_2-fixing plant-bacterium symbioses that have ever evolved in land plants (see above for cyanobacteria; ref. *34* for rhizobia and *Frankia*), it is safe to say that it will still be some time before any artificially engineered plant-bacterium N_2-fixing symbiotic system comes into existence. Therefore, it will be more efficient to study, improve, and make use of the naturally occurring systems like those of *Azolla-Anabaena* and legumes-rhizobia in order to solve the nitrogen supply problem for agricultural production. The fact that no monocot has any symbiotic association with N_2-fixing bacteria further highlights the potential difficulty to genetically engineer a cereal-bacterium N_2-fixing system. Thus, *Azolla-Anabaena* symbiosis will likely remain as the only choice for supplying the fields with nitrogen in a self-renewable and environmentally sustainable way for rice and rice-wheat growing countries. It is of a pragmatic necessity, not just an intellectual curiosity, to understand and improve this system. Initiating an *Azolla* genome-sequencing project will itself provide an essential amount of basic genomic information about this organism, opening a way for its future research. More importantly, the project will make the *Azolla-Anabaena* system more attractive to many experimental biologists for intensive studies of both the plant and bacterium, which have not been as well studied as the legumes and rhizobia. Historically, designation of an organism as a model naturally leads to intensive characterization of every aspect of the organism. The recent examples are *Lotus japonicus* (*35*) and *Medicago truncatula* (http://medicago.toulouse.inra.fr/). An increased understanding of symbiotic relationships between plants and bacteria or fungi, as can be obtained from studies of *Azolla-Anabaena*, legumes-rhizobia, and plant-mycorrhizae, will provide the much-needed knowledge to help formulate environmentally sustainable policies and practices in agricultural production.

The last major goal of making *Azolla* a model organism for genomic studies is to gather information from the species that represent a major and intermediate step in plant evolution so that comparative

analyses can be conducted to shed light on many aspects of the nuclear genome evolution in green plants. One such aspect concerns genome size. The plant nuclear genome seems to have increased steadily in size at every major step of phyletic evolution, e.g., from green algae to land plants, from bryophytes to vascular plants, and from pteridophytes to seed plants (Table 1). This increase may have been achieved through polyploidization (*36*), as a gradually lengthened sporophyte (diploid) generation in the life cycle provides an opportunity for manifesting the evolutionary advantage of multiple copies of the same or similar genes. Along this major trend, occasional dramatic genome size expansion (e.g., *Chara,* Table 1) or reduction (e.g., *Selaginella,* Table 1) is seen in some lineages, possibly due to further polyploidization and/or multiplication or elimination of particular classes of transposons and other repetitive sequence elements. Such phenomena have been seen in the sequenced genomes of human (*24*) and fission yeast (*37*). The 720-Mb *Azolla* genome is smaller than most pteridophyte genomes (*38*), but significantly larger than those of *Arabidopsis* (*4*) and rice (*5, 6*). Comparisons between the two angiosperm genomes have yielded several interesting insights into possible mechanisms of the genome size variation in plants. First, plant genomes are organized in such a way that the transposon-derived repetitive sequences are scattered between gene-islands where the plant genes are clustered; thus, the increase of genome size due to transposon expansion is largely intergenic (*5, 39, 40*). Second, the rice genome has twice as many genes as predicted for the *Arabidopsis* genome, indicating that an ancient genome duplication event had happened after the split of monocot and eudicot plants. Surprisingly, the extra set of genes in the rice genome lacks homologs in any other known genomes sequenced so far but is definitely transcribed; some genes are expressed at high levels in rice tissues (*5, 41*). Third, the average size of the rice genes is larger than that of the *Arabidopsis,* attributable mainly to the gradual intron size increase (not due to transposon insertions) over the evolutionary time scale. Our preliminary analysis has shown that the difference is extendable between the genes of monocot and eudicot plants in general, regardless of the actual genome sizes. Finally, a GC-content gradient increase starting from the 5' end of gene transcripts is found in most of the rice or monocot genes (*5, 41*). Such a maneuver in DNA composition suggests that evolutionary forces, such as mutation and selection at the DNA sequence level, are working on the genes constantly. Taken together, *Azolla* should have fewer genes than the two sequenced angiosperms, as it has not evolved many features of the latter two, e.g., seed, pollen, flower, and many secondary compounds. It may have lineage-specific transposons that are propagating or deteriorating to affect the genome size. However, we will not know the answer with certainty until the *Azolla* genome is fully sequenced.

Another aspect regarding genome evolution involves gene content. A few studies that have examined evolution of multigene families in land plants show that the copy number increases steadily from charophytes to angiosperms, e.g., actin genes (*42*), MADS box genes (*32*), and phytochrome genes (*43*). It is not known whether the gene copy number increases via individual gene duplication, chromosome segmental duplication, or polyploidization. If the last two mechanisms are responsible for evolution of the multigene families, the next question is to what extent the synteny has been maintained. Comparisons among three flowering plants, *Arabidopsis*, tomato, and soybean show that there are major syntenic blocks conserved on the chromosomes after duplication (*44, 45*). It is tempting to ask how far back during land plant evolution this kind of large-scale syntenic relationships was maintained, and what kind of evolutionary forces were behind the maintenance.

The final aspect concerning genome evolution is the role of "junk" DNA, i.e., transposons and introns. It has been known for a long time that most eukaryotic genomes, especially those of high plants and animals, are packed with highly and moderately repetitive DNAs from DNA denaturation and renaturation (Cot curve plotting) studies. Now whole genome sequence analyses have confirmed this aspect of genome organization. At least one quarter of the rice genome is of recognizable transposons in origin (*5*), whereas for a large genome like that of human, transposons account for about half of the genome (*24*). What are the evolutionary and functional roles of this massive amount of transposon sequences in the genome? It has been suggested that they are involved in origins and functions of centromeres (*46 – 48*), and telomeres (*49, 50*). Further, they are responsible for chromosome rearrangements (*51*). Hence, these seemingly useless DNAs, as judged by geneticists in the traditional sense of coding capacity for phenotype, are actually fundamental forces in maintaining faithful inheritance of the information in the genome from generation to generation, and at the same time breaking

up syntenic relationships among loci after replicated chromosomes are partitioned into daughter species and generating genetic diversity via recombination and independent assortment. In other words, transposons are a "twin engine" that powers the evolution of life (*52*)! Similarly, introns have been dubbed molecular parasites, but now we know that they play an extremely important role in eukaryotes to generate protein diversity by making alternative splicing possible in animal genomes, such as the human genome (*24*). It has been suggested that plants may not use much of the cellular alternative splicing machinery to the same extent as animals (*5*), but this conclusion is drawn from comparisons of two small, perhaps atypical plant genomes (*Arabidopsis* and rice, see Table1) with the large human genome. These are just a few clues seen from a very small sample of diversity of life that has been subjected to whole genome sequencing and analysis. It would be ideal to see if the validity and generality of the conclusions drawn from these studies can be extended to all life when a larger diversity of representative organisms is investigated. Thus, a plant like *Azolla* makes an ideal candidate for future genome sequencing projects to help us understand the evolution of the nuclear genome in plants. In a broader perspective, as we turn over more rocks (sequenced genomes), we will certainly see more surprises (new insights into genome biology and evolution)!

Technical Planning and Cost Estimate

The *Azolla* genome most likely will be sequenced through a whole genome shotgun approach due to the genome size and the readiness of sequence-assembly software tools. It can be carried out in two phases, a genome survey sequencing phase and a gene-map construction phase. In the first phase of the project, sequencing reads (or sequencing traces, usually about 500 bp in average length) of a low coverage (0.1 to 2X coverage of the genome, depending on its repetitive sequence content) should be acquired and analyzed to provide independent evaluation of the genome constituents, such as the type of repeats and the fraction of gene-coding sequences. An adequate amount of expressed sequence tags (ESTs) or cDNAs (say, representing over 10,000 unique genes) from different tissues or developmental stages (to provide enough sample diversity) of the plant should also be sequenced. In order to localize genes on chromosomes, some large-insert clone libraries, such as those of bacterial artificial chromosomes (BACs) and cosmids, should be constructed concurrently for subsequent physical mapping. In the second phase of the project, significant sequence coverage should be achieved, usually in a range of 4-7X of the genome equivalents. For the *Azolla* genome, 1X coverage is equal to 1.4 million sequencing reads. A draft sequence assembly should be obtained at this point together with some clone-end sequences from large-insert clone libraries to provide a framework in the sequence-assembling process. Genes now can be identified based on gene-prediction software and verified by the acquired EST sequences. Nearly 95% of the genes in the genome can be identified from a working draft sequence. A vigorous completion strategy has to be applied for the final 5% genes, adding significant cost to the grand total. It is still debatable to what extent a genome project is defined as "finished" or "complete" since the plant genomes desirable for sequencing are mostly too large to be even considered as tangible targets.

Based on current academic institutional costs (including both direct and indirect costs) and the cost reduction trend over the past 5 years in large-scale sequencing efforts, the total cost of the project should be around 6 to 10 million US dollars for a start-date in the current fiscal year. The phase one goals of the project may only cost about 2 million US dollars to fulfill. Therefore, the genome project of this magnitude in scientific and agriculture significance is very much a policy decision for the funding agencies rather than a scientific one. Scientists who are eager to use the basic genomic information yield from the *Azolla* Genome Project should not shy about expressing their thoughts about the prospective scientific and practical yield of this endeavor.

Acknowledgments

We thank Dina Mandoli and Aru K. Arumuganathan for their help with determining the genome size of *Azolla* sp. We also thank Anton Manuilov, Olena Dombrovska, Dale A. Callaham, and The Central Microscopy Facility at University of Massachusetts, Amherst for their help in photographing *Azolla* sp. and *Anabaena azollae*. Y.Q. is supported by a NSF CAREER Award (DEB 0093012). J.Y. is supported by two grants from Chinese High-tech Program "863" of the Ministry of Science and Technology

(2001AA231061 and 2002AA229021).

References

1. Sanders, R. M. K. and Fowler, K. 1993. The supraspecific taxonomy and evolution of the fern genus *Azolla* (Azollaceae). *Pl. Syst. Evol.* 184: 175-193.
2. Wagner, G. M. 1997. *Azolla*: a review of its biology and utilization. *Bot. Rev.* 63: 1-26.
3. Vaishampayan, A., *et al.* 2001. Cyanobacterial biofertilizers in rice agriculture. *Bot. Rev.* 67: 453-516.
4. The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796-815.
5. Yu, J., *et al.* 2002: A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79-92.
6. Goff, S. A., *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92-100.
7. Sasaki, T., *et al.* 2002. The genome sequence and structure of rice chromosome 1. *Nature* 420: 312-316.
8. Feng, Q., *et al.* 2002. Sequence and analysis of rice chromosome 4. *Nature* 420: 316-20.
9. Sato, S., *et al.* 2001. Structural analysis of a *Lotus japonicus* genome. I. Sequence features and mapping of fifty-six TAC clones which cover the 5.4 Mb regions of the genome. *DNA Res.* 8: 311-318.
10. Schneller, J. J. 1990. Azollaceae. In *The Families and Genera of Vascular Plants, Vol. I, Pteridophytes and Gymnosperms* (ed. Kramer, K. U and Green, P. S), pp. 57-60. Springer-Verlag, Berlin, Germany.
11. Hasebe, H., *et al.* 1995. Fern phylogeny based on rbcL nucleotide sequences. *Amer. Fern J.* 85: 134-181.
12. Archangelsky, A., *et al.* 1999. Paleoazolla, a new heterosporous fern from the Upper Cretaceous of Argentina. *Amer. J. Bot.* 86: 1200-1206.
13. Reid, J. D., *et al.* 2002. Systematics of the genus *Azolla* (Azollaceae). *Amer. J. Bot.* 88: (abstract) http://www.2002.botanyconference.org/section11/abstracts/4.shtml.
14. Vancoppenolle, B., *et al.* 1995. Genetic diversity and phylogeny analysis of *Anabaena-azollae* based on RFLPs detected in *Azolla-Anabaena azollae* DNA complexes using *nif* gene probes. *Theor. Appl. Genet.* 91: 589-597.
15. Zheng, W. W., *et al.* 1999. Genetic diversity and classification of cyanobacteria in different *Azolla* species by the use of PCR fingerprinting. *Theor. Appl. Genet.* 99: 1187-1193.
16. Shi, D. J. and Hall, D. O. 1988. The *Azolla-Anabaena* association: historical perspective, symbiosis and energy metabolism. *Bot. Rev.* 54: 353-386.
17. Watanabe, I. and Liu C. C.1992. Improving nitrogen-fixing systems and integrating them into sustainable rice farming. *Pl. & Soil* 141: 57-67.
18. Forni, C., *et al.* 2002. Sulphadimethoxine and *Azolla filiculoides* Lam.: a model for drug remediation. *Water Res.* 36: 3398-3403.
19. Singh, N. and Sethunathan, N. 1999. Degradation of carbofuran by an enrichment culture developed from carbofuran-treated *Azolla* plot. *Pesticide Science* 55: 740-744.
20. Antunes, A. P. M., *et al.* 2001. Batch studies on the removal of gold (III) from aqueous solution by *Azolla filiculoides*. *Biotech. Lett.* 23: 249-251.
21. Cohen-Shoel, N., *et al.* 2002. Biofiltration of toxic elements by Azolla biomass. *Water Air and Soil Pollution* 135: 93-104.
22. Socolow, R. H. 1999. Nitrogen management and the future of food: lessons from the management of energy and carbon. *Proc. Natl. Acad. Sci., USA* 96: 6001-6008.
23. Vance, C. P. 2001. Symbiotic nitrogen fixation and phosphorous acquisition: plant nutrition in a world of declining renewable resources. *Pl. Physiol.* 127: 390-397.
24. International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
25. Hedges, S. B. 2002. The origin and evolution of model organisms. *Nat. Rev. Genet.* 3: 838-849.
26. Qiu, Y.-L. and Palmer, J. D. 1999. Phylogeny of early land plants: insights from genes and genomes. *Trends Plant Sci.* 4: 26-30.
27. Meeks, J. C. 1998. Symbiosis between nitrogen-fixing cyanobacteria and plants. *Bioscience* 48: 266-276.
28. Rai, A. N., *et al.* 2000. Cyanobacterium-plant symbioses. *New Phytol.* 147: 449-481.
29. DeLuca, T. H., *et al.* 2002. Quantifying nitrogen-fixation in feather moss carpets of boreal forests. *Nature* 419: 917-920.
30. Stracke, S., *et al.* 2002. A plant receptor-like kinase required for both bacterial and fungal symbiosis. *Nature* 417: 959-962.
31. Gifford, E. M. and Foster, A. S. 1989. *Morphology and Evolution of Vascular Plants, 3rd ed.* Freeman and Company, New York, USA.
32. Theissen, G., *et al.* 2000. A short history of MADS-box genes in plants. *Plant Mol. Biol.* 42:115-149.
33. Soltis, D. E., *et al.* 2002. Missing links: the genetic architecture of flower and floral diversification. *Trends Plant Sci.* 7: 22-31.
34. Soltis, D. E., *et al.* 1995. Chloroplast gene sequence data suggests a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. *Proc. Natl. Acad. Sci. USA* 92: 2647-2651.
35. Asamizu, E., Watanabe, M., and Tabata, S. 2000. Large scale structural analysis of cDNAs in the model legume *Lotus japonicus*. *J. Plant Res.* 113: 451-455.

36. Stebbins, G. L. 1950. *Variation and Evolution in Plants.* Columbia University Press, New York, USA.
37. Wood., V., *et al.* 2002. The genome sequence of *Schizosaccharomyces pombe. Nature* 415: 871-880.
38. Bennett, M. D. and Leitch, I. J. 2001. Nuclear DNA amounts in pteridophytes. *Ann. Bot.* 87: 335-345.
39. Wong, G.-K., *et al.* 2000. Is "junk" DNA mostly intron DNA? *Genome Res.* 10: 1672-1678.
40. Wong, G.-K, Passey, D.A., Yu, J. 2001. Most of the human genome is transcribed. *Genome Res.* 11: 1975-1977.
41. Wong, G.-K., *et al.* 2002. Compositional gradients in Gramineae genes. *Genome Res.* 12: 851-856.
42. Bhattacharya, D., *et al.* 2000. Actin gene duplication and the evolution of morphological complexity in land plants. *J. Phycol.* 36: 813-820.
43. Schneider-Poetsch, H. A. W., *et al.* 1998. Non-angiosperm phytochromes and the evolution of vascular plants. *Physiol. Plantarum* 102: 612-622.
44. Grant, D., Cregan, P., and Shoemaker, R. C. 2000. Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis. Proc. Natl. Acad. Sci. USA.* 97: 4168-4173.
45. Ku, H.-M., *et al.* 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci. USA.* 97: 9121-9126.
46. Copenhaver, G. P., *et al.* 1999, Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* 286: 2468-2474.
47. Cheng, Z., *et al.* 2002. Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Pl. Cell* 14: 1691-1704.
48. Zhong, C. X., *et al.* 2002. Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. *Pl. Cell* 14: 2825-2836.
49. Levis, R. W., *et al.* 1993. Transposons in-place of telomeric repeats at a *Drosophila* telomere. *Cell* 75: 1083-1093.
50. Arkhipova, I. R. and Morrison, H. G. 2001. Three retrotransposon families in the genome of *Giardia lamblia*: two telomeric, one dead. *Proc. Natl. Acad. Sci., USA* 98: 14497-14502.
51. Lonnig, W.-E. and Saedler, H. 2002. Chromosome rearrangements and transposable elements. *Annu. Rev. Genet.* 36: 389-410.
52. McClintock, B. 1984: The significance of responses of the genome to challenge. *Science* 226: 792-801.
53. Reski, R., *et al.* 1994. Genome analysis of the moss *Physcomitrella patens* (Hedw.) B.S.G. *Mol. Gen. Genet.* 244: 352-359.
54. Ito, M., *et al.* 2000. Genome and chromosome dimensions of *Lotus japonicus. J. Plant Res.* 113: 435-442.
55. Qiu, Y.-L., *et al.* 1998. The gain of three mitochondrial introns identifies liverworts as the earliest land plants. *Nature* 394: 671-674.
56. Qiu, Y.-L., *et al.* 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402: 404-407.
57. Soltis, D. E., *et al.* 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL* and *atpB* sequences. *Bot. J. Linn. Soc.* 133: 381-461.
58. Karol, K. G., *et al.* 2001. The closest living relatives of land plants. *Science* 294: 2351-2353.
59. Pryer, K. M., *et al.* 2001. Horsetails and ferns are a monophyletic group and the closest relatives to seed plants. *Nature* 409: 618-622.

Received: 27 December, 2002
Accepted: 8 January, 2003