The Comparative Genomics of Orphan Crops

Andrew H. Paterson[1], Barbara K. Mashope[2] and Lifeng Lin[1]

[1]Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602, USA;[2]Early Detection and Evaluation of Invasive Alien Plants, South African National Biodiversity Institute, Kirstenbosch Gardens, P/Bag X7, Claremont 7735, Cape Town, South Africa

(Email: paterson@uga.edu, Mashope@sanbi.org, lifeng_lin@yahoo.com)

Abstract

The common ancestry of all flowering plants, reflected in the DNA sequences of those which have been sequenced to date, provides opportunities for 'translational biology' -- leveraging information and tools from major crops and botanical models in the study and improvement of many 'orphan crops' that are understudied at the genomic level although they are important sources of food , medicine, or other ecosystem services. We discuss some general approaches for reaping the benefits of translational genomics in orphan crops and offer a small sampling of examples of orphan crops that might benefit from this approach, also noting limitations of the translational approach in understanding unique features of particular plants. Translational genomics offers opportunities for many mutually-fruitful partnerships between African and non-African scientists, requiring ethical and responsible actions from both sides which acknowledge the global forces at play in the arena within which scientists practice.

Keywords: domestication, Nikolai Vavilov, DNA sequence, genetic map, translational genomics, medicine

Introduction

Flowering plants, known to botanists as angiosperms, are the Earth's dominant vegetation, sustaining humanity by providing 'ecosystem services' including oxygen, food, feed, fiber, fuel, medicines, spirits, erosion and flooding control, soil regeneration, urban cooling and green-space, wildlife habitat, and other benefits. Tracing to a common ancestor thought to have existed 140-180 million years ago [1, 2], about 250,000-400,000 angiosperm species are thought to exist [3, 4, 5], and a remarkable average of 2,350 new ones are discovered each year [6].

A tiny subset of about 200 angiosperms have become 'domesticated', over about the past 10,000 years literally co-evolving with humans to such a great degree that " changes in the (crop) population's genotypes ... makes them more useful ... and better adapted to human intervention" [7]. Domestication and ongoing improvement of crop plants ranks among the greatest of human achievements, permitting small numbers of people to feed much larger numbers, and freeing the remainder to pursue science and technology, arts and culture [8] Even after several centuries of progress, modern plant breeding is estimated to offer a 35% rate of return on public research investments [9].

Although often thought of as a Neolithic activity, crop domestication and improvement is an ongoing process driven by changing human needs and priorities, and constraints on agriculture. In particular, growing attention to the needs of Africa is re-awakening the merits of further improving native plants already reasonably well suited to local cultivation, and which have been subjected to degrees of domestication ranging from none to advanced. Even those African crops that have received some scientific plant breeding lag behind other leading crops -- for example, yields of sorghum, a staple for large populations in the African Sahel that is used for feed or ethanol (fuel) production on other continents, only gained a total of 6% from 1961-1963 to 1999-2001[10, 11], much less than maize, rice, and wheat and also far outstripped by population growth.

The sequencing and detailed functional analysis of the genomes of a few select botanical models opens new doors into comparative biology of flowering plants, with especially great potential benefits for improvement of many 'orphan crops' that are under-studied at the genomic level although they feed large populations already [12] and could feed many more people, in some cases with less inputs and at less cost than established crops.

Many orphan crops are also a first line of defense against illness. In South Africa, it is estimated that 60% of the indigenous people consult a traditional healer (sangoma) as a first attempt to get medical attention for a variety of illnesses [14]. Given that more than 50% of all drugs in clinical use in the world are derived from natural products or their derivatives [15], it is alarming that very little breeding or development of these plants has taken place even in developing countries, although a substantial part of their population are highly dependent on these genetic resources for basic healthcare. For example, *Artemisia afra* (wildeals in Afrikaans, lengana in Setswana) is widely used in South Africa for the treatment of a variety of ailments including, coughs, colds, influenza and also fever and malaria [15]. The isolation of the well known anti-malaria drug artemisinin from the Chinese equivalent *Artemisia annua* highlights the potential application of these plants in agricultural biotechnology [16]. Encouragingly, South Africa's Department of Science and Technology has adopted the 'farmer to pharma' strat-

egy which recommends the cultivation of genetically engineered crops for the production of biopharmaceuticals [17]. Artimisin, could very well be one such biopharmaceutical.

Among 27 orphan crops collectively planted to 250 million ha/yr and yielding US$100 billion/yr farm gate value in the developing world, only 4 (barley, sorghum, cassava, and sunflower) had appreciable numbers of sequences (>10,000) in Genbank as of 11 Feb 2005 [12]. Pearl millet and tef are prime examples of such orphan crops having utmost importance in feeding millions of people [18a, 18b], yet with limited resources in Genbank. Sorghum recently became the first of these, and indeed the first plant of African origin, to have its genome fully sequenced [19]. Efficient methods to leverage genomic knowledge for botanical models in the study and improvement of orphan crops will play a central role in translation of genomic discovery research into increased agricultural development of indigenous medicinal and food crops by developing countries.

Shared ancestry of all flowering plants in the common language of DNA

Nearly a century ago, Nikolai Vavilov suggested an underlying commonality to exist in the hereditary information of diverse crop plants [20]. Today, based on many different lines of evidence we believe that all flowering plants trace to a common ancestor thought to have existed 140-180 million years ago [1, 2], with most modern lineages of flowering plants having been established by the mid-late Cretaceous period [21, 22]. While Vavilov's observations were based on similarities in the appearance (phenotype) of different plants, the transition to DNA-based plant genetics about 25 years ago has revealed commonality among the hereditary information of diverse crop (and other) plants at three levels:

i) Genetic and/or physical mapping. Measuring recombination and/or physical quantity of DNA between reference 'DNA markers' along the chromosomes of different species, shows common DNA markers to exhibit parallels in chromosomal affinities ('synteny') and orders along the chromosomes ('collinearity'). Very strong parallels in gene arrangement are generally found among different species within common genera (for example *Solanum tuberosum*, potato and *Solanum lycopersicon*, tomato, thought to be separated by only a few million years of evolution [23]); lesser but still clear parallels are common among divergent genera within a taxonomic family (for example, *Sorghum bicolor*, sorghum, and *Oryza sativa*, rice, members of the Poaceae family that are thought to be separated by ~50 million years of evolution [19]); and 'islands' of non-random similarity remain discernible even among clades such as monocots and eudicots that diverged early in angiosperm evolution [24, 25]. These parallels can be obscured to varying degrees in different lineages by polyploidy and associated gene loss, by structural rearrangement of chromosomes, and by duplication and/or transposition (movement) of individual genes or small groups of genes.

ii) Whole-genome sequences. Once a daunting task, requiring an international consortium and many tens of millions of dollars to sequence the first angiosperm genome [26], as of this writing primary descriptions of ten angiosperm genome sequences have been published, including the eudicots *Arabidopsis thaliana* [26], *Populus trichocarpa* (poplar, or black cottonwood) [27], *Vitis vinifera* (grapevine) [28, 29], *Carica papaya* (papaya) [30], *Cucumis sativus* [31] and *Glycine max (*soybean) [32]; and the monocots *Oryza sativa* (rice, including each of two subspecies) [33, 34, 35, 36], *Sorghum bicolor* (sorghum) [19], *Zea mays* (maize) [37], and *Brachypodium distachyon* [38]. Primary descriptions and/or sequencing are in progress for many more. Growing evidence shows that cytologically-distinguishable domains of chromosomes classically referred to as 'euchromatin' and 'heterochromatin' are also distinctive in their DNA content and evolutionary histories. Euchromatin is relatively rich in genes and DNA transposons, and relatively poor in rapidly-evolving elements such as retrotransposons and other repetitive DNA such as centromere-associated elements. As with genetic and/or physical maps, parallels among genomes in euchromatin range from near-identity of gene arrangements in the respective subspecies of rice to 'islands' of non-random similarity between monocots and eudicots. In contrast, heterochromatin is gene-poor, rich in repetitive DNA (indeed, accounting for most of the difference in physical size of some genomes) and is rapidly evolving in both structure and individual DNA elements, with little discernible synteny and/or colinearity among taxa [39]. Despite its rapid evolution in DNA content and order, the physical location of the heterochromatic domains of genomes appears to remain approximately constant for tens of millions of years [19, 32, 39]. Because heterochromatin is largely recalcitrant to recombination, it accounts for very little of genetic maps, often being reflected as clusters of DNA markers that are poorly resolved, although they may be physically very far apart.

iii) Individual genes and their sequences. Most flowering plants share most of their basic gene sets with one another. A recent comparison of Arabidopsis, poplar, grape, and papaya yielded estimates of 'ancestral gene number' varying from 10149 for *Carica* to 13043 for *Populus*, a range very similar to the 12,000-14,000 from a previous estimate based on an independent gene birth model [40]. Another independent analysis comparing OrthoMCL-defined gene families (identified based on overall conservation of inferred protein sequence – [41]) in an independent genome (sorghum), to those of *Arabidopsis*, rice and poplar, likewise suggested 11,502 ancestral angiosperm gene families represented in at least one contemporary grass and rosid genome [19]. However, duplication of individual genes and entire genomes has conferred substantial variations among lineages in 'copy number' of genes that share common ancestry, with actual gene numbers in most angiosperm genomes typically being 2-3x the number of ancestral gene families. Moreover, genes are classified into common gene families based on well-conserved features that are often important to their function – however, outside these features, and even within them, the se-

quence ('spelling') of individual family members can range from some with near-identity to others no longer recognizable as being related to one another.

'Translational genomics' – leveraging information and tools from major crops and botanical models in improvement of less-studied crops

Through the common 'language' of DNA, shared ancestry provides approaches by which to accelerate progress in improvement of indigenous 'orphan crops' by utilizing information from the growing set of botanical models and well-studied crops in several ways. Three broad categories of such approaches follow.

i) *Targeted assays of candidate genes or genomic regions*. The collective efforts of agricultural and life scientists worldwide have revealed the functions of an appreciable set of plant genes already, and progress is accelerating. Many genes function in similar ways in different plants, and the function of a gene in a botanical model is predictive of its function in crops that are separated by the model by tens, even hundreds of millions of years of evolutionary history. In a particularly prominent example, a wheat gene conferring reduced height and increased harvest index of wheat that contributed to the 'Green Revolution' is clearly recognizable as a homolog of an *Arabidopsis* gene conferring insensitivity to the plant hormone gibberrellin [42], notwithstanding that the wheat and *Arabidopsis* genes have experienced ~150 million years of independent evolutionary history. Hence, an attractive means for leveraging genomic information is analogy – recognition that a trait of interest in a less-studied crop is similar to a well-known trait in a well-studied plant, followed by targeted evaluation of candidate genes. For example, envision that one had identified in funde (*Digitaria exilis*, a drought tolerant West African grain crop that is taxonomically a member of the panicoid cereals) a strain with reduced height and increased harvest index. By crossing the new strain to a conventional strain and inbreeding the F_1 hybrid, one would expect to find some F_2 progeny with the conventional (tall, lower yielding) phenotype and others with the desired trait (shorter, higher yielding). One could very simply test the hypothesis that a funde homolog of the Green Revolution gene accounted for the new trait. Briefly, one would scrutinize the Green Revolution gene to identify feature of its DNA sequence that were common to a wide range of known cereal genes (for example, wheat, rice, sorghum, and maize) and tacitly assume that those 'conserved' features were likely to also be shared by funde. Experience has shown this assumption to be reasonable, especially when several candidate features can be targeted [43]. DNA primers for the now-ubiquitous 'polymerase chain reaction' (PCR) can be designed and applied to DNA from each member of the funde population, to test the hypothesis that an allele from the short, high yielding parent correlates very closely with the phenotype in the progeny. Identification of such a diagnostic marker might be of high value in permitting a breeder to identify plants that will exhibit a desired trait at the seedling stage, avoiding the need to grow undesirable plants to maturity (saving time and space), or to distinguish true-breeding homozygotes from heterozygotes, cutting generations from the breeding cycle. This general approach could be applied in principle to quickly seek diagnostic markers for a wide range of additional traits, including other aspects of plant development (flowering time, for example), resistance or tolerance to biotic (diseases and pests) or abiotic (drought, soil conditions) stresses, grain quality, and others.

ii) *Whole-genome searches*. While functionally-known genes may explain some important mutations in less-studied crops, the majority of differences among plant genotypes within a species are thought to be determined by the collective effects of numerous genes, each one of which has only a small effect. Such genes, often referred to as quantitative trait loci, have been the subject of literally thousands of studies in plants and animals over the past two decades. While there is some predictive value of the locations of QTLs for some traits across taxa, QTL mapping most commonly involves searching an entire genome with DNA markers at closely-spaced intervals to identify genomic regions that make the largest contributions to the genetic control of a complex trait.

The fact that most flowering plants share most of their genes, and that gene sequences change relatively slowly because proper gene function is intolerant of many changes, provides a means to design the large numbers of pan-taxon DNA markers that are required for genome-wide surveys of plant genomes, even in plants for which no *de novo* sequence information exists. As for the single 'Green Revolution' gene above, one can computationally scrutinize tens of thousands of genes to identify subsets that have islands of sequence shared by a wide range of taxa. Such islands of conserved provide targets for designing synthetic oligonucleotide primers that can be used by the PCR to amplify corresponding sequences from a wide range of taxa, even those lacking sequence information. Such cross-utilization of genomic tools to study genetic diversity requires resolution of a fundamental conflict between the need to identify genomic sequences that are conserved (largely or wholly) across many divergent taxa, and the need to identify DNA-level differences that reflect diversity at its most elemental level. The relatively high level of conservation of the locations [44], but not the sequences of introns, provides a resolution to this dilemma. The identification of vast numbers of probable gene and intron locations in the sequences of botanical models is routine. 'Conserved intron scanning primers (CISP)' within relatively conserved exons located near exon-intron boundaries, can be used to scan introns for variation suitable for DNA marker identification. permitting systematic sampling of entire genomes for well-distributed markers, or targeted enrichment of particular regions containing a gene of interest. This approach has been particularly well studied in the monocots, using information from rice and sorghum

to identify DNA markers useful in pearl millet and other less-studied cereals [43], as well as in non-cereal monocots [45]. Variations on the same method have been applied to many other plant families (legumes, crucifers, nightshades) and in principle are generally applicable to all flowering plants, as well as animals.

iii) Deductions about probable gene repertoire and function. Let us take the funde example a step further. Let us assume that the trait could NOT be accounted for by a homolog of the Gren Revolution gene. Accordingly, a whole-genome scan was done and the trait associated with several genetic markers on a different chromosome. The genetic markers all show correspondence a small region of same rice chromosome, that contains a different gene known to confer reduced height and increased harvest index. In other words, the initial candidate gene did not explain the phenotype but other candidates of known function from different regions of the genome may do so. As the number of genes for which functions are known increases, naturally the number of candidates to consider in such comparative studies does, too.

Limitations – Lineage specific genes

Plants are selected to become crops because of some unique feature(s) that attracted indigenous peoples, that were able to be improved by either conscious or unconscious selection. As such, each of our crops becomes a sort of botanical model for some extreme feature of plant growth and development. To understand and manipulate the features that make a crop unique, i.e. for which we cannot draw upon analogy to other plants, will require crop-specific enabling tools, technologies, and resources; in particular targeting genes that are substantially different from those of other organisms. How will we recognize the genes that confer these features, and how will we determine how they work?

A curious finding in virtually all eukaryotic genomes sequenced to date is 'lineage-specific' genes, for which an ortholog cannot be discerned in closely related species. Among 18447 'deduced ancestral loci' in Arabidopsis, poplar, grape, papaya and rice [46a], 3680 (20%) were specific to only one species. Among OrthoMCL gene families in sorghum, 3,983 (24%) appear to be absent from the dicot genomes sequenced to date; and 1,053 (6.4%) appear to be absent from the other monocot genome, rice. One might expect lineage-specific genes to be relatively frequent in angiosperms due to recurring polyploidization/diploidization cycles [46a] that do increase regulatory and may increase morphological complexity [46b]; and to transposition of gene segments that may evolve new genes [47, 48], albeit rarely [49]. Nonetheless, recent analyses of well-groomed mammalian genomes [50] report higher lineage-specific gene frequencies than found in sorghum versus rice.

Lineage-specific genes are a tantalizing target for early functional analysis because they may perhaps relate to features that differentiate closely related taxa, but caution is especially warranted in selecting such genes for analysis. Rapid gene evolution may be due to a lack of structural or functional constraint (indicating a *lack of important function*) or to strong positive selection for functional divergence (indicating *important new function*), possibilities that can be distinguished statistically [51, 52, 53, 54]. Genes under strong positive selection in *Drosophila*, mammals, and several other species are vital to reproductive success, cell-cell recognition, and cellular response to pathogens [e.g., 45, 55, 56]. However, identifying lineage-specific genes is inherently error-prone. Transposable-element associated genes have been a major contributor to inflated gene number estimates and false positive inferences of lineage-specific angiosperm genes [57]. Whole-genome shotgun sequencing approaches (often including many gaps), and occasional errors in accurate inferences of intron-exon boundaries, each may artifactually truncate or elongate gene models. Indeed, gene annotations of virtually all angiosperm genomes, even *Arabidopsis* which has been scrutinized by hundreds of scientists for nearly a decade, remain works in progress and different annotations of even one genome invariably differ. While much may be learned from functional analyses of lineage-specific genes, their careful manual annotation should precede investments in their analysis. Each additional genome available to study using comparative approaches further improves power to expose false-positive cases of apparent lineage-specific genes, for example by gross phylogenetic incongruities in inferred gene sets [19].

Exemplary opportunities to increase knowledge of African crops

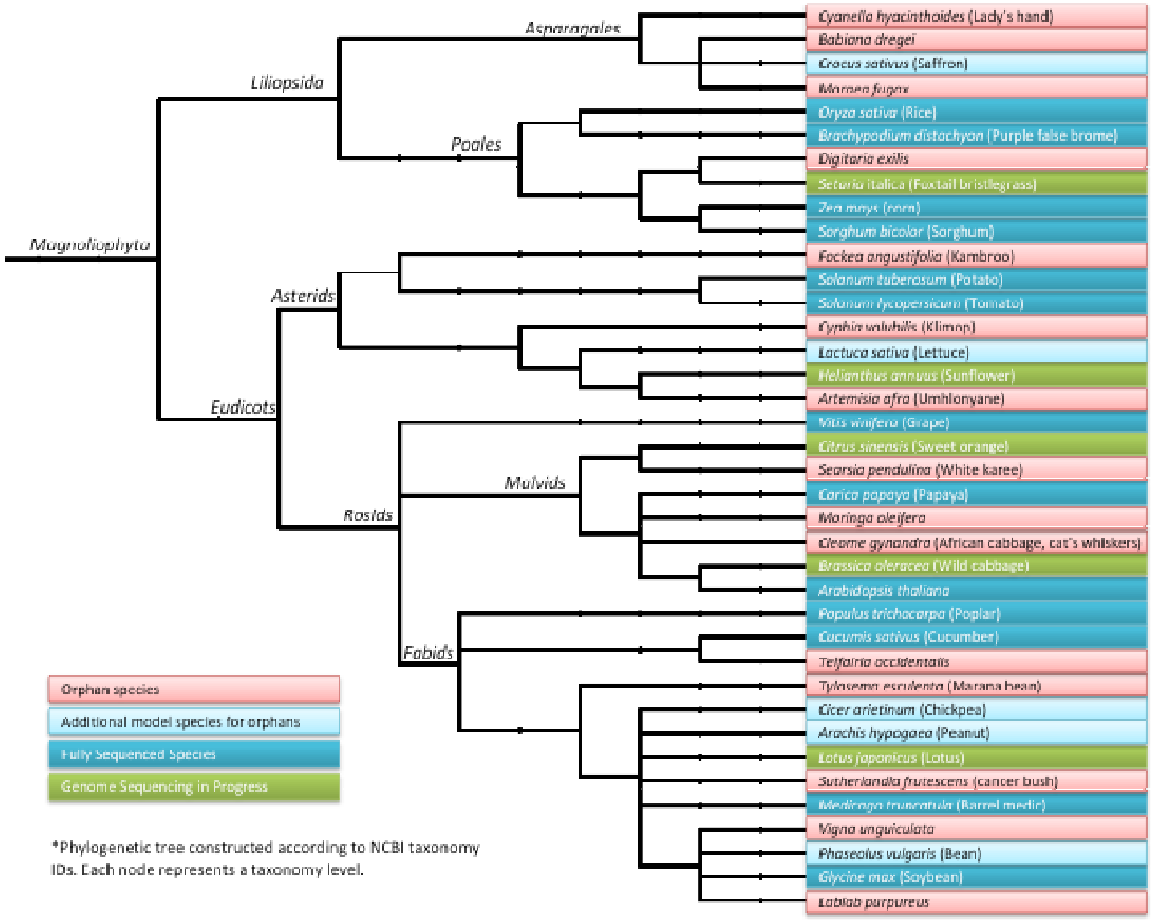
While a host of indigenous and introduced crops would benefit from comparative approaches to their improvement, below and in **Figure 1** we briefly offer a few exemplary cases. For example, *Cleome gynandra*, a leaf vegetable from South Africa, and *Moringa oleifera*, introduced from India but widely grown in Africa and referred to in a prominent US National Academies report [58] as "a sort of supermarket on a trunk", are both members of the plant order Brassicales which includes the fully-sequenced models *Arabidopsis thaliana, Carica papaya*, and *Brassica rapa* (turnip). In addition, extracts from the leaves of *Cleome* spp. have demonstrated insecticidal, repellent, antifeeding properties [59,

60a] displaying their potential use as additives in insect control agents [60b].

Digitaria exillis, noted above, is a panicoid cereal as are fully-sequenced sorghum and maize, and *Setaria italica* (Foxtail millet) for which sequencing is in progress. *Telfairia occidentalis, a leaf vegetable, is a member of the* Cucurbitaceae, closely related to the fully-sequenced cucumber.

Numerous nitrogen-fixing legumes are important indigenous crops, perhaps reflecting the relatively nitrogen-poor soils of much of Africa, including *Tylosema esculenta, Vigna unguiculata*, and *Lablab purpureus*, *all of which might derive considerable benefit from the fully-sequenced genome of Glycine (soybean), and expected*

Figure 1. Examples of orphan crops and their relationships to botanical models and/or other crops which have information that might be utilized for translational genomics.



sequences for Medicago (alfalfa), Lotus, Phaseolus (common bean), and Arachis (groundnut or peanut). The large seeds of *Tylosema esculanta* (marama bean in English, ombanui in Herero) are eaten in South Africa, Botswana and northern Namibia. This is seen as a significant crop for arid regions [61] with comparable protein content to existing crops. It is also being developed in Texas, Australia and Israel [15]. *Sutherlandia frutescens* (cancer bush in English, musa-pelo in Sesotho) is a widely used indigenous medicinal crop in South Africa, Botswana, Lesotho and Namibia [62]. *In vitro* studies indicate anticancer activity [63], and antidiabetic effects [64].

Study of *Searsia pendulina* (White karee), formerly *Rhus pendulina*, might benefit from analogy to the relatively small genome of members of the genus Citrus (orange). The fruits of S. *pendulina* are mixed with the gum of *Acacia karoo* to form a high energy (1202 kJ/100 g) confection [65]. Several other asterids, *Fockea angustifolia, Cyphia volubilis,* and *Artemisia afra* might benefit from a planned genome sequence for Helianthus (sunflower), and to a lesser degree the fully-sequenced genomes of more distantly-related tomato and potato.

Several corm-producing members of the Asparagales such as *Cyanella hyacinthoides* , *Moraea fugax* and *Babiana dregei* are highly regarded as staple foods in the drier parts of Southern Africa [65]. Other tuber species that are consumed during periods of drought include several *Fockea* species such as *F. angustifolia medulis* and to a lesser extent *F. camaru* and *crispa.* The watery, white flesh is consumed by inhabitants of the Kalahari desert when no surface water is available [15]. While these may benefit from the albeit limited EST resources available for the Asparagales taxon saffron crocus, the lack of genomic information for this clade points to the need for greater investment in its genomic characterization.

What else might be out there? (species and alleles)

About 250,000-400,000 angiosperm species are thought to exist [3, 4, 5], and a remarkable average of 2,350 new ones are discovered each year [6]. The vast majority of such newly-identified species are either confined to small areas, or are exceedingly rare, although one recent discovery in Eastern Ethiopia dominates the vegetation over at least 8000 km^2, an area nearly the size of the island of Crete [66]. This area has been seldom visited by botanists because of its inaccessibility and general unrest, and the newly-named plant, *Acacia fumosa*, does not seem to have any noteworthy uses, but it begs the question of what else might be found in such 'terra incognita' [67]. Even plants that lack obvi-

ous uses to researchers and non-native people may harbour sources of novel 'versions' (alleles) of genes with effects that are useful or informative, and which are ever-more accessible to discovery through genomics.

More immediate opportunities lie in the deeper exploration of the gene pools of indigenous African plants, and greater utilization of their full genetic potential. A few better-studied crops of African origin, such as sorghum, are well represented in the world's 'gene banks', and many of the extant alleles in the species will be discovered in due course by 'resequencing' projects. Nonetheless, there are various known gaps and probably additional unknown gaps in collections that warrant further sampling. Moreover, association of rare alleles with their functions would be expedited by access to greater numbers of genotypes containing the alleles. Many of the 'orphan' crops lack large-scale collections, or even any collections. Indigenous communities accumulate immeasurable knowledge of the various uses of the flora and fauna they have co-existed with for millennia, and the documentation and cataloging of such indigenous knowledge may lead to identification of new ecosystem services and their causal genes.

A further challenge is that restrictions on flow of germplasm due to intellectual property issues and the privatisation of agricultural biotechnology often hinder what could be very fruitful partnerships between African and non-African scientists to bring the full spectrum of biotechnologies to bear on these crops. Transfer of cutting edge/ frontier technologies which would empower researchers in developing countries to advance independently has often lagged transfer of germplasm from developing to developed countries [68]. With the privatization of agricultural biotechnology, scientists in technology rich (developed countries) and gene rich (developing countries) find themselves under the cooperation—comptition paradox [69]. Indigenous people have in the past cooperatively exchanged germplasm and information on plant use with researchers and managers, with little reciprocal exchange in technology that would allow the development of profitable products from such plants. The way forward requires ethical and responsible actions from both sides which acknowledge the global forces at play in the arena within which scientists practice.

References

1. Sanderson MJ, Thorne JL, Wikstrom N, Bremer K (2004) Molecular evidence on plant divergence times. American Journal of Botany 91: 1656-1665
2. Bell CD, Soltis DE, Soltis PS (2005) The age of the angiosperms: A molecular timescale without a clock. Evolution 59: 1245-1258
3. Thorne RF (2002) How many species of seed plants are there? Taxon 51: 511-522
4. Govaerts R (2003) How many species of seed plants are there? - a response. Taxon 52: 583-584
5. Scotland RW, Wortley AH (2003) How many species of seed plants are there? Taxon 52: 101-104
6. Prance GT, Beentje H, Dransfield J, Johns R (2000) The tropical flora remains undercollected Ann. Miss. Bot. Gard. 87: 67-71
7. Clement CR (1999) 1492 and the loss of Amazonian crop genetic resources. I. The relation between domestication and human population decline. Economic Botany 53: 188-202
8. Raven P, Evert R, Eichhorn S (1992) Biology of Plants. Worth Publishers, Inc., New York
9. Thro AM, Zankowski P (2003) Classical plant breeding is the route to food security. Nature 422: 559
10. Anonymous (1997) Sahel Region Program. *In*. USAID
11. Anonymous (2002) World agriculture: towards 2015/2030. *In*. FAO, Rome
12. Naylor RL, Falcon WP, Goodman RM, Jahn MM, Sengooba T, Tefera H, Nelson RJ (2004) Biotechnology in the developing world: a case for increased investments in orphan crops. Food Policy 29: 15-44
13. Van Wyk BE, Van Oudtshoorn B, Gericke N (2009) Artemisia afra. *In* Medicinal Plants of South Africa. Briza Publication, Pretoria, South Africa
14. Kinghorn AD, Balandrin MF (1993) Human medicinal agents from plants. *In* ACS Symposium Series, Vol 534, Washington D.C.
15. Van Wyk BE, Gericke N (2000) People's plants; A guide to useful plants of Southern Africa. BrizaPublications, Pretoria, South Africa
16. Liu NQ, Van der Kooy F, Verpoorte R (2009) Artemisia afra: A potential flagship for African medicinal plants? South African Journal of Botany 75: 185-195.
17. DST (Department of Science and Technology) (2009) 2009-2010 Corporate Strategy. http://www.dst.gov.za.

18a. Ketema S (1997) Promoting the conservation and use of underutilized and neglected crops. *In*. Institute of Plant Genetics and Crop Plant Research, Gatersleben/ International Plant Genetic Resources Institute, Rome, Italy, p 50

18b. Qi X, Pittaway TS, Lindup S, Liu H, Waterman E, Padi FK, Hash CT, Zhu J, Gale MD, Devos KM (2004) An integrated genetic map and a new set of simple sequence repeat markers for pearl millet, Pennisetum glaucum. Theor Appl Genetics 109: 1485-1493

19. Paterson A, H., Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Lyons E, Maher C, Narechania A, Penning B, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein PE, Kresovich S, McCann MC, Ming R, Peterson DG, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS (2009) The Sorghum bicolor genome and the diversification of grasses Nature 457: 551-556
20. Vavilov N (1922) The law of homologous series in variation. J Genet 12: 1
21. Doyle JA, Donoghue MJ (1993) Phylogenies and angiosperm diversification. Paleobiology 19: 141-167
22. Crane PR, Friis EM, Pedersen KR (1995) The Origin and Early Diversification of Angiosperms. Nature 374: 27-33
23. Bonierbale M, Plaisted, RL, Tanksley, SD (1988) RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. Genetics:

1095-1103

24. Paterson AH, Lan TH, Reischmann KP, Chang C, Lin YR, Liu SC, Burow MD, Kowalski SP, Katsar CS, DelMonte TA, Feldmann KA, Schertz KF, Wendel JF (1996) Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. Nature Genetics 14: 380-382

25. Tang H, Bowers JE, Wang X, Paterson AH (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proceedings of the National Academy of Sciences of the United States of America 107: 472-477

26. Initiative TAG (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796-815

27. Tuskan GA, DiFazio S, et al (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313: 1596-1604

28. Jaillon O, Aury JM, et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 463-467

29. Velasco R, Zharkikh A, et al (2007) A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. Plos One 2

30. Ming R, Hou S, Feng Y, et al (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452: 991-997

31. Huang S, Li R, Zhang Z, Li L, al e (2009) The genome of the cucumber, Cucumis sativus L. Nature Genetics 41: 1275-1281

32. Schmutz J, Cannon SB, Schlueter J, Ma JX, Mitros T, Nelson W, Hyten DL, Song QJ, Thelen JJ, Cheng JL, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu SQ, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du JC, Tian ZX, Zhu LC, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463: 178-183

33. Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchinson D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong JP, Miguel T, Paszkowski U, Zhang SP, Colbert M, Sun WL, Chen LL, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu YS, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (Oryza sativa L. ssp japonica). Science 296: 92-100

34. Yu J, et al (2002) A draft sequence of the rice genome (Oryza sativa L. ssp indica). Science 296: 79-92

35. Matsumoto T, et al (2005) The map-based sequence of the rice genome. Nature 436: 793-800

36. Yu J, et al (2005) The Genomes of Oryza sativa: A history of duplications. Plos Biology 3: 266-281

37. Schnable PS, et al (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326: 1112-1115

38. Initiative TIB (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463: 763-768

39. Bowers JE, Arias MA, et al (2005) Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. Proceedings of the National Academy of Sciences of the United States of America 102: 13206-13211

40. Lieven Sterck SR, Klaas Vandepoele, Pierre Rouze and Yves Van de Peer (2007) How many genes are there in plants (... and why are they there)? Current Opinion in Plant Biology: 199-203

41. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 13: 2178-2189

42. Peng JR, Richards DE, Hartley NM, Murphy GP, Devos KM, Flintham JE, Beales J, Fish LJ, Worland AJ, Pelica F, Sudhakar D, Christou P, Snape JW, Gale MD, Harberd NP (1999) 'Green revolution' genes encode mutant gibberellin response modulators. Nature 400: 256-261

43. Feltus FA, Singh HP, Lohithaswa HC, Schulze SR, Silva T, Paterson AH (2006) Conserved intron scanning primers: Targeted sampling of orthologous DNA sequence diversity in orphan crops. Plant Physiology 140: 1183-1191

44. Quax-Jeuken Y, Quax W, van Rens G, Khan PM, Bloemendal H (1985) Complete structure of the alpha B-crystallin gene: conservation of the exon-intron distribution in the two nonlinked alpha-crystallin genes. Proc Natl Acad Sci U S A 82: 5819-5823

45. Lohithaswa HC, Feltus FA, Singh HP, Bacon CD, Bailey CD, Paterson A, H. (2007) Leveraging the rice genome sequence for comparative genomics in monocots. Theoretical and Applied Genetics 115: 237-243

46. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH (2008) Unraveling Ancient Hexaploidy through Multiply-aligned Angiosperm Gene Maps. Genome Research 18: 1944-1954

46. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008) Synteny and colinearity in plant genomes. *Science* 320: 486-488

46a. Freeling M, Thomas BC (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Research 16: 805-814

47. Jiang N, Bao ZR, Zhang XY, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569-573

48. Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Current Opinion in Genetics & Development 15: 621-627

49. Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE (2005) The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. Genome Research 15: 1292-1297

50. Church D, Goodstadt L, Hillier L, Zody M, Goldstein S, She X, Bult C, Agarwala R, Cherry J, DiCuccio M, Hlavina W, Kapustin Y, Meric P, Maglott D, Birtle Z, Marques A, Graves T, Zhou S, Teague B, Potamousis K, Churas C, Place M, Herschleb J, Runnheim R, Forrest D, Amos-Landgraf J, Schwartz D, Cheng Z, Lindblad-Toh K, Eichler EE, Ponting CP, Consortium TMG (2009) Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse. PLOS Biology 7: e1000112

51. Yang ZH (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences 13: 555-556

52. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929-936

53. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15: 568-573

54. Yang ZH, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431-449

55. Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proceedings of the National Academy of Sciences of the United States of America 98: 7375-7379

56. Swanson WJ, Zhang ZH, Wolfner MF, Aquadro CF (2001) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proceedings of the National Academy of Sciences of the United States of America 98: 2509-2514

57. Bennetzen JL, Coleman C, Liu RY, Ma JX, Ramakrishna W (2004) Consistent over-estimation of gene number in complex plant genomes. Current Opinion in Plant Biology 7: 732-736

58. Council NR (`2006) Lost Crops of Africa. Volume II: Vegetables. National Academies Press, Washington, DC

59. Fletcher R (1999) Cleome gynandra (Cat's whiskers). The Australian New Crops Newsletters 11

60a. Lazzeri L, Leoni O, Manici LM (2004) Biocidal plant dried pellets for biofumigation. Industrial Crops and Products 20: 59-65

60b. Somboom S, Pimsamarn S (2006) Biological activity of Cleome spp.extracts against the rice weevil, Sitophilus oryzae L. Agricultural Science Journal 37: 232-235

61. Boshid A (1979) Tropical legumes: Resources for the Future. National Academy of Sciences, Washington DC, USA

62. Moshe D (1998) A biosystematic study of the genus Sutherlandia R. Br. (Fabaceae, Galegeae). University of Johannesburg

63. Tai J, Cheung S, Chan E, Hasman D (2004) In vitro culture studies of Sutherlandia frutescens on human tumor cell lines. Journal of Ethnopharmacology 93: 9-19

64. Chadwick WA, Roux S, de Venter MV, Louw J, Oelofsen W (2007) Anti-diabetic effects of Sutherlandia frutescens in Wistar rats fed a diabetogenic diet. Journal of Ethnopharmacology 109: 121-127

65. Archer FM (1990) Planning with people: Ethnobotany and African uses of plants in Namaqualand, South Africa. Mitteilungen aus dem Institut fuer Allgemeine Botanik Hamburg 23: 959-972

66. Thulin M (2007) Acacia fumosa sp nov (Fabaceae) from eastern Ethiopia Nordic Journal of Botany 25: 272-274

67. Mabberley DJ (2009) Exploring Terra Incognita. Science 324: 472

68. Sori B, Wilkinson J (1994) Biotechnologies, multinationals and the agrofood systems of development countries. *In* A Bonanno, L Busch, W Friedland, L Gouveia, E Mingione, eds, From Columbus to ConAgra: the globilisation of agriculture and food. University of Kansas, Lawrence, Kansas, pp 85-101

69. Silva JdS (1997) Agricultural biotechnology transfer to developing countries under the cooperation-competition paradox. Cardenos de Ciência & Technologia, Brasilia 14: 91-112