Tailoring Bioinformatics for the Genetic Improvement of Orphan Crops

Melaku Gedil
International Institute of Tropical Agriculture (IITA), PMB 5320, Oyo Road, Ibadan, Nigeria
(Email: m.gedil@cgiar.org)

Abstract

The advent and rapid advance of genomics technologies can enable plant breeders to design cost-effective and efficient breeding strategies by exploiting the ability of molecular plant breeding to increase favorable gene action and efficiency of selection. The unparalleled scientific progress in the fields of genomics and bioinformatics can successfully be harnessed to address the challenges of small holding farmers in developing countries. The power of molecular breeding extends to orphan crops with little DNA sequence information through comparative genomics methods. This growing abundance of genomic resources necessitates that plant scientists be equipped with fundamental genomic analysis tools for genomics assisted crop improvement. The role of bioinformatics as a pivotal tool for molecular breeding is growing steadily, particularly in identification of nucleotide variants associated with key traits. Basic bioinformatics skills to utilize selected public databases and integrated resources are outlined. Online resources for self-paced tutorials and other skill building opportunities were suggested. Particular emphasis was made to comparative genomics techniques to develop genomic resources for molecular breeding. Research institutions in developing nations should invest in bioinformatics capacity building in terms of human resources and infrastructure development in addition to forging strong partnerships with advanced research institutes.

Key words: molecular breeding, comparative genomics, database, bioinformatics, orphan crops, markers

Introduction

Agriculture, the main stay of Africa's economy and livelihood, is beset by a web of interacting and interrelated factors, exacerbated by climate change, posing a threat to food security which calls for innovative and effective breeding strategy.A number of recent reviews have provided detailed account of how the advent of genomics and its derived 'omics' technologies can enable plant breeders to design cost-effective and efficient breeding strategies by exploiting the ability of molecular plant breeding to increase favorable gene action and efficiency of selection among other things [15, 29]. The rapid accumulation of genomic data and the ensuing development of functional genomics techniques, tools, and databases ushered the era of molecular breeding as a new paradigm [51]. Numerous powerful molecular tools have been and are being developed to understand fundamental processes underlying key physiological traits desired for germplasm enhancement [15]. A wide variety of markers have been developed and progressively improved for cost-effectiveness, efficiency, and increased throughput. Nucleotide variation in the forms of SNP and SSR have been broadly utilized to study genetic diversity and to genetically map traits of economic importance across a wide range of crops [1, 2, 16]. The unparalleled scientific progress in the fields of genomics and bioinformatics can successfully be harnessed to address the challenges of small holding farmers in developing countries where orphan crops are grown as staple food or cash crops. Given the meager agricultural input in developing countries genetic improvement is the most plausible option to raise crop productivity for the resource-poor farmers. The advent of new technologies in molecular biology and the parallel evolution of bio-computational tools offer broader opportunities for devising an efficient and effective breeding strategy. In order to extend the power of molecular breeding to orphan crops with little DNA sequence information, plant scientists should be equipped with fundamental genomic analysis tools including comparative genomics. This paper reviews selected bioinformatics tools, databases, and services suitable for plant biologists engaged in improvement of under-researched crops. An attempt has been made to provide a flavor of potential application of bioinformatics databases and tools for a novice molecular breeder in the developing country, taking into account the limited resources and infrastructure in most national agricultural research institutions. While focus is on orphan crops, breeders working on non-orphan crops such as maize, soybean, and rice are also urged to start applying these techniques in their breeding scheme with earnest.

Plant Genome Projects

The completion of genome sequences of the model plant *Arabidopsis* and the first crop plant, rice, heralded the dawn of the genomics era. Following these landmark achievement, the research community is aggressively taking on the challenges of integrating molecular breeding into the existing breeding programs [13]. Knowledge of the genome sequence of plants is of paramount importance in understanding the physiological processes underlying plant traits which can be manipulated to create desirable cultivar. The technology of genome sequencing has dramatically improved as evidenced by the steadily growing amount of genomic information and the completion of vast number of organisms [6, 7, 26]. In fact, with the current trend of rapid development of sequencing technology, it will not be too long before the genome sequence of all agricultural plants will be determined.

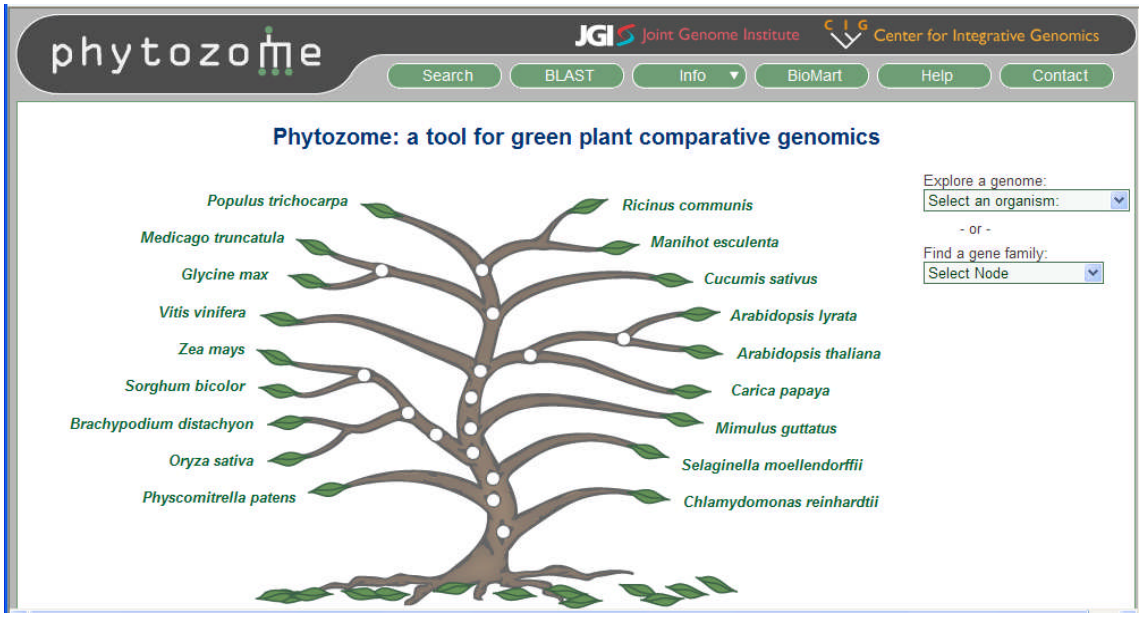
Current views on opportunities for tackling the challenge of food security vis-à-vis the increasing world population

and climate change range from exerting the power of science to break the yield barrier [35] to broadening the wheat-rice-maize dominated source of food by improving underutilized crops [14]. The aforementioned major crops have undergone substantial extensive genomic investigations. According to the NCBI-GenBank release note 176.0 (Feb. 2010), six plant species are among the 20 most sequenced organisms based on the number of entries and bases of nucleotide sequences in GenBank/EMBL/DDBJ. The most sequenced plant species, maize (*Zea mays*), has 3.9 million nucleotide entries followed by rice (*Orzya sativa* Japonica group; 1.2 billion bases). In comparison, crops listed in **Table 1** have a fraction of the amount of genomic resources available for the major crops and model plants such as *Arabidopsis*. For instance, the number of entries

Table 1. Available genomic resources at the NCBI databases for selected orphan crops and alternative resources for comparative studies

Crops	Cow-pea	Pea-nut	Grass Pea	Chick Peas	Tef	Finger millet	Cof-fee	Rape seed	Sun-flower	Yam	Lentil	Ba-nana Musa sp
Genus	Vigna	Arachis	Lathy-rus	Cicer	Er-agrosti s	Eleusin e	Coffea	Bras-sica	Helian-thus	Dio-scorea	Lens	Musa
Family	Fa-baceae	Fa-baceae	Fa-baceae	Fa-baceae	Poacea e	Poacea e	Rubia ceae	Brassi-caceae	As-teracea e	Dio-scorac eae	Fa-baceae	Musac eae
Related species	Soy-bean/ Lotus	Soy-bean/ Lotus	Soy-bean/ Lotus	Soy-bean/ Lotus	Rice/ maize	Rice, Maize	Dicot	Arabi-dopsis	Lettuce	Mono-cots	Soy-bean/ Lotus	Mono-cots
Resources	LIS	LIS	LIS	LIS	Multi-ple	Multiple	GPD	TAIR*	Compo-sitdb	PlantG DB	LIS	PlantG DB
Nucleotide	507	1,767	25	934	564	188	680	10,817	9,448	683	184	4,210
Nucleotide EST	187,48 3	87,002	178	34,208	2,816	1,927	43,61 9	643,94 3	133,68 2	31	9,513	31,268
Nucleotide GSS	54,123	9,347		50,853	40		3,875	102,61 9	573		485	7,186
Protein	365	966	22	756	6	65	453	9,711	5,575	622	127	2,510
Structure	8	28	1			3		7	5		5	6
Genome Sequences				1			1	2	1	1		
Genome Projects	1			1	1	1	1	2	1	2	1	1
Popset	32	16	6	6	5	8	17	8	197	54	2	34
3D Domains	29	124	4			4		10	5		36	10
GEO Datasets	5	3	11	14				22	2		9	3
UniGene	15,740	11,909						27,139	12,216			
UniSTS	75	203	5	54	4	42	17	284	1,627	4		69
PubMed Central	645	443	27	217	24	34	142	1,514	693		20	265
Gene				108			140	106	138		213	
Taxonomy	11	3	1	1	1	3	1	3	1		4	77

Figure 1. Version 5.0 of Phytozome comprises twenty genomes (sequenced at the Joint Genome Institute and other institutions) the phylogenetic relationships of the species to facilitate comparative genomic studies.



Source: www.phytozome.net

for the multi species Genus *Dioscorea*, which consists of such important cultivated yam species as water yam (*D. alata*, 31 sequences), yellow yam (*D. cayenensis*, 6 sequences), and white yam (*D. rotundata*, 3 sequences) is less than a thousand GenBank sequences in total. The paucity of genomic resources considerably hampers the application of marker assisted breeding in orphan crops. However, advances in technology and ultra high throughput genotyping technologies are changing the landscape of genomic research.

The list of completed genome sequences is publicly available for several crops and tree plants at the Entrez Genome database (http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html) and at the Phytozome web site (www.phytozome.net). **Figure 1** displays the list of crops with complete draft genome in the database of Phytozome. The resources cover different families of green plants representing cereals, legumes, fruits, vegetable, root crop, and tree plants, the majority of these being completed in the past few years. Mega initiatives such as "The 1000 Plant/Animal Genomes De novo Sequencing Project" of the Beijing Genomics Institute (BGI; http://www.genomics.cn/en/) is aiming to sequence 500 plant and 500 animal species.

Genomics for marker development and gene discovery

The new technologies furnished a new set of molecular markers that are amenable for high throughput discovery and genotyping [2]. Besides the number of completed and ongoing genome sequencing projects, numerous large-scale plant EST sequencing projects were launched to generate molecular data that can be used for marker development (http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html#IP_SEQ). As a result of these massive genomic data, there is a trend of shifting from the first generation DNA-based markers such as RFLP, RAPD, and AFLP towards functional or gene-targeted markers such as EST-derived SSRs and single nucleotide polymorphism (SNP) markers [1]. Preference for SNP is largely driven by the cost-effectiveness of its discovery, availability of high throughput genotyping technologies, and the range of applications for genetic and genomic studies [4]. One of the notable outcomes of developments in genomics is the rapid rate of discovery and characterization of novel genes. Advances in functional genomics have begun shedding light on the mechanism underlying physiological processes relevant to agricultural productivity thereby elucidating a repertoire of stress-induced genes and pathways, pathogen defense genes, nutritional quality traits such as pro-vitamin A carotenoids [21]. Knowledge of genes and pathways opened a new avenue of research for enhanced efficiency and effectiveness of breeding programs through development of gene targeted markers (GTM) and functional markers (FM) [1, 51]. The potential use of markers for germplasm management, trait conversion, and trait stacking and pyramiding are becoming widely accepted.

There are large number of orphan crops that are neglected by global research community and the private sector owing to their negligible or unknown importance outside their locality or region [3]. The CGIAR, a

* (Table 1) Cassava not included because genome has been sequenced; General Plant Databases (NAR vol 38, 2010)

Legume: In addition to soybase, Legume Information center.

Cereals/grasses: Graingenes, gramene, plantGDB

coalition of 15 research centers (www.cgiar.org), whose mission is to achieve sustainable food security and reduce poverty in developing countries through scientific research and research-related activities, strives to improve 19 crops of which 12 could be considered as orphan crops including millet, chickpea, lentil, pigeon pea, cassava, yam, cowpea, sweet potato, plantain and banana, coconut, and groundnut. Among these, the draft genome of cassava has been released recently triggering several projects in its wake (http://www.phytozome.net/cassava.php).

Below are some useful resources for plant genomes and comparative analysis

⇒ *GreenPhylDB* (http://greenphyl.cirad.fr/): a platform for full genome comparison of Arabidopsis and rice but also includes GOST (GreenPhyl Orthologs Search Tool) which assists the identification of orthologous and paralogous genes for any plant gene.

⇒ *PlantGDB* (http://www.plantgdb.org/): is a comprehensive resource for comparative plant genomics with tools and tutorials for downloading, comparing, and annotating sequences.

⇒ *The Genome On Line Database, GOLD* (http://www.genomesonline.org/): ar comprehensive access to information regarding complete and ongoing genome projects, as well as metagenomes and metadata. This tool allows viewing of species phylogenetically which is very important in orphan crops to find genomic-rich related plant.

⇒ *Phytozome* (http://www.phytozome.net/): Release version 5.0 of Phytozome features 20 sequenced and annotated plant genomes which have been clustered into gene families at fifteen evolutionarily significant nodes (see **Figure 1**). Phytozome provides various tools for similarity search and viewing as well as links to external resources.

With the new technologies, numerous genotypic data are being generated in short period of time. The challenge, however, is how to utilize such genomic data to improve the efficiency and effectiveness of crop improvement strategy. Broadly, molecular breeding comprises three components: phenotyping, genotyping, and data management. For under-researched crops, the limited genomic information currently available as well as innovative molecular tools discovered in other crops could be adapted and incorporated into the existing breeding programs through comparative genome analysis. The paucity of genomic resources in orphan crops could hamper the implementation of molecular breeding. Below are basic bioinformatics end-user tools that are accessible to scientists in developing countries and that could help reduce the gap in genomic knowledge and pave the way for the application of molecular breeding.

Bioinformatics

Bioinformatics can be viewed broadly as the development and application of computational tools to acquire, store, organize, retrieve, and analyze large amount of biological information. In the context of this review, bioinformatics refers to the search, evaluation, and utilization of biological computational tools and databases for candidate gene discovery and/or marker development. The spectacular advances in genome sequencing and the subsequent generation of large amount of biological data triggered the development of tools for data management, visualization, integration, analysis, modeling, and prediction [38]. At the moment, the number of scientists involved in bioinformatics are too few to meet the increasing demand for tools and methods to make sense out of the mounting data. On the other hand, substantial number of biological scientists is not in a position to utilize the existing tools due to lack of information on the existence of such tools or the function of the tools. A number of recent books [34, 51] and review articles [31, 36, 38] are testimony to the growing importance of bioinformatics skill. Rhee *et al*. [38] described basic and vital areas of bioinformatics such as sequence analysis, transcriptomics, proteomics, ontology, and databases. More specific presentation of a set of tools and databases relevant to weed science was provided by Larrinua *et al*. [24]. More recently, Armstead *et al*. [3] discussed the challenges and opportunities of using bioinformatics in the improvement of orphan crops, represented by three forage crops. The focus of the present review is limited to orphan food crops of Sub-Saharan Africa (SSA).

The role of bioinformatics as a pivotal tool for molecular breeding is growing steadily, particularly in identification of nucleotide variants associated with key traits [25]. The first step towards variant discovery is the mining of data in public databases. Subsequently the retrieved data would be subjected to compare nucleotides, perform similarity search, deduce protein sequences, and understand the function of the protein. Many users are either unaware of the presence of myriads databases and tools or intimidated by the idea of getting into such bioinformatics research. Here, I provide highlights of the relevant databases and end-user tools and services that can be employed in the breeding of orphan crops with limited genomic resources.

In silico Marker development

Availability of nucleotide sequences is the prerequisite for the application of marker-assisted breeding. In the past two decades, numerous labs were engaged in generating molecular markers such as RFLP probes, AFLP, SSR, and SNP using laborious and capital intensive protocols. This was not affordable by many institutions in the developing countries. Even when funds are provided by charity organization/donors, the lack of skilled personnel and infrastructure hampers the introduction of molecular techniques. Nowadays, genome sequences and associated functional genomics studies have become the primary source of genomic resources for comparative genomics. Experts in data mining are able to perform *in silico* research to develop molecular

markers from public databases using a combination of search and computational techniques [2, 40].

The most popular contemporary sequence-based markers are SSR and SNP.

SSR: Since its advent in mid 1980 [47], SSRs have been used in a variety of applications and crops. Variability is generated when a sequence is amplified by a pair of primers flanking a mono-, di-, tri-, or tetra-nucleotide repeats due to variable number of repeats in different individuals. For small scale number of sequences, manual designing of primers is possible. For large number of sequences, manual prediction is not only cumbersome and time consuming but also error-prone. Several software packages were developed to identify SSRs and design flanking primers including FastPCR [22] and Repeatfinder [49].

SNP: The free Dictionary (http://encyclopedia.thefreedictionary.com/Single-Nucleotide+Polymorphism) defines SNP as "A single nucleotide polymorphism (SNP, pronounced *snip*), is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual). For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide. In this case we say that there are two *alleles*: C and T (bold font). Almost all common SNPs have only two alleles. For a variation to be considered as SNP, it must occur in at least 1% of the population".

The following sections introduce a novice bioinformatics user how to discover nucleotide variations such as SSR and SNP and convert them to molecular markers for gene-tagging, linkage mapping, diversity analysis and other applications. More information on high throughput marker discovery and genotyping can be obtained [2].

Basic skills in bioinformatics

Besides sequence-based markers such as SSR and SNP, researchers working with RAPD and AFLP, may require some basic skills in bioinformatics in such a situation as a RAPD and AFLP fragment/band cosegregating with an interesting trait and the investigator wants to convert this fragment into a PCR-based, single-locus specific genetic markers such as sequence characterized amplified region (SCAR) [33] or cleaved amplified polymorphic sites (CAPS) [23]. Such conversion entails basic bioinformatics end-user skills such as sequence editing, similarity search, primer design, among others. In this section, a brief guide and example of web-services and open source software is outlined.

Choosing databases and web servers

It is challenging for biologists to choose the right tool amid the explosive growth of web sites and tools. It is absolutely important that users distinguish between personal web sites and reliable peer-reviewed and up-to-date resources. The challenge is how to choose from the multitude of databases and tools available. A good starting point could be the collection of molecular biology databases published in the journal Nucleic Acid Research, the latest being volume 38, Database issue (http://www.oxfordjournals.org/nar/database/c/). The most recent update of molecular biology databases feature over a thousand databases of which several hundreds are on plants [12]. Two world renowned organizations, NCBI (The National Center of Biotechnology Information) and EMBL (European Molecular Biology Laboratory), provide access to a comprehensive and integrated collection of biological data worldwide. NCBI [43] maintains many database resources including primary nucleotide and protein sequences, derived databases, bibliography, books, software, and tutorials. GenBank, the nucleotide sequence database of NCBI [6], comprises nucleotide sequences for more than 300,000 organisms, submitted by individual laboratories and batch submissions from large-scale sequencing projects. Two other public databases with whom GenBank daily exchanges data with are the EMBL Nucleotide Sequence Database in Europe and the DNA Data Bank of Japan (DDBJ). Entrez, the query and retrieval system at NCBI can be used to access several linked and integrated databases including DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed. Equivalent comprehensive collection of databases and tools can also be found at the European Bioinformatics Institute (EBI) [7] web site (http://www.ebi.ac.uk/) including SRS, data integration platform for easy access to large amount of biological data across 100s of databases. While the primary nucleotide sequence databases are identical with GenBank, the data retrieval system or the user interfaces are different. Users' choice of databases is a matter of preference or ease of learning the tools. Furthermore, the Online Bioinformatics Resources Collection (OBRC) which contains annotations and links for 2681 bioinformatics databases and software tools have compiled 148 plant specific databases [10]. (http://www.hsls.pitt.edu/guides/genetics/obrc/plant). Other systems with integrated querying are BioMart [44] and PLAZA [37].

Sequence retrieval and manipulation

Finding sequences in one of the above public databases is basically the same. Searches begin with keywords, accession number, gene name, species name, etc. The Entrez search engine at NCBI, in addition to retrieving sequences, returns pre-computed lists of data elements such as related sequences, gene, protein, taxonomy, and others. Search can be performed in all databases or restricted to nucleotide in the drop down menu. The result can be displayed in different format or downloaded. The most common download format is FASTA format. Description of FASTA format at NCBI is as follows:

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column.

It is recommended that all lines of text be shorter than 80 characters in length. An example sequence in FASTA format is:

>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENR
T
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFY
CK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK

It should be noted that FASTA is just one of several formats that is accepted by many sequence analysis software. However, many of the software have a choice of input format and also allow exporting sequences in various formats. Further details on sequence format can be found in one of the resources listed below.

Sequence alignment

Sequence alignment is the prerequisite of virtually all forms of sequence analysis ranging from search, to assembly, and to phylogenetics. Various algorithms have been developed to produce optimal alignment, a topic which is beyond the scope of this review. It is suffice to know that many softwares have been developed to perform nucleotide or protein sequence alignments. Two examples of widely used open access softwares, namely BioEdit [19], and MEGA [46], are freely downloaded and installed with easy-to-understand user's manual. A pair of sequences or multiple sequences saved, for example in FASTA format, can be used as an input. However, sequence alignment can also be done on the web at one of the resources listed in this review (e.g. EMBL-EBI) using the ClustalW program or other methods.

Phylogenetics

Phylogenetic analysis is the basis of taxonomical and evolutionary studies. In the context of this paper, phylogenetic analysis is performed to cluster multiple sequences based on genetic distances. This is a broad topic and a subject of 100s of articles and books. A deluge of tools and web services can also be found online (e.g. http://evolution.genetics.washington.edu/phylip/software.html). For beginners, stand alone programs such as MEGA can do excellent job of phylogeny tree construction. In addition, web services such as EMBL-EBI provide similar tools.

Similarity search

Sequence comparison is essential for understanding evolutionary relationship between genes. The most common and widely used similarity search tool is BLAST (Best Local Alignment Search Tool [53]. BLAST is a set of programs used to compare a nucleotide or protein query sequence to all of the available sequence databases. NCBI and EBI provide many different types of BLAST. Information on how to access BLAST services on WWW, choosing the right type of BLAST, interpreting BLAST results, how to do batch BLAST jobs, and others can be found at NCBI-BLAST home page (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Primer design

There are several applications in which primer designing is required for marker development. Such cases include, but not limited to, retrieved sequences containing simple sequence repeats suitable for SSR marker development. Orphan crops lack sequence information in which comparative genomics approaches such as homologous sequences are used to design degenerate primers, or re-sequence the gene of interest. The most widely used program for primer designing is PRIMER 3.0 (http://frodo.wi.mit.edu/primer3/) with several versions of web interface. The web-site provides user-friendly web interface and user manual describing the underlying principle of the program.

Advanced Skills

The major areas of high-end bioinformatics include the development of databases and algorithms for multiple sequence alignment, analysis and annotation of various types of microarray platforms, high-density oligonucleotide chips, variety of mass spectrometry, and diverse platforms of next generation sequencing. Computer savvy researchers who aspire to become bioinformatics tool developer should consider learning a scripting language program such as Perl (community web site: http://www.cpan.org/). Some genomics tasks such as discovery of SNPs or SSR in thousands of sequences, filtering sequences with the target motif and designing assay reagent (e.g. primer), filtering the result of BLAST, and annotation of thousands of EST sequences is a daunting job. Programming skill allows automation of such large scale and complex jobs.

Comparative Genomics

Comparative and functional genomics tools greatly facilitate the transfer of knowledge from thoroughly-studied model plants to orphan crops. Discovery of genes involved in flowering in model plants such as *Arabidopsis* have been successfully utilized to identify homologous genes in garlic [39] and in cauliflower [41]. Comparative genomic analysis tools have been used to investigate functional diversification and evolutionary mechanisms of plant genes [27, 55]. Most importantly, these tools provide an insight into the biochemical mechanism underlying economically important traits such as lignin biosynthesis for biofuel researchers [52], the cellulose synthase superfamily [54], improvement of the balance of essential amino acids and starch quality and quantity [15]. In the cases of orphan crops, where sequence information is meager or lacking, a host of comparative genomics techniques can be employed to tap into the benefit of genomics advances [17]. In groundnut, for instance, several hundred SNPs were identified in the Conserved Cross-Legume Orthologs (unpublished data).

Investigators could identify curated databases of the genus, family, or other category to which the crop in

question belongs. **Table 1** provides an example of crops and relevant databases and potential information that can be obtained. Such databases may be very useful for scientists looking for markers, QTL information, etc. However, many of the other tools and databases could be searched for conserved genes such as NBS-LRR disease resistance genes, pVAC, drought tolerance, etc.

Identification of candidate disease resistance genes could be used as an illustration of the successful application of comparative genomics. Methods of enriching the repertoire of gene level knowledge in orphan crops can take approaches such as the identification of resistance gene analogs (RGAs) for identification of genes involved in plant defense [30]. This technique capitalizes on the presence of conserved regions of resistance genes for designing degenerate primers and isolating resistance gene homologues from different plant genomes using the polymerase chain reaction (PCR). This homology-based approach has led to the identification of thousands of partial sequences of NBS-LRR genes in a wide array of plant species [5, 8, 9, 18, 48]. In general, understanding the structure, localization, function, variation, and evolution of resistance genes will provide the basis for devising an efficient breeding strategy for disease resistance [20, 45]. The RGA techniques can easily serve as an entry point to bioinformatics in which beginners can retrieve sequences, design primers, amplify candidate R gene regions, and characterize it by similarity search and other sequence manipulation tools. The database 'PRGdb' [42] provides a manually curated database of well characterized and candidate plant disease resistance genes belonging to nearly two hundred plant species. Users can download reference genes of interest to design degenerate primers to amplify homologous genes in their species of interest or simply follow the various links provided for further information on domains, motifs, bibliography.

Capacity building opportunities

A recent article on mobilizing science to break yield barrier, emphasized the role of emerging technologies could play in improving agricultural productivity in Sub-Saharan Africa and South Asia [36]. It also advocates investing in human resources and briefly discussed the effort of CGIAR and the donor community towards training young scientists in developing economies. Bioinformatics has established itself as the cornerstone of modern molecular biology research. It is vital to initiate various forms of training, through non-credit courses and workshops, as well as degree awarding academic programs [28] to keep developing countries scientists abreast of current technologies. Below are listed some avenues for training:

Free Online courses and tutorials

In the field of bioinformatics it is not uncommon to find free online courses such as one offered by S*Star, an alliance of eight universities, spanning five continents (http://s-star.org). Webinars on specific topic or analysis method are also available from the private sector. However, the most important resources to get started are tutorials provided by worldwide renowned institutions such as NCBI and EBI. Open access journals published by Public Library of Science (e.g. PLoS computational biology) and Biomed Central (e.g. BMC Bioinformatics) are a good source of full text articles as well as numerous tools and hyperlinks.

NCBI: The NCBI handbook, one of the 257 free online books available on the BookShelf, provides information about the various databases and tools available at the site (http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook).

2Can Support portal at EBI: offers a number of tutorials on different topics. http://www.ebi.ac.uk/2can/tutorials/index.html

Nucleic Acid Research, volume 38, Database issue contains list of curated databases as well as open access full text articles.

Tips on online search engines

While this seems trivial, nowadays, using the most popular search engines such as Google and Yahoo for any information could be frustrating and often unlikely to furnish the needed information. Filtering through the results can be a daunting task if appropriate search strategy is not used. Too often, the results could be unwanted aggressive marketing web sites with flashy popups or even indecent websites which have no relevance to the query word. While many web sites provide site-specific search engine, the following alternatives help for effective search of scientific terms; i) *Google Scholar*; ii) *Scirus* (www.scirus.com) is a comprehensive scientific research tools with options for filtering data and refining search results; iii) *Wikipedia* is an online free-content encyclopedia that anyone can edit and contribute to. Even though the quality of information is not good enough to be cited, it can help as a stepping stone through the links to primary sources when available (http://en.wikipedia.org/wiki/Main_Page); and iv) *Bioforums* such as Protocol-online.org, Biotechnique's Protocol wiki allow posting of questions in a specific subjects where members provide useful information thoes are hard to find otherwise.

Courses and workshops

African scientists should be on the look out for trainings by the CGIAR centers such as International Institute of Tropical Agriculture (IITA, www.iita.org), BecA-ILRI (http://hub.africabiosciences.org/), generation challenge program (GCP, http://www.generationcp.org/), and other centers (www.cigar.org). Bioinformatics trainings tailored for plant breeding had been organized by the International Centre for Advanced Mediterranean Agronomic Studies, in Zaragoza, Spain (http://www.iamz.ciheam.org/ingles/cursos09-10/) with broad spectrum of topics covering the most common bioinformatics tools relevant to plant breeders. In Africa, the West African Biotechnology Workshop Series of Nigeria facilitates such

trainings (http://www.wabws.org/workshops.htm) in collaboration with international research centers such as IITA and advanced labs in North America. Scientists in advanced labs such as Ontario Institute for Cancer Research collaborate with African institutions to offer bioinformatics education to many African young scientists (http://www.oicr.on.ca/Portalnews/Vol2_Issue4/africa.htm). Aside from the South African National Bioinformatics Institute (SANBI, http://www.sanbi.ac.za/), the author has no knowledge of African university offering undergraduate or graduate degree, a certificate, or other degrees with emphasis on bioinformatics. It is high time for African academic institutions to incorporate bioinformatics in their curriculum.

International institutions

The Generation Challenge Program (GCP, www.generationcp.org) provides various learning materials those are relevant to molecular breeding in its 'capacity building corner' (http://mbp.generationcp.org/). The Global Partnership Initiative for Plant Breeding (GIPB) at the Food and Agriculture Organization (FAO) strives to promote capacity building in main-stream plant breeding in Sub Saharan Africa (SSA) and in other continents. Young scientists trained in main-stream plant breeding programs should be equipped with complementary bioinformatics skill simultaneously for successful application of molecular plant breeding. Furthermore, in the electronic and computer age, applied bioinformatics can greatly appeal to young breeders with a knack for computer.

Conclusion

Advances in the genotyping technology has accelerated the growth of bioinformatics as evidenced by recent increase of publications in existing journals (e.g. Nucleic Acids Research vol 37, Web server Issue and vol 38, Database Issue), dozens of dedicated bioinformatics-specific journals (e.g. Bioinformatics, Briefings in Bioinformatics, BMC-Bioinformatics, PLoS Computational Biology), and books [11, 34, 51]. In appreciation of the role of bioinformatics in life science research, a number of research institutions in developing countries have embarked on the development of their computational biology capacity including China [50], and Mexico [32]. However, there is no substantial effort to develop computational capacity in Africa. The aim of this review is to urge individual scientists to consider capacity building in bioinformatics, on one hand, and to implore policy makers and national institutions to devise a strategy to benefit from these technological advances. It may not seem reasonable to promote bioinformatics capacity enhancement when only a handful of ill-equipped biotech laboratories in developing countries, particularly in SSA, are struggling to generate molecular data. However, it should be noted that several start-up companies are offering affordable genotyping services that orphan plant breeders, with limited resources, can tap into for innovative breeding approach to accelerate the process of variety development.

The myriads problems in agricultural production in developing countries offer extensive avenue for research. The new molecular technologies are revolutionizing crop improvement. Advances in genomics technologies and the associated computational resources are consistently evolving towards cost-effectiveness and accessibility thereby increasing the potential to be adopted by resource-poor countries. The rapidly growing and expanding advances in information and communication technology (ICT) such as World Wide Web greatly facilitate accessibility of these scientific advances. Bioinformatics is one of the remarkable achievements of this century that scientists in developing countries can mobilize to leapfrog agricultural productivity. However, these resources are largely unknown to scientists in low income nations. The field of genomics has become a vast, information-intensive discipline, sparking the development of numerous databases and tools in its wake. The frustratingly large number and variety of databases and tools calls for end-user support in identifying appropriate and reliable resources [28].

The rapid growth of sequencing and genotyping technology and the parallel growth of bioinformatics and online biological resources further broaden crop improvement strategies for well-studied and under-studied crops alike. Nowadays, genome sequencing of an organism does not entail large sum of money and long time. The rate limiting step is rather, mining the genome to unravel the genes and pathways underlying economically important traits. This entails a strong team of multidisciplinary genomics and informatics scientists. Africa has to go long way to build the critical mass of scientists dedicated to improvement of orphan crops. With the plethora of existing and emerging web resources, the sky is the limit for scientists. African governments, institutions, and policy makers should gear up to accelerate the development of bio-computational human resources if African agriculture is to benefit from the current 'omics' boom. In the interim, strong partnerships with advanced research institutions around the world should be fostered to leverage genomics and bioinformatics for accelerated improvement of Africa's neglected crops.

African bioinformatics scientists, however small in number, should join forces to mobilize funds to help create shared resources and expertise and build synergy. Regionally organized professional groups will have the opportunity to play a policy advocacy role to enhance government funding and also to liaise with ARIs.

References

1. Andersen,J.R., andT.Lubberstedt. (2003) "Functional markers in plants", Trends in Plant Science, 8, 554-560.

2. Appleby,N., D.Edwards, and J.Batley. 2009. New Technologies for Ultra-High Throughput Genotyping in Plants. p. 19-39. *In* D.J.Somers (ed.) Plant Genomics. Humana Press,

3. Armstead,I., L.Huang, A.Ravagnani, P.Robson, and H.Ougham. (2009) "Bioinformatics in the orphan crops", Brief Bioinform, 10, 645-653.

4. Ayeh, K. O. Expressed sequence tags (ESTs) and single

nucleotide polymorphisms (SNPs): Emerging molecular marker tools for improving agronomic traits in plant biotechnology. African Journal of Biotechnology 7[4], 331-341. (2008).
Ref Type: Generic

5. Bai,J., L.A.Pennill, J.Ning, S.W.Lee, J.Ramalingam, C.A.Webb, B.Zhao, Q.Sun, J.C.Nelson, J.E.Leach, and S.H.Hulbert. (2002) "Diversity in nucleotide binding site-leucine-rich repeat genes in cereals", Genome Res., 12, 1871-1884.

6. Benson,D.A., I.Karsch-Mizrachi, D.J.Lipman, J.Ostell, and E.W.Sayers. (2010) "GenBank", Nucleic Acids Res., 38, D46-D51.

7. Brooksbank,C., G.Cameron, and J.Thornton. (2010) "The European Bioinformatics Institute's data resources", Nucleic Acids Res., 38, D17-D25.

8. Budak,H., S.Su, and N.Ergen. (2006) "Revealing constitutively expressed resistance genes in Agrostis species using PCR-based motif-directed RNA fingerprinting", Genet. Res., 88, 165-175.

9. Chen,G., D.Pan, Y.Zhou, S.Lin, and X.Ke. (2007) "Diversity and evolutionary relationship of nucleotide binding site-encoding disease-resistance gene analogues in sweet potato (Ipomoea batatas Lam.)", J. Biosci., 32, 713-721.

10. Chen,Y.B., A.Chattopadhyay, P.Bergen, C.Gadd, and N.Tannery. (2007) "The Online Bioinformatics Resources Collection at the University of Pittsburgh Health Sciences Library System--a one-stop gateway to online bioinformatics databases and software tools", Nucleic Acids Res., 35, D780-D785.

11. Claverie,J.-M., and C.Notredame. 2006. Bioinformatics For Dummies. Wiley Publishing, Inc., New York.

12. Cochrane,G.R., andM.Y.Galperin. (2010) "The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources", Nucl. Acids Res., 38, D1-D4.

13. Collard,B.C., C.M.Vera Cruz, K.L.McNally, P.S.Virk, and D.J.Mackill. (2008) "Rice molecular breeding laboratories in the genomics era: current status and future considerations", Int. J. Plant Genomics, 2008, 524847.

14. Deane,C., G.Ejeta, R.Rabbinge, and J.Sayer. (2010) "Science for Development: Mobilizing Global Partnerships", Crop Sci, 50, v.

15. Dennis,E.S., J.Ellis, A.Green, D.Llewellyn, M.Morell, L.Tabe, and W.J.Peacock. (2008) "Genetic contributions to agricultural sustainability", Philos. Trans. R. Soc. Lond B Biol Sci, 363, 591-609.

16. Dwivedi,S., J.Crouch, D.Mackill, Y.Xu, M.Blair, M.Ragot, H.Upadhyaya, and R.Ortiz. (2007) "The molecularization of public sector crop breeding: prgress, problems, and prospects", Adv Agron, 95, 163-318.

17. Feltus,F.A., H.P.Singh, H.C.Lohithaswa, S.R.Schulze, T.D.Silva, and A.H.Paterson. (2006) "A comparative genomics strategy for targeted discovery of single-nucleotide polymorphisms and conserved-noncoding sequences in orphan crops", Plant Physiol, 140, 1183-1191.

18. Gedil,M.A., M.B.Slabaugh, S.Berry, R.Johnson, R.Michelmore, J.Miller, T.Gulya, and S.J.Knapp. (2001) "Candidate disease resistance genes in sunflower cloned using conserved nucleotide-binding site motifs: genetic mapping and linkage to the downy mildew resistance gene Pl1", Genome, 44, 205-212.

19. Hall,T.A. (1999) "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT", Nucl. Acids. Symp. Ser., 41, 95-98.

20. Hammond-Kosack,K.E., andJ.E.Parker. (2003) "Deciphering plant-pathogen communication: fresh perspectives for molecular resistance breeding", Curr. Opin. Biotechnol., 14, 177-193.

21. Harjes,C.E., T.R.Rocheford, L.Bai, T.P.Brutnell, C.B.Kandianis, S.G.Sowinski, A.E.Stapleton, R.Vallabhaneni, M.Williams, E.T.Wurtzel, J.Yan, and E.S.Buckler. (2008) "Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification", Science, 319, 330-333.

22. Kalendar R. FastPCR: a PCR primer and probe design and repeat sequence searching software with additional tools for the manipulation and analysis of DNA and protein. (www.biocenter.helsinki.fi/bi/programs/fastpcr.htm) . (2007).
Ref Type: Electronic Citation

23. Konieczny,A., andF.M.Ausubel. (1993) "A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers 449", The plant journal, 4, 403-410.

24. Larrinua,I.M., andS.B.Belmar. (2009) "Bioinformatics and Its Relevance to Weed Science", Weed Science, 56, 297-305.

25. Lee,I., B.Ambaru, P.Thakkar, E.M.Marcotte, and S.Y.Rhee. (2010) "Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana", Nat. Biotechnol, 28, 149-156.

26. Liolios,K., I.M.Chen, K.Mavromatis, N.Tavernarakis, P.Hugenholtz, V.M.Markowitz, and N.C.Kyrpides. (2010) "The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata", Nucleic Acids Res., 38, D346-D354.

27. Liu,Q., H.Wang, Z.Zhang, J.Wu, Y.Feng, and Z.Zhu. (2009) "Divergence in function and expression of the NOD26-like intrinsic proteins in plants", BMC. Genomics, 10, 313.

28. Messersmith,D.J., D.A.Benson, and R.C.Geer. (2006) "A Web-based assessment of bioinformatics end-user support services at US universities", J. Med. Libr. Assoc., 94, 299-87.

29. Moose,S.P., andR.H.Mumm. (2008) "Molecular plant breeding as the foundation for 21st century crop improvement", Plant Physiol, 147, 969-977.

30. Moroldo,M., S.Paillard, R.Marconi, L.Fabrice, A.Canaguier, C.Cruaud, B.De, V, C.Guichard, V.Brunaud, C.Le, I, S.Scalabrin, R.Testolin, G.G.Di, M.Morgante, and A.F.dam-Blondon. (2008) "A physical map of the heterozygous grapevine 'Cabernet Sauvignon' allows mapping candidate genes for dis-

ease resistance", BMC Plant Biol., 8, 66.

31. Nielsen,C.B., M.Cantor, I.Dubchak, D.Gordon, and T.Wang. (2010) "Visualizing genomes: techniques and challenges", Nat. Methods, 7, S5-S15.

32. Palacios,R., andJ.Collado-Vides. (2007) "Development of Genomic Sciences in Mexico: A Good Start and a Long Way to Go", PLoS Comput Biol, 3, e143.

33. Paran,I., andR.W.Michelmore. (1993) "Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce
81", Theoretical and Applied Genetics, 85, 985-993.

34. Pevsner,J. 2009. Bioinformatics and Functional Genomics. John Wiley & Sons.

35. Phillips,R.L. (2010) "Mobilizing Science to Break Yield Barriers", Crop Sci, 50, S-99.

36. Procter,J.B., J.Thompson, I.Letunic, C.Creevey, F.Jossinet, and G.J.Barton. (2010) "Visualization of multiple alignments, phylogenies and gene family evolution", Nat. Methods, 7, S16-S25.

37. Proost,S., B.M.Van, L.Sterck, K.Billiau, P.T.Van, P.Y.Van de, and K.Vandepoele. (2009) "PLAZA: a comparative genomics resource to study gene and genome evolution in plants", Plant Cell, 21, 3718-3731.

38. Rhee,S.Y., J.Dickerson, and D.Xu. (2006) "Bioinformatics and its applications in plant biology", Annu. Rev. Plant Biol., 57, 335-360.

39. Rotem,N., E.Shemesh, Y.Peretz, F.Akad, O.Edelbaum, H.D.Rabinowitch, I.Sela, and R.Kamenetsky. (2007) "Reproductive development and phenotypic differences in garlic are associated with expression and splicing of LEAFY homologue gaLFY", J. Exp. Bot., 58, 1133-1141.

40. Rudd,S., H.Schoof, and K.Mayer. (2005) "PlantMarkers–a database of predicted molecular markers from plants", Nucleic Acids Res, 33, D628-D632.

41. Saddic,L.A., B.Huvermann, S.Bezhani, Y.Su, C.M.Winter, C.S.Kwon, R.P.Collum, and D.Wagner. (2006) "The LEAFY target LMI1 is a meristem identity regulator and acts together with LEAFY to regulate expression of CAULIFLOWER", Development, 133, 1673-1682.

42. Sanseverino,W., G.Roma, S.M.De, L.Faino, S.Melito, E.Stupka, L.Frusciante, and M.R.Ercolano. (2010) "PRGdb: a bioinformatics platform for plant resistance gene analysis", Nucleic Acids Res, 38, D814-D821.

43. Sayers,E.W., T.Barrett, D.A.Benson, E.Bolton, S.H.Bryant, K.Canese, V.Chetvernin, D.M.Church, M.Dicuccio, S.Federhen, M.Feolo, L.Y.Geer, W.Helmberg, Y.Kapustin, D.Landsman, D.J.Lipman, Z.Lu, T.L.Madden, T.Madej, D.R.Maglott, A.Marchler-Bauer, V.Miller, I.Mizrachi, J.Ostell, A.Panchenko, K.D.Pruitt, G.D.Schuler, E.Sequeira, S.T.Sherry, M.Shumway, K.Sirotkin, D.Slotta, A.Souvorov, G.Starchenko, T.A.Tatusova, L.Wagner, Y.Wang, W.W.John, E.Yaschenko, and J.Ye. (2010) "Database resources of the National Center for Biotechnology Information", Nucleic Acids Res., 38, D5-16.

44. Smedley,D., S.Haider, B.Ballester, R.Holland, D.London, G.Thorisson, and A.Kasprzyk. (2009) "BioMart - biological queries made easy", BMC Genomics, 10, 22.

45. Stange,C. (2006) "Plant-virus interactions during the infective process", Cien. Inv. Agr., 33, 1-18.

46. Tamura,K., J.Dudley, M.Nei, and S.Kumar. (2007) "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0", Mol. Biol. Evol., 24, 1596-1599.

47. Tautz,D. (1989) "Hypervariability of simple sequences as a general source for polymorphic DNA markers 101", Nucl. Acids Res., 16, 6463-6471.

48. van der Linden,C.G., D.C.Wouters, V.Mihalka, E.Z.Kochieva, M.J.Smulders, and B.Vosman. (2004) "Efficient targeting of plant disease resistance loci using NBS profiling", Theor. Appl. Genet., 109, 384-393.

49. Volfovsky,N., B.J.Haas, and S.L.Salzberg. (2001) "A clustering method for repeat analysis in DNA sequences", Genome Biol, 2, RESEARCH0027.

50. Wei,L., andJ.Yu. (2008) "Bioinformatics in China: a personal perspective", PLoS. Comput. Biol., 4, e1000020.

51. Xu,Y. 2010. Molecular plant breeding. CAB International.

52. Xu,Z., D.Zhang, J.Hu, X.Zhou, X.Ye, K.L.Reichel, N.R.Stewart, R.D.Syrenne, X.Yang, P.Gao, W.Shi, C.Doeppke, R.W.Sykes, J.N.Burris, J.J.Bozell, M.Z.Cheng, D.G.Hayes, N.Labbe, M.Davis, C.N.Stewart, Jr., and J.S.Yuan. (2009) "Comparative genome analysis of lignin biosynthesis gene families across the plant kingdom", BMC. Bioinformatics., 10 Suppl 11, S3.

53. Ye,J., S.McGinnis, and T.L.Madden. (2006) "BLAST: improvements for better sequence analysis", Nucleic Acids Res, 34, W6-W9.

54. Yin,Y., J.Huang, and Y.Xu. (2009) "The cellulose synthase superfamily in fully sequenced plants and algae", BMC. Plant Biol., 9, 99.

55. Zhang,X.C., S.B.Cannon, and G.Stacey. (2009) "Evolutionary genomics of LysM genes in land plants", BMC. Evol. Biol., 9, 183.