Technical assistance to strengthen national agricultural research organizations' capacity to use digital sequence information

A submission from CGIAR

Technical assistance to strengthen national agricultural research organizations' capacities to use digital sequence information

A submission from CGIAR for consideration by the Governing Body of the International Treaty

Contents

List of boxes:

Box 1: Supporting partner organizations with decision-making in relation to DSI generation and use

Box 2: Sequencing services

Box 3: Examples of Bioinformatics courses involving CGIAR Centers

Box 4: Public repositories maintained by CGIAR Centers

Box 5: Ongoing whole genome sequencing activities and resulting DSI made available in public repositories

Box 6: Increasing capacities for applying DSI in genebanks

Box 7: Training national partners on the use of "omics" technologies in breeding

Box 8: Open access platforms for modern and collaborative breeding

1. Introduction

CGIAR submits this report in response to an open request from the Plant Treaty Secretary, received June 1, 2023, for submissions regarding, a) contracting parties' and stakeholders' capacity building needs for accessing and using digital sequence information (DSI)/genomic sequence data (GSD) and b) 'technical assistance' and 'actions taken' by stakeholders (including CGIAR) 'to reduce the existing gap on capacity regarding DSI/GSD'.

The primary objective of this paper is to respond to the second part of the request by sharing information about how CGIAR Centers and Initiatives have been assisting organizations outside CGIAR to access, generate, share, analyse, and use DSI for the conservation of plant genetic resources for food and agriculture (PRGFA) and for the use of PGRFA in pre-breeding and breeding.[1] This paper does not provide an exhaustive account of *all* of the Centers' relevant activities, but it does provide a general overview of the kinds of activities in which the Centers have been engaged. CGIAR very much appreciates the Governing Body's initiative, as expressed in Resolution 16/2022, to ask the Plant Treaty Secretariat to gather and synthesize information about both demand for, and supply of, capacity strengthening related to DSI linked to plant genetic resources for food and agriculture, with the overall objective of working to close the capacity gap between developed and developing countries. It is our hope that, based on the outcomes of this exercise, CGIAR will be able to further adapt and improve its own approach to capacity sharing in response to needs prioritized by the Governing Body.

CGIAR has already submitted, to the Plant Treaty's Governing Body, the Conference of the Parties to the Convention on Biological Diversity (CBD) and the FAO Commission on Genetic Resources for Food and Agriculture (CGRFA), detailed reports concerning how CGIAR scientists are using DSI in their own conservation and crop improvement programs.[2] We will repeat only as much of that information here as

---

[1] This paper does not address the unresolved issue of the definition of the term Digital Sequence Information (DSI). For the purposes of this submission, we are using the term to refer to sequences derived from DNA or RNA, including short and long reads and all derived molecular markers such as Single Nucleotide Polymorphisms (SNPs), Presence Absence variations, insertions and deletions (InDels), chromosomal translocations and rearrangements detected through sequencing and associated annotations setting out current knowledge of the functions of different parts of the genome. This corresponds to the narrowest potential definition of DSI as set out by the CBD Secretariat (2020a), and comprises a subset of potential broader definitions. We have adopted this definition for matters of convenience, because it corresponds to the scope of activities described in this paper. We do not want to be understood as taking a position on the scope of information that can or should be included in a definition of DSI that the international could possibly adopt in the future. From this point forward in this paper, we will refer to DSI and not DSI/GSD.

[2] Sackville Hamilton et al. 2020. '*Digital sequence information is changing the way genetic resources are used in agricultural research and development: implications for new benefit-sharing norms: a submission from CGIAR to the CBD/COP 15*', available at https://cgspace.cgiar.org/handle/10568/125749, and

CGIAR. 2023. *'Issues for further consideration concerning digital sequence information: a submission from CGIAR'*, available at https://www.cbd.int/notifications/2023-003 (Number 24, under 'Observers').

is necessary to set the context for our capacity strengthening engagement with organizations outside CGIAR.

2. CGIAR

CGIAR is a global research and innovations partnership;[3] its mission is to deliver science and innovation that advance the transformation of food, land, and water systems in a climate crisis. The primary focus of CGIAR's research and development are resource-poor farmers and consumers in developing countries. CGIAR seeks to scale up research outputs and innovations through capacity development and policy advice, delivered through regional and global initiatives in close collaboration with partners around the world, including national and regional research institutes, civil society organizations, academia, development organizations and the private sector.

CGIAR includes 15 international agricultural research Centers, as depicted in Figure 1, with more than 9,000 scientists working with more than 3,000 partner organizations in 90 countries around the world. Eleven of the Centers maintain, collectively more than 750,000 accessions of 'in trust' crops and forages, pursuant to agreements that those Centers signed with the Plant Treaty's governing body in 2006. CGIAR Centers coordinate breeding programs for more than 25 crops, forages and trees.

Figure 1: map showing CGIAR Centers' locations

As of January 1, 2022, CGIAR research is organized under three Action Areas: System Transformation, Resilient Agrifood Systems, and Genetic Innovation. Each Action Area is made up of Research Initiatives.

---

[3] https://www.cgiar.org/how-we-work/strategy/

Most research and development activities working directly with genetic resources for food and agriculture and associated information (including DSI) take place under the Genetic Innovation Action Area. This Action Area includes the CGIAR genebanks, genomics research and the plant breeding programs.

In the following sections of this submission, we describe how the generation and use of DSI can be useful for the conservation and sustainable use of PGRFA, and we include examples of how CGIAR Centers provide assistance to other organizations to use DSI, either as partners in CGIAR Research Initiatives, participants in training courses, or as downstream users of the tools, technologies, methods, knowledge and data that CGIAR Centers have generated.

The following sections of this paper are organized around fundamental activities or steps that involve DSI, from deciding what materials to sequence and how to sequence them, to using genomic tools in genebank operations, pre-breeding and plant breeding activities. Of course, in reality, these activities are usually not discrete. Most of the efforts of CGIAR to assist other organizations using DSI do not focus on a single step as presented below, but address knowledge and technology gaps along comprehensive research and development chains.

3. Generating DSI

DNA sequence information is extremely helpful for understanding genetic variability among cultivars and populations across and within species, and for identifying and deploying key genes controlling important agronomic traits.

Sequencing and annotating a genome accurately from scratch is time consuming and expensive. Although time and costs have been reduced dramatically in recent years, it remains expensive to undertake on a large scale. A single genome that has been sequenced with high accuracy and reliability may be used as a "reference genome" enabling cheaper low-cost high-throughput genotyping or resequencing (using next generation sequencing -NGS- approaches) of many related genomes to identify genetic diversity. This low-cost genotyping and resequencing data then becomes a recurrent and dynamic resource; new data can be added incrementally over time and re-used by multiple researchers simultaneously around the world for a wide variety of purposes including trait discovery, mapping and incorporation into breeding programmes.

In order to build the knowledge that they need in an efficient manner, research organizations must first address fundamental questions such as what to sequence and what for, and whether or not the desired genetic information has been already generated and is available. Depending on the answers to these questions, sequencing efforts may need to cover whole collections of PGRFA or subsets; they may focus on whole genome resequencing or targeted sequencing; they may require more or less sequencing coverage and hence different technologies and expertise. Many research organizations located in the developing countries lack the necessary infrastructure, expertise and knowledge to address these questions in an informed manner.

CGIAR Centers and Initiatives deliver courses that include training sessions aimed at improving decision-making in relation to DSI generation and use in conservation and breeding activities. We present some of these courses in subsequent sections of this paper. In addition to these courses, medium to long term partnerships and collaborations between CGIAR Centers and Initiatives, and national agricultural research organizations often help strengthen the capacity of those partners to make informed decisions about what sequencing practices and genomic tools to use as part of their countries' national conservation and plant breeding programs. Examples of these collaborations are presented below, in Box 1.

Box 1: Supporting partner organizations with decision-making in relation to DSI generation and use

Strategic advanced genotyping in Mozambique

Collaborative activities between the CGIAR (through the 'Excellence in Breeding' Platform), the International Center for Maize and Wheat Improvement (CIMMYT), the Instituto de Investigação Agrária de Moçambique (IIAM) and Zambeze University have sought to implement effective genotyping practices in the context of diversity characterization and crop improvement in the Zambezia, a central province of Mozambique where periods of drought have become longer and more frequent.

The collaboration included heterotic group characterization from the Mozambique Genebank for improvement of rice, cowpea, and maize populations in view of current production challenges.

The collaboration contributed to better prioritization of genotyping efforts, faster delivery of results, and avoidance of work duplication among institutions.

Genotyping activities in West Africa for yams (Nigeria, Ghana, Benin and Cote d'Ivoire)

As part of the 'Excellence in Breeding' Platform of CGIAR, CGIAR Centers carried out the AfricaYam project, which was funded by the Bill and Melinda Gates Foundation, and implemented together with national agricultural research organizations in West Africa such as National Roots Crop Research Institute (NRCRI), Nigeria; Ebonyi State University, Nigeria; CSIR-Crops Research Institute (CRI), Ghana; CSIR-Savannah Agricultural Research Institute (SARI), Ghana; University of Abomey-Calavi, Benin Republic; and Centre National de Recherche Agronomique (CNRA), Cote d'Ivoire. The project developed strategies to select samples (both landraces and breeding lines) for genotyping using advanced NGS technologies with the objective to advance pre-breeding and breeding of yam in these countries.

Through this collaboration, several yam genotypes have been selected and genotyped and trait-marker associations have been established using GWAS (genome-wide association studies) and QTL (quantitative trait analyses) for several targeted traits. Several training programs have been organized for sample collection, sample preparation and data analyses for NARS partners. This collaborative effort has provided platform to NARS partners to strategize their sampling efforts including selection of genotypes for genotyping and identification of duplicates across different institutes.

Not all organizations involved in PGRFA conservation and use can afford to have their own sequencing or genotyping platforms. In addition to acquiring the sequencing equipment and paying for its maintenance, organizations need to cover the costs of instruments, reagents and work that are necessary to process each sample (including, for example, instruments for nucleic acid quantitation, quality analysis and isolation; instruments and expertise for library preparation; and sequencing reagents). Moreover, to keep up with the rapid evolution of sequencing and genotyping technologies and avoid becoming dependent

on old, slow, expensive equipment, they need to invest frequently in upgrading the equipment. Contracting external services is often more effective than 'in-house' options, but the public genebanks and breeding programs do not always have access to the optimal services, lacking knowledge about providers, and not having sufficient demand for sequencing or genotyping to be able to negotiate advantageous terms. To address these gaps, CGIAR Centers and Initiatives offer sequencing and genotyping services, both in-house and through agreements with external providers, at low financial and administrative costs. Box 2 provides some details.

Box 2: Sequencing services

Center of Excellence in Genomics and Systems Biology

The Center of Excellence in Genomics and Systems Biology of the International Crops Research Institute for the Semi Arid Tropics (CEGSB of ICRISAT; https://cegsb.icrisat.org/) offers genomics and informatics services to CGIAR programs and national research organizations on a cost-to-cost basis. The CEGSB houses genomics platforms including HiSeq 2500, MiSeq, ABI3730, Microarray Spotter & Scanner, and more. It also has a high-performance computational analysis platform with 600 cores, 6 TB RAM, and 830 TB storage capacity. The services offered include *De novo* and re-sequencing of genomes, RNASeq/Transcriptome sequencing and small RNA sequencing, ChiP-Seq, Methylation/Bisulfite sequencing, SNP (KASPar assay) and SSR genotyping, DArT genotyping, qRT-PCR analysis, Sequence alignment and variant calling, basic genome and transcriptome analysis and Genotyping-by-sequencing analysis. Until now, CEGSB has generated 56 TB data through sequencing DNA/RNA and 22 million datapoints from more than 16000 samples using different genotyping platforms. The objective is to scale-up genomics applications in breeding and research at ICRISAT and partner organizations.

ILRI's Genomics and Bioinformatics platforms

The Genomics Platform of the International Livestock Research Institute (previously the BecA-ILRI Hub), which is based in Kenya, offers standardized protocols and last-generation technologies to Africa-based research organizations and researchers. A wide range of services on offer include nucleic acid extraction, quantification and amplification, library preparation, RT-qPCR analysis and DNA/RNA sequencing using Sanger, Illumina (MiSeq, HiSeq & NextSeq) and Oxford Nanopore (MinION & Flongle) technologies. In terms of molecular biology, the platform also offers DNA cloning, bacterial transformation, and transfection, and random mutagenesis services and single-cell sequencing and gene expression analysis.

ILRI's Genomics platform is complemented with a Bioinformatics platform offering services which include data storage and data analysis using an on-site High-Performance Cluster equipped with four compute nodes totalling 128 CPU cores and 5,120 GPU cores for massively parallel computing. The team has developed a strong capacity in whole-genome sequencing and assembly at chromosome level, using various techniques to aid scaffolding (long reads, chromosome conformation capture) and to enhance genome annotation (transcriptomics, repeat analysis and masking, annotation transfer, etc).

SAGA

The Genetic Analysis Service for Agriculture (SAGA) located at CIMMYT is a research and service laboratory that uses the DArTseq method of genotyping, offering high, mid and low density DArTseq services for agriculture. The sequencer at SAGA can read up to 100,000 samples annually and generates large quantities of data at a comparatively low price. SAGA provides genotyping services to national and international collaborators, including CIMMYT, the International Center for Agricultural Research in the Dry Areas

(ICARDA), the International Center for Tropical Agriculture (CIAT), the Mexican Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (INIFAP), national genebanks in Latin America, private companies and universities. In addition to genotyping service, staff at SAGA offer support to clients to explore, understand and interpret data generated, facilitated through workshops, small group and 1:1 training focusing on the appropriate generation, use and interpretation of genotypic data for research and germplasm development.

CGIAR Initiative 'Breeding Resources'

The CGIAR Initiative 'Breeding Resources' negotiates contracts with validated DNA extraction and genotyping service vendors based on anticipated bulk demand across eligible CGIAR and national agricultural research organizations. The Initiative receives and aggregates demands for genotyping to negotiate competitive genotyping prices with vendors, using one centralized administrative and contracting process. This simplifies paperwork and reduces the price of genotyping considerably. The CGIAR Initiative 'Breeding Resources' has published manuals and video tutorials explaining how to make use of this service effectively.

The low-density genotyping service is based on Kompetitive allele specific PCR (KASP). Low-KASP genotyping is highly specific to a target: it can be used for QC analysis, dropdown selection, marker assisted selection and forward breeding. The markers are known, validated and in use. They are in the public domain and accessible at https://excellenceinbreeding.org/module3/kasp. The list of markers is continuously updated and improved. To illustrate the volume of this service, Figure 1 shows that 13,000 samples of 10 crops in 8 African countries were genotyped from September to December 2019.
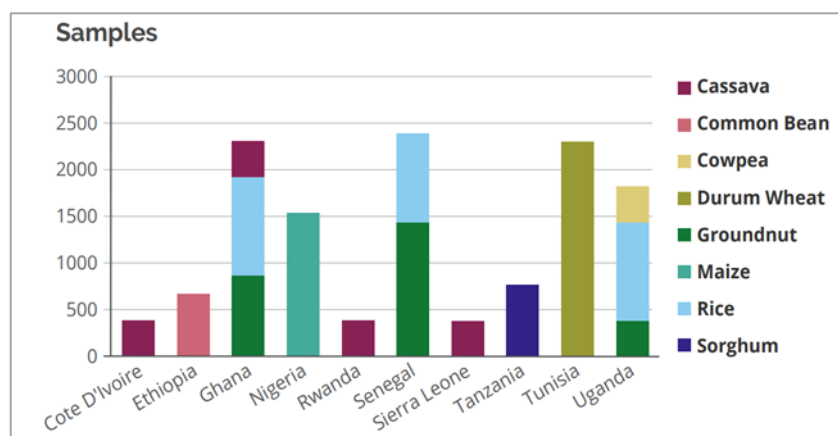

Figure 2. Number of samples, crops they belong to and provenance of the request for genotyping through the Low-density KASP genotyping service in September-December 2019.

The mid-density genotyping service is based on DArTag method providing a fixed, pre-defined set of markers for select species. The service is primarily intended for genomic selection applications, but mid density panels can also be used for background recovery in marker assisted selection.

By June 2023 mid-density DArTAg panels have been implemented for common bean, cowpea, groundnut, maize, pigeonpea, potato, rice, sorghum and wheat.

Recently, high-density genotyping services have started to be provided through a service agreement with Texas A&M University. This is being used for example in generating 3x skim-sequencing data on rice samples for approximately $35/sample.

4. Maintaining and analyzing DSI

Bioinformatics tools to manage and analyze DSI data are essential for plant genome research. Many plant genome databases have been established and continue to expand. Analytical methods based on bioinformatics are also well developed in many aspects of plant genomic research including comparative genomic analysis, phylogenomics, transcriptomics, genome-wide association studies, etc. The need to constantly upgrade computational infrastructures including high-capacity data storage and high performing analysis software, and the need to count on a skilled bioinformatics workforce are challenges for the effective use of DSI. Many research organizations, often those located in developing countries, lack the necessary expertise and infrastructure to process large genomic datasets.

Bioinformatics training in most parts of the world occurs primarily in university programs which are usually tailored for those who want to specialize in bioinformatics. To gain proficiency in programming languages, data management and processing, which are essential for effectively handling various genomics-related tasks, bioinformatics training typically spans several months or even years. This comprehensive approach may not be optimal for individuals seeking to concentrate on a specific area, such as analyzing genotyping data in plants. In such cases, it becomes crucial to have tailored training materials and tools that do not necessitate extensive bioinformatics expertise. Some CGIAR Centers provide short trainings that equip plant scientists from developing countries with the necessary bioinformatics skills to analyze DSI.

Box 3: Examples of Bioinformatics courses involving CGIAR Centers

Bioinformatics and modern breeding methods courses at ICRISAT by the Center of Excellence in Genomics & Systems Biology (CEGSB)

CEGSB of ICRISAT (https://cegsb.icrisat.org/) has continuously been engaged in building expertise in bioinformatics and computational biology to effectively analyse and interpret large scale sequencing and genotyping datasets. So far, CEGSB has conducted 16 training programs in the area of crop genomics and sequencing data analysis for national agricultural research organizations, and trained in total 492 participants from 36 countries in the area of crop breeding modernization and improvement. These training programs have also fostered collaborations and partnerships with different universities and institutes to promote knowledge exchange, and resource and expertise sharing.

Bioinformatics courses at ILRI

Since 2013, ILRI has regularly organized short training courses on bioinformatics for African researchers in collaboration with other expert organizations, where participants are provided some form of scholarship.

Capacity development in the Genomics Platform of the International Livestock Research Institute (previously the BecA-ILRI Hub) includes annual hands-on training workshops in key skill areas; research placements enabling early career scientists to spend up to six months at the Hub; and institutional support to key laboratories and individuals in the region. As part of network building, the hub has established 13 communities of practice, linking African scientists of different disciplines with similar interests, some of them related to genomics. Recently, two unique and in-depth bioinformatics training programs have been implemented at ILRI Genomics Platform which aimed at helping early-career African researchers harness the power of long-read DNA sequencing to unlock the genetic potential of crops and livestock in Africa. The second iteration of these training programs was entitled '3rd Generation Genomics and Bioinformatics in

---

Africa' and was jointly organized by the International Livestock Research Institute (ILRI, Kenya), the John Innes Centre (JIC, UK) and the Earlham Institute (EI, UK) with support from UK Research and Innovation (UKRI) Biotechnological and Biological Science Research Council (BBSRC) and Oxford Nanopore Technologies. It was delivered in three phases: 'Develop', 'Demonstrate' and 'Deploy'. The 'develop' phase was a two-month long hybrid (residential and online) training program that introduces all the major steps in long-read sequencing from DNA extraction, library preparation, sequencing, and bioinformatics analyses. The 'demonstrate' phase allowed for project-based learning where trainees got hands-on experience from three short projects. The 'deploy' phase will extend the impact of the training to the home institutions of the trainees across the continent. In this phase, the trainees will implement a long-read sequencing project they design themselves and will also train others in their home institutions.

As a tangible output of these two in-depth training programmes in bioinformatics, the trainees have led an effort whose outcome was the publication of the first chromosome-level genome for *Lablab purpureus* (hyacinth bean), and they are working on assembling and annotating complete genomes for two other neglected African crops, *Sphenostylis stenocarpa* (African yam bean) and *Dovyalis caffra* (kei apple). The 29 early-career trainees of these two cohorts are now occupying prominent positions in institutes with regional or continental reach in Africa and the Middle East, or completing their PhD studies.

Bioinformatics Training Courses at IITA

The International Institute of Tropical Agriculture (IITA) conducts two bioinformatics workshops annually, one as 'Basic Bioinformatics Workshop' and the other as 'Advanced Bioinformatics Workshop'. These workshops target research scientists, lecturers and teachers, graduate students, technicians, and all those who require knowledge of bioinformatics within Africa and beyond. These workshops are being provided from 2019 and through applications from interested participants after the announcements is made every year twice. The registration fees for these courses are $100 for each course and covers tuition fees, course materials and training certificate. The aim of these training workshops is to equip participants with the basic knowledge and skills of bioinformatics. These training comprises of 3-day workshop that includes theoretical and hands-on training in basic bioinformatics techniques. Participants are also introduced to online repositories and trained on how to work with them. The 'Basic Bioinformatics training course' includes the following sessions: Introduction to Bioinformatics, Hands-on training on DNA repositories (National Center for Biotechnology Information of the National Institutes of Health -NCBI, the European Nucleotide Archive -ENA, Hands-on training on protein repository (UniProt), BLAST sequence alignment, Hands-on BLAST search, PCR primer design and evaluation, Phylogenetics tree construction, Introduction to next generation sequencing, and Introduction to RNA-seq data analysis. At the end of the training, the participants are evaluated based on their experience in the use of online tools like BLAST for sequence search, knowledge and experience in primer design and evaluation, and knowledge and experience in the use of online repositories. Similarly, the 'Advanced Bioinformatics training course' includes sessions such as Analysis of next-generation sequencing, RNA-sequence analysis, Metagenomics analysis, De-novo sequencing, and Gene annotations. At the end of this training, participants are evaluated based on their knowledge and experience on analyzing and interpreting genetic data and information.

CABANA project

The project 'Capacity building for bioinformatics in Latin America' (CABANA) was designed to strengthen local capacity for data-driven biomolecular research by creating a sustainable capacity-building program focusing on three challenge areas – communicable disease, sustainable food production and protection of biodiversity. CABANA was orchestrated by an international consortium of ten organizations - nine in Latin America (including the International Potato Center - CIP) and one in the UK. The project promoted

networking, collaboration and open sharing of information. CABANA activities included: 1) Secondment of Latin American scientists to the European Bioinformatics Institute; 2) Train the trainer activities to disseminate knowledge; 3) Workshops/short courses – one per participating country per year – to build a community of practice; and 4) E-learning resources to catalyze the roll-out of bioinformatics education and support ongoing collaboration. By the end of the project, over 100 bioinformatics trainers were trained and more than 800 scientists attended one of the project's bioinformatics workshops. Bioinformatics positions had increased by over 80% amongst participating institutions, whereas the number of projects generated by partners rose by an average of 5 per year compared to baseline; international collaborations of partner institutions increased by almost 30%. The project generated 20 publications and 95% of partners reported submitting all their data to publicly available databases. After the project ended, the community of practice continued and managed to generate a follow up project 'CABANAnet'.

Bioinformatics courses from South Green platform

The South Green platform delivers a wide range of courses in genomic data analyses on an annual basis. These courses cover various topics such as transcriptomics, genome sequencing, scientific programming, and reproducible research. The platform is strongly supported by French institutions IRD, CIRAD, INRAE, and the Alliance of Bioversity International & CIAT.

Most of these courses are conducted in Agropolis, Montpellier, France, where a rich pool of expertise is available. However, South Green has expanded its reach beyond France and has successfully organized courses in other countries as well, including Senegal, Burkina Faso, and Vietnam. This global outreach has allowed the platform to engage with diverse scientific communities and foster collaboration on an international scale. Over the past 10 years, South Green has organized more than 50 courses. These courses are designed to be hands-on and offer practical training experiences to participants.

In addition to the courses, the platform also serves as a hub for hosting students and visiting scientists who wish to undertake longer training periods. This facilitates the exchange of knowledge and expertise between researchers from different backgrounds. These individuals have the opportunity to engage in joint research activities, enabling them to collaborate on cutting-edge projects and contribute to the advancement of genomics research.

5. Sharing DSI

Data volumes, data backup, and site design are massive challenges that make the maintenance of DSI repositories extremely expensive. Developing and maintaining DSI repositories is well beyond the means of most researchers and labs in both developed and developing countries. By incurring the costs of maintaining DSI repositories and making these resources freely accessible, CGIAR Centers help lower and overcome other organizations' barriers to entry into research and development that depends on access to and use of such DSI.

CGIAR Centers develop and host repositories for DSI related to genebank and/or breeding material. The following text box describes some examples of DSI repositories that have been developed by CGIAR Centers, or that are maintained by CGIAR Centers.

Box 4: Public repositories maintained by CGIAR Centers

SNP-seek database

The International Rice Research Institute (IRRI) has developed the Rice SNP-seek database that allows quick retrieving of SNP alleles for all varieties in a given genome region, finding different alleles from predefined varieties and querying basic passport and morphological phenotypic information about sequenced rice lines (Mansueto *et al.,* 2017). SNP-seek was initially populated with data from the 3000 rice genomes sequencing project and the earlier 2000 accessions genotyped on the high-density 700,000-SNP HDRA. It is currently being extended with data from resequencing of thousands of elite breeding lines, the 10,000 rice genomes sequencing project, and skim-sequencing of 5000 georeferenced accessions. This data is directly traceable to source accessions in the IRRI genebank, facilitating use of this data and these accessions by researchers in many countries.

CicerSeq

CicerSeq is a public repository developed by ICRISAT that includes a global map of genome variation based on sequencing of 3,366 Cicer genomes representing 3,171 cultivated species accessions and 195 accessions of seven wild species (Varshney et al, 2021). Information about genotype, passport data and different variants including SNPs and various structural variations (SVs) can be accessed through this repository. This database provides information to facilitate genomics-informed decisions in chickpea improvement.

Dataverse collections

The Dataverse Project is an open-source web application to share, preserve, cite, explore, and analyze research data. It was developed by the Harvard University's Institute for Quantitative Social Science (IQSS), along with many collaborators and contributors worldwide. Most Centers maintain Dataverse Collections that facilitate access to DSI data under permissive Creative Commons licenses. For example, the CIMMYT Dataverse Research Data and Software Repository Network and CIMMYT wheat and maize Germinate Instances, managed and maintained by CIMMYT, together house DSI and associated datasets on over 100,000 accessions from CGIAR and partner genebanks together with data from germplasm from CGIAR and partner breeding programs.

Genotype Investigator for Genome-Wide Analyses (GIGWA)

GIGWA software was developed by CIRAD. It helps explore large amounts of genotyping data by filtering based on variant features, including functional annotations, and on genotypes. To date, GIGWA has been adopted to create and maintain repositories of DSI for *musa* collections maintained by the Alliance of Bioversity International and CIAT, and for all collections maintained by ICARDA. There has been an agreement that all the other CGIAR Centers hosting collections under Article 15 of the International Treaty on Plant Genetic Resources for Food and Agriculture will also adopt GIGWA for sharing their own DSI repositories.

CGIAR scientists use also online public platforms that are maintained by other organizations to publish DSI and make available genomic tools such as markers. Examples of these platforms are the Germinate platform. Germinate is an open-sourced platform for plant genetic resource information. It hosts data for a wide range of crops in collaboration with international partners like CIMMYT and the Crop Trust. Germinate repository includes genotypic data of maize and wheat from the Seeds of Discovery project, which was carried out by MasAgro (Mexico) and CIMMYT (Raubach *et al.,* 2020), as well as the information

arising from the CIMMYT-led Allele Mining Initiative. This initiative focuses on targeted trait discovery and deployment activities for climate change adaptation from the immense germplasm bank resources of the CGIAR. Using the power of GWAS and genomic prediction approaches the initiative leverages the intrinsic value of collection site historic climate data to identify genomic features of landrace collections associated with climatic features of breeding interest. In addition, when publishing scientific results based on DSI, CGIAR scientists submit raw DSI datasets to the International Nucleotide Sequence Database Collaboration database, as expected by academic publishers and in line with current good practices in the scientific community.

Without these repositories, researchers would be limited to just genomic data they themselves have generated. Providing access to these datasets ensures their use and does not remain limited to the organisations that generated them. It has been demonstrated that when data are made available through open access infrastructures, scientists in both developed and developing countries access and use, in roughly the same proportions, DSI generated from genetic resources outside their borders and DSI generated from genetic resources located within their borders (Scholz et al, 2021). While there are significant differences among countries in the number of scientists and the volume of use and reuse of DSI, this finding nonetheless underscores the utility of increasing investments for scientists in low and middle-income countries (LMICs) to access, generate, make available, and use DSI.

Box 5 describes whole genome sequencing activities and genotyping activities in which CGIAR Centers have participated and which are leading to new DSI being made public and easily browsable on online repositories, including the ones mentioned above. It is worth noting that many of these activities focus on less widely used crop and tree species which are crucial for food and nutrition security, and production diversification and sustainability.

Box 5: Ongoing whole genome sequencing activities and resulting DSI made available in public repositories

African underutilized crops and trees

The African Orphan Crops Consortium (AOCC) is hosted by the World Agroforestry (CIFOR-ICRAF) at the headquarters in Nairobi, Kenya. The Consortium is sequencing whole genomes of 101 African underutilized trees and crops. The sequencing services are available only for the trees and crops mandated by the then New Partnership for African Development (NEPAD), now the African Union's Development Agency (NEPAD-AUDA) in a project-based partnership model. Until now, the consortium has published 14 whole genome sequences, eight of which are available at a dedicated branch of the Online Resource for Community Annotation of Eukaryotes (https://bioinformatics.psb.ugent.be/orcae/aocc/; yssel et al, 2019) and the remaining will be added soon. Currently, AOCC partners are sequencing 30 more genomes. The AOCC operates in a public-private partnership model which currently comprises 38 partners which offer their expertise spread across knowledge generation, data analytics, sequencing and complementary technologies, instrumentation, capacity building, and advocacy & fund raising.

Minor grains and legumes

Together with collaborators and partners from different countries, ICRISAT's Center of Excellence in Genomics and Systems Biology (CEGSB) has generated sequencing data for developing reference genome assemblies for each of ICRISAT's mandate crops, and whole genome resequencing has taken place for 6632

---

germplasm accessions (3366 chickpeas; 300 pigeonpea, 1800 groundnuts, 994 pearl millet and 172 finger millet). All the sequencing data are available on NCBI.

Tropical forage species

Tropical forages are among the most neglected or forgotten crops and it remains difficult to get digital sequence information on most of the species. ILRI has been working on reference genome sequencing of a few of the species: *Lablab purpureus* (Njaci et al. 2023), *Melilotus albus* (Wu et al., 2022) and Napier grass (Yan et al. 2021) in collaboration with Chinese researchers at Lanzhou University. The genomes of these species are publicly available on the NCBI and the Chinese National Genomics Data Center (NGDC).

Root and Tuber Crops

IITA in collaboration with Japan International Research Centre for Agricultural Science (JIRCAS), the Iwate Biotechnology Research Centre (IBRC) Japan, the Earlham Institute, UK and the Sainsbury Laboratory, UK generated the first ever whole genome reference sequencing of white guinea yam (*Dioscorea rotundata Poir*) (Tamiru et al., 2017). The high-quality draft genome sequence is available in the public databases DNA Data Bank of Japan (DDBJ) and NCBI. Similarly, a highly contiguous chromosome-scale genome assembly of water yam or greater yam (*Dioscorea alata L.*) was generated by IITA in collaboration with the National Root Crops Research Institute of Nigeria and the University of Berkeley through funding from the National Science Foundation (Bredeson et al, 2022). The genome sequence, annotation, and SNP data are browsable at Phytozome or YamBase. The *D. alata* nuclear genome, transcriptome, plastid, and mitochondrion assemblies were deposited in the NCBI GenBank database. Recently, an IITA researcher received 'Greater Good Initiative' award from Illumina to carry out whole genome resequencing of 1000 yam genotypes. This data will allow genome-wide mining for targeted genes for economically important traits.

Other crops

A public-private partnership between the CGIAR Initiatives 'Crops to End Hunger' and 'Breeding Resources' and Corteva Agriscience has led to high quality reference genomes of eight banana varieties (three wild banana diploid varieties, one edible diploid banana variety, three improved diploid banana parents, and one plantain variety), one early maturing, photo-insensitive, disease resistant variety of cowpea, and two potentially heat, drought and pathogen resistant common bean varieties. Currently other crops are in the pipeline. The genomes are publicly available through CIMMYT Dataverse and are meant to address a key bottleneck of lack of good reference genomes that are needed to develop molecular tools for breeding.

Viromes of food security crops (sweetpotato in Africa, Potato in Peru & Burundi)

CIP in collaboration with a US university and NARS in 12 African countries, generated the African sweetpotato virome database, identifying and containing sequence data for all viruses in the crop. A similar effort was done for potato in Peru and Burundi. These data are openly available and provide an overview of identity and prevalence of viruses in different regions that serves as a resource for breeders, seed producers and researchers alike. Several new viruses were identified for both crops that may represent potential threats in the future.

6. Applying DSI in PGRFA genebanks

DSI can be used for the rational, efficient and effective conservation and use of biological diversity in *ex situ* collections. Genome sequences can be processed to identify and catalogue functionally relevant variants in accessions. Many sequence variants have no known effect on the regulation of genes and

contribute little to functional diversity; focusing specifically on functionally distinct variants transforms the task of optimising conservation based on DSI into an achievable as well as an effective task.

With a catalogue of functionally distinct variants, the composition of a collection can be altered to maximise the functional diversity that can be conserved within a fixed budget, by archiving accessions that bring little unique diversity to the collection and creating space to bring in different accessions with greater diversity. Addressing the challenge of limited budgets for *ex situ* conservation has been a priority for many years, but without DSI it is essentially unsolvable.

DSI data that are generated with genotyping platforms that minimize ascertainment bias (i.e., those that do not rely excessively on insufficient reference genomes) shed light on the genetic relationships among accessions. These data also facilitate identification of gaps, as well as genetically redundant accessions, both within and across multiple germplasm collections conserved by different genebanks. Such cross-genebank analyses can enhance the efficiency of the global system of conservation of PGRFA by 1) contributing to reduce redundancies within and among genebank collections; and 2) supporting decisions to collect new germplasm based on the total diversity conserved across multiple genebanks.

DSI can also be used to study the diversity within accessions. This helps genebanks adopt appropriate conservation and regeneration strategies to preserve the genetic integrity of accessions.

Under the CGIAR Initiative on Genebanks, CGIAR Centers are providing support to selected national agricultural research organizations in developing countries for the integration of DSI in the work of national genebanks.

Box 6: Increasing capacities for applying DSI in genebanks

Strengthening DSI Capacity in Latin America: Collaborative Efforts of the CoP for PGRFA Conservation and Utilization

The Alliance Bioversity International & CIAT and CIMMYT have collaborated since 2021 to establish a Latin America and the Caribbean Community of Practice (CoP) focused on enhancing capacity in DSI for the conservation and utilization of PGRFA. Their primary focus is on crops with centers of origin in the region, including common bean, maize, potato, and cassava. The first step in forming the CoP involved conducting a comprehensive survey to understand the needs and interests of national genebanks in Latin America. The survey received responses from 48 participants representing 14 countries and 33 institutions, providing valuable insights. The survey revealed that genebank collections varied in size, ranging from 1 to 50,000 accessions, with the majority falling between 1,000 and 10,000 accessions. Most genebanks had genotyped only a small proportion (0-20%) of their collections, primarily for genetic diversity studies, trait discovery, association analysis, and germplasm characterization. Challenges related to sharing datasets between collaborators were identified, including issues of trust, policies, intellectual property, data size, permits, and agreements. To address these capacity gaps, during 2022 the CoP conducted two virtual and one in-person workshops where participants shared experiences and knowledge on sample traceability and genotyping methods for generating DSI. During the in-person workshop organized in Mexico, and sponsored by the CGIAR Genebank Initiative, the CoP reviewed different tools and methods for DSI curation, analysis and data interpretation and provided an opportunity for participants to use the DArTseq genotyping platform at the

Genetic Analysis Service for Agriculture (SAGA) at CIMMYT to genotype accessions of various crops. In 2023, the CoP reconvened in April to socialize their experiences and results of DSI from their collections. With genotyping data now available to the genebanks, the CoP aims to compare the conserved diversity among national genebanks in the Latin American region. Based on initial results, additional funding may be sought to conduct more comprehensive cross-genebank diversity studies. These collective efforts have not only fostered a community of practice among Latin American genebanks but have also paved the way for addressing other genebank-related topics beyond DSI in the future. The CoP is actively planning collaborative initiatives and seeking opportunities for further advancement in the field of plant genetic resources conservation and utilization.

Training and supervision of students for DSI generation and use in genebanks

Training and mentoring of students are very common in CGIAR Centers' genebanks. More and more often, the students' work involves generating and applying DSI in accession characterization. Some recent examples: AfricaRice genebank supervised two PhD studies by Benin PhD candidates "Genetic Resources Characterization of African Rice", which involved the development and validation of diagnostic SNP markers for quality control genotyping in a collection of four rice (*Oryza*) species and the comparisons of sampling methods for assessing intra and inter accession genetic diversity in three rice species using genotyping by sequencing; and "Screening of traditional accessions from *Indica* group for their resistance to rice yellow mottle virus (RYMV)", which applied DSI to identify RYMV resistant rice landraces.

7. Modernizing breeding

Advancements in genomics and genetic research have revolutionized the field of plant breeding, from the development of molecular markers and their use in marker assisted selection to the application of more efficient and targeted approaches to trait selection. One such approach is Genome Wide Association Studies (GWAS), which utilizes whole-genome genotyping data to identify genetic markers associated with specific traits. GWAS allows for the identification of chromosomal regions that may be involved in controlling target traits, making it a valuable tool in phenotyping trials. It is scalable and transferable, and, where genome sequences have been resolved into genes and regulatory regions, it can be used to identify candidate genes (i.e. genes that may control, but that have not been proven to control, the target trait). Building on GWAS, various techniques can be used to further enhance understanding of the genotype-phenotype relationship. For example, fine mapping can be used to refine genotype-phenotype correlations to smaller genome regions, thus reducing the number of candidate genes; and gene editing can be used to knock out (or disable) a candidate gene and thus prove whether or not it does indeed control the target trait.

Genomic selection is a new approach for improving quantitative traits in large plant breeding populations. It combines whole-genome molecular marker data with phenotypic and pedigree data (when available) to estimate breeding value through genomic prediction. Genomic selection by-passes the need for trait-specific knowledge and uses advanced algorithms and machine learning techniques to guide breeding work based on the genome. It helps to increase genetic diversity and speed up genetic gains in crop improvement programmes. Similarly, haplotype-based breeding (HBB), which utilizes information about

specific combinations of alleles, known as haplotypes, holds great potential to improve the efficiency and precision of plant breeding programs.

GWAS, fine mapping, gene editing and genomic selection/prediction have become important tools in CGIAR pre-breeding and breeding programmes. Recently, CGIAR Centers that apply these tools have started collaboration with breeding programs in Africa, Asia and Latin America for their application in collaborative breeding with national programs. Some Centers have made great advances in this direction. For example, ICRISAT, together with partners from national agricultural research organizations from Asia and Africa, established and implemented trait mapping pipeline for crop breeding resulting in the development of 21 genomics-bred varieties and hybrids (12 in chickpea, 6 in groundnut and 3 in pearl millet) for release and commercial cultivation.

Box 7 includes examples of training activities led by, or involved CGIAR Centers, that have focused on the use of DSI in crop breeding. As mentioned in Box 3, ICRISAT's Center of Excellence in Genomic and Systems Biology (CEGSB) has long experience in raising national partners' capacities in bioinformatics and the application of "omics" technologies in breeding. Since this experience has been presented in Box 3 above, we won't mention it again in Box 7.

Box 7: Training national partners on the use of "omics" technologies in breeding

African Plant Breeding Academy (AfPBA)

The African Plant Breeding Academy has trained 151 plant breeders across 28 African countries in five cohorts over the span of last seven years for deploying genomics technologies in pre-breeding and breeding programs. These plant breeders work on 125 crops including 60 underutilized African crops and trees. This fully sponsored program by the academy is managed and governed by the University of California Davis (UCD) and hosted by the World Agroforestry (CIFOR-ICRAF) at Nairobi, Kenya. The candidates are selected in an open and highly competitive process by giving due consideration to gender and regional representation. The academy is primarily supported by Alliance for Green Revolution in Africa (AGRA), Mars Incorporated, and other philanthropic organizations.

Recently, the AfPBA, in partnership with Innovative Genome Institute (IGI) at the University of California, Davis and International Institute of Tropical Agriculture (IITA), has started a multi-year program to train 80-100 African scientists in genome editing technologies, specifically the CRSIPR. Currently, the first cohort of 10 scientists representing six African countries has completed two sessions at Nairobi, Kenya, and will complete the third and the last session by the end of 2023.

Hands-on Training on Molecular Techniques at IITA

The Bioscience Center of IITA at Ibadan, Nigeria, organizes a five-day 'Hands-on Training Course on Molecular Techniques' quarterly every year. These training courses target laboratory scientists, research technicians, graduate students, and lecturers in Molecular Biology across Universities and National Agricultural Research Institutes in Nigeria and other countries. The participants register by paying a fee of N90,000 ($100) that covers the cost of consumables, training, course materials and training certificate. The training involves both theory and practical sessions covering different topics including the fundamentals of research techniques, including real-time PCR/gene expression studies, DNA sequencing analysis, introduction to DArTseq and its applications, and other techniques. The practical sessions

include training on molecular biology techniques such as DNA extraction, Polymerase Chain reaction (PCR) analysis, sequencing, NanoDrop spectrophotometry analysis, and Agarose gel electrophoresis. In addition, the participants are trained on laboratory safety procedures, the use and proper handling of laboratory equipments, and basic biostatistics analysis of molecular data. At the end of the training, participants are evaluated based on their knowledge and experience from the training course.

In general, most of the students and even researchers coming from national universities and research organizations have only the theoretical knowledge. The quarterly training on Molecular Biology techniques was initiated in 2018 to provide hands-on training to young graduates including interns, National Youth Service Corp (NYSC) members, Masters and PhD students as well as lecturers or researchers from various African universities and national institutes, representing different disciplines, including molecular biology, genetics, biotechnology, biochemistry, and microbiology. The aim of this training is to equip participants with the basic knowledge and skills required to function in a molecular biology laboratory.

In addition, the Bioscience Center carries out individual trainings to young researchers who enrol themselves in Molecular Breeding and needs training on molecular biology techniques to complete their research work. The center thus provides individual training on specific areas of interest that helps these young researchers to get trained as well as finish their degree.

Training on CRISPR/Cas-9 gene editing technologies in India

Recently, ICRISAT's BioNcube[4] concluded a hands-on training program on 'CRISPR/Cas9-based Gene editing technologies in plants'. Twenty-three participants from different parts of India, including scientists, industry representatives and post-doctoral/doctoral students attended the training with national-level experts serving as resource personnel, providing valuable insights into the diverse tools, approaches, and applications of gene-editing technologies in agriculture. This highly popular training program is the third in the series since October 2022.

GWAS training for African scientists and students

AfricaRice with the support of USDA provided an online practical training to scientists from national agricultural research organizations and African students on the use of whole-genome genotyping data to identify genetic markers associated with specific traits through the Genome-Wide Association Studies (GWAS) approach. A total of 17 participants from West and East Africa, including 6 women and 11 men were involved in this training. The training was designed to equip participants with skills in the following areas: (a) Understanding the concept of GWAS; (b) Manipulating and filtering genomic and phenotypic datasets (c) Implementing different statistical GWAS models to identify candidate SNPs with significant p-values for highly associated traits (d) Utilizing GWAS results in crop genetic improvement research or pre-breeding. The training enabled participants to acquire the expertise necessary to effectively leverage GWAS techniques in their research endeavors.

Training postgraduate students in Ethiopia

A team of ILRI scientists working with staff at the ILRI forage genebank organized a series of postgraduate student training workshop on 'basic data wrangling in the R environment' for national partners in Ethiopia (Ethiopian Institute of Agricultural Research, Addis Ababa University, Wachamo University and

---

[4] BioNcube is an incubator for start-ups in the area of agricultural biotechnologies. BioNcube is supported by the Biotechnology Industry Research Assistance Council (BIRAC) of the Department of Biotechnology of India.

Ambo University). The workshop covered genetic diversity analysis, developing markers-trait associations, applying GWAS, QTL identification, and marker-assisted selection.

Basics of genomic selection

The training webinar "Basics in Genomic Selection", which is available online through the CGIAR Platform 'Excellence in Breeding', explains the methodology and presents the different scenarios where the methodology can be used are presented. Experts from the CGIAR system including CIMMYT, IRRI, and ICRISAT present case studies on the proper deployment of genomic selection in breeding programs. Software packages for genomic selection using dense molecular markers and pedigree can be obtained for free from CIMMYT.

Huge advances have been made in the last decade to develop and adopt breeding informatic solutions that facilitate the sharing, management and utilization of DSI along the breeding pipeline, from trait discovery to genomic selection. CGIAR Centers have been involved in the development of various applications that are freely available for adoption by any breeding programme. These applications come with huge amounts of DSI that were generated by CGIAR Centers and other organizations during the course of the application development, testing and utilization. Various CGIAR Centers support national agricultural research organizations from developing countries that are interested in adopting these applications for collaborative breeding.

Box 8: Open access platforms for modern and collaborative breeding

Breedbase

Developed by Boyce Thompson Institute together with IITA, CIP, the Alliance of Bioversity International and CIAT and other partners, Breedbase is an open-source, web-based application that provides breeding workflows, data management procedures, and analysis tools to address breeder informatics needs. The system allows users to upload and use genotypic and phenotypic data of various crops. Breedbase is based on Cassavabase, which was developed as part of NextGen Cassava, a project that sought to modernize cassava breeding in Africa through the use of cutting-edge tools. Cassavabase has accumulated an immense amount of cassava breeding data, consisting of information on more than 500,000 cassava accessions, characterized in over 4,000 trials, and nearly 35,000 genotyping experiments (Morales *et al.,* 2022; Agbona et al. 2023). Similarly, through YamBase, SweetpotatoBase and MusaBase a large number of yam, sweetpotato and Musa breeding data including genotyping as well as phenotyping has been deposited.

Enterprise Breeding System

The Enterprise Breeding System (EBS) is a comprehensive platform designed by the CGIAR rice, maize and wheat breeding programs to support collaborative, multi-country and multi-institute breeding networks. The vast majority of germplasm records have pedigrees traceable to founding accessions. It currently stores over 600,000 germplasm records and over 3 million trial datapoints for rice alone, and is the cornerstone for information management in the global CGIAR rice breeding partnership. The maize and wheat EBS instances also house substantial collections of germplasm information with over 1.7 million germplasm records for maize and over 8.4 million germplasm records for wheat. The availability of phenotypic data for maize and wheat is also increasing both from active usage of the system and from migration of historical data into the EBS. The EBS supports the collaborative exchange of information between CGIAR Centers working on the same crop and is also designed to enable data sharing within breeding networks.

Tools for Polyploids project

Several researchers across CGIAR Centers are involved in the Tools for Polyploids project. Funded by USDA NIFA under Speciality Crop Research Initiative, this project responds to the realization that polyploid specialty crops were not capitalizing on the rapid advances in genomics as were diploid crops due to the lack of computational tools needed to use the data in these complex genetic systems. This project coordinates efforts to resolve the common analytical issues faced by all polyploid crops across multiple plant families (Asteraceae, Actinidiaceae, Dioscoraceae, Convolvulaceae, Ericaceae, Poaceae, Rosaceae, Rubiaceae, Solanaceae) representing many economically important food, fodder, and ornamental crops throughout the world. The list of polyploid specialty crops used for food includes roots and tubers (potato, sweet potato, yam), fruit (strawberry, blackberry, blueberry, European plum, tart cherry, kiwi, persimmon, banana), vegetables (leek, watermelon), and others (coffee, basil, hops). This project aims to develop a suite of genomic/genetic/analytical tools and training materials that integrate genomic information into applied polyploid plant breeding programs. This project leverages on the existing datasets and current genomic tools available in rose, blackberry, tart cherry, strawberry, blueberry, potato, sweetpotato to train and adopt to other crops.

8. Strengths, weaknesses, and opportunities for capacity building

In the pursuit of its mission, CGIAR already plays a significant role in technology transfer, knowledge sharing, and capacity building for use of DSI in agricultural research and development in developing countries.

Research and development partnerships between CGIAR Centers and Initiatives and national agricultural research organizations, universities and other actors are perhaps the most effective means for effective tech transfer, capacity building and knowledge sharing in both directions. These partnerships entail direct interaction between experts from CGIAR Centers and national research organizations, facilitating effective knowledge transfer and mentorship. Partnering with national institutions has helped to ensure that capacity building components of projects are tailored to the specific needs and priorities of the respective countries or regions. This approach promotes ownership and sustainability, as the acquired skills can be integrated into national research agendas and long-term breeding programs.

In addition, CGIAR Centers and Initiatives have developed and coordinated capacity strengthening programs in the form of training programs, workshops, and knowledge-sharing activities that aim to enhance the understanding and application of genomics techniques. They include training in areas such as genotyping, diversity analyses, bioinformatics, marker-assisted selection, and genomic data analysis such as GWAS or Genomic selection, data management as well as related policies for access and benefit sharing.

Some of the trainings and other capacity strengthening activities highlighted in this submission are organized within the scope of specific projects or grants, and as such, represent isolated efforts that are not integrated into a long term, broad-based capacity strengthening programs. To ensure sustainable capacity building to close capacity building gaps in line with the expectations of many contracting parties

and stakeholders engaged in discussions at the level of the Plant Treaty's governing body, and the Convention on Biological Diversity, a broader, more long-term, coordinated approach is required, building on evidence-based analyses of capacity strengthening needs of research and development organizations in developing countries. CGIAR looks forward to the additional clarity that will result from contracting parties' and other stakeholders' submissions concerning their capacity strengthening needs to be better able to generate, access, and use DSI. It is our hope that, with a sufficiently well-defined set of priorities developed by the Governing Body, CGIAR will be able respond by tailoring its own contributions to address those prioritized needs.

References

Agbona, A. et al. (2023). Data Management in Multi-disciplinary African RTB Crop Breeding Programs. In: Williamson, H.F., Leonelli, S. (eds) Towards Responsible Plant Data Linkage: Data Challenges for Agricultural Research and Development. Springer, Cham. https://doi.org/10.1007/978-3-031-13276-6_5

Bredeson, J.V., Lyons, J.B., Oniyinde, I.O. *et al.* 2022. Chromosome evolution and the genetic basis of agronomically important traits in greater yam. *Nat Commun*. 13, 2001. https://doi.org/10.1038/s41467-022-29114-w

CBD Secretariat (2020a) Digital Sequence Information on genetic resources: concept, scope and current use. CBD/DSI/AHTEG/2020/1/3. https://www.cbd.int/doc/c/fef9/2f90/70f037ccc5da885dfb293e88/dsi-ahteg-2020-01-03-en.pdf

CGIAR, 2023. 'Issues for further consideration concerning digital sequence information: a submission from CGIAR', available at https://www.cbd.int/notifications/2023-003

Mansueto, L., Fuentes, R.R., Borja, F.N., *et al*. 2017. Rice SNP-seek database update: new SNPs, indels, and queries. Nucl Acids Res 45:D1075-D1081. DOI: 10.1093/nar/gkw1135.

Morales, N., Ogbonna, AC., Ellerbrock, BJ. *et al*. 2022. Breedbase: a digital ecosystem for modern plant breeding. G3 Genes|Genomes|Genetics; doi: jkac078.

Njaci, I., Waweru, B., Kamal, N. *et al.* Chromosome-level genome assembly and population genomic resource to accelerate orphan crop lablab breeding. *Nat Commun* 14, 1915 (2023). https://doi.org/10.1038/s41467-023-37489-7

Sackville Hamilton, R.; Halewood, M.; Lopez Noriega, I.; Rouard, M.; Tiambo, C.; Azevedo, V.; Bhattacharjee, R.; Cuellar, W.; Gisel, A.; Hearne, S.J.; Heider, B.; Hendre, P.; Kreuze, J.; Muchugi, A.; Mwai, O.; Ndjiondjop, M.-N.; Odeny, D.; Patel, E.; Perset, E.; Platten, J.; Ramaiah, V.; Roa, A.C.; Singh, K.; Wenzl, P.; Vermeulen, S.; Zhang, X. (2022) Digital sequence information is changing the way genetic resources are used in agricultural research and development: implications for new benefit-sharing norms. A discussion paper from CGIAR for consideration by delegates to the 15th Session of the Conference of the Parties to the Convention on Biological Diversity. CGIAR Genebank Initiative, available at https://cgspace.cgiar.org/handle/10568/125749

Scholz et al. 2021. Myth-busting the provider-user relationship for digital sequence information, *GigaScience*, Volume 10, Issue 12. https://doi.org/10.1093/gigascience/giab085

Tamiru, M., Natsume, S., Takagi, H. *et al.* 2017. Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination. *BMC Biol* **15**, 86 https://doi.org/10.1186/s12915-017-0419-x

Varshney, R.K., Roorkiwal, M., Sun, S. *et al.* 2021. A chickpea genetic variation map based on the sequencing of 3,366 genomes. *Nature* 599,. https://doi.org/10.1038/s41586-021-04066-1

Wu, F., Duan, Z. et al. 2021. Genome and systems biology of *Melilotus albus* provides insights into coumarins biosynthesis. *Plant Biotechnology Journal* Volume 20, Issue 3

Yan, Q., Wu, F., Xu, P., Sun, Z., Li, J., Gao, L., et al. 2021. The elephant grass (Cenchrus purpureus) genome provides insights into anthocyanidin accumulation and fast growth. Mol. Ecol. Resour. 21, 526–542. doi:10.1111/1755-0998.13271.

Yssel, A.E.J.; Kao, S.-M.; Van de Peer, Y.; Sterck, L. 2019. ORCAE-AOCC: A Centralized Portal for the Annotation of African Orphan Crop Genomes. *Genes*, *10*, 950. https://doi.org/10.3390/genes10120950

Contributors