E

Food and Agriculture Organization of the United Nations

International Treaty on Plant Genetic Resources for Food and Agriculture

INTERNATIONAL TREATY ON PLANT GENETIC RESOURCES FOR FOOD AND AGRICULTURE

TWELFTH MEETING OF THE AD HOC OPEN-ENDED WORKING GROUP TO ENHANCE THE FUNCTIONING OF THE MULTILATERAL SYSTEM

Rome, Italy, 16–19 September 2024

Policy brief on digital sequence information/genetic sequence data: Generation, Use and Sharing of Digital Sequence Information in Crop Improvement

Note by the Secretary

At its Tenth Session, the Governing Body requested the Co-Chairs to give early attention to three identified "hotspots", one of which is digital sequence information/genetic sequence data (DSI/GSD). The Governing Body also took note of decision 15/9 of the Conference of the Parties to the Convention on Biological Diversity (CBD) on Digital Sequence Information on Genetic Resources and urged the Working Group to take this decision and related developments into account when addressing the issue in the context of the process to enhance the functioning of the Multilateral System.

At its eleventh meeting, the Working Group agreed to consider the possibility of developing a specialized approach for DSI/GSD on PGRFA under the International Treaty, while monitoring the relevant processes under the CBD, to ensure mutual supportiveness. Any solution should not restrict facilitated access to PGRFA or open access to DSI/GSD on PGRFA and should seek to exclude double payments by users.

This document contains a policy brief on DSI/GSD on plant genetic resources for food and agriculture commissioned by the Co-Chairs to further inform the discussions of the Working Group.

The policy brief was prepared by the CGIAR Initiative on Genebanks and the Secretariat of the International Treaty.

Food and Agriculture Organization of the United Nations

International Treaty on Plant Genetic Resources for Food and Agriculture

CGIAR INITIATIVE ON Genebanks

TWELFTH MEETING OF THE AD HOC OPEN-ENDED WORKING GROUP TO ENHANCE THE FUNCTIONING OF THE MULTILATERAL SYSTEM

GENERATION, USE AND SHARING OF DIGITAL SEQUENCE INFORMATION IN CROP IMPROVEMENT

POLICY BRIEF

Rome, 16–19 September 2024

This policy brief was commissioned by the Co-chairs of the Working Group to Enhance the Functioning of the Multilateral System of Access and Benefit-sharing under the International Treaty on Plant Genetic Resources for Food and Agriculture (International Treaty), with the aim of informing international discussions on the sharing of benefits arising from the use of digital sequence information (DSI) in the agriculture sector. The document was prepared by the CGIAR Initiative on Genebanks and the Secretariat of the International Treaty. Digital sequence information is transforming research, conservation and breeding for sustainable agriculture and food security. This policy brief seeks to shed light on the benefits and challenges that the current information revolution is generating.

INTRODUCTION

Each individual species has a distinctive genome, which is the entire set of DNA instructions found in a cell. The genome varies enormously among species. For example, in humans, the genome consists mainly of 23 pairs of chromosomes; common wheat's genome is organized in 7 groups of chromosomes, each group containing a set of 6 chromosomes in 3 pairs originating from 3 different ancestors.

Sequencing the genome means determining the order of the four chemical building blocks known as 'bases' that make up the DNA molecule. DNA sequences are unique from one organism to the next. Scientists analyse genetic sequences to discover which stretches of DNA contain genes, and which stretches carry regulatory instructions, turning genes on or off, and to differentiate between those regions of genes that encode proteins (coding regions) from those that do not. In addition, and importantly, sequence data can highlight differences between genes that may translate into different characteristics or traits in the organism.

Plant varieties (including varieties developed by farmers, research organizations and private companies) and their wild ancestors represent combinations of genetic sequences that underpin the traits of each particular variety in interaction with the environment where the variety grows. While a variety is unique as a whole, portions of genetic sequences (such as those coding an early-maturing or late-maturing trait) may be the same in many different varieties. The reverse can also occur: the same trait, for example, early-maturing, can also be under the control of different sequences in different varieties. Traits interact with the environment where the plant grows, and hence the expression of any gene can be widely influenced by the environment. The uniqueness of a variety comes from the combination of those genetic sequences and the environment's influence. The combination of DNA sequences in each particular variety is the result of hundreds and thousands of years of directed and random selection, by the environment, by farmers and by plant breeders.

In crop research and development, the generation, analysis and use of genetic sequence data or digital sequence information are inextricably linked to the management of plant genetic material,[1] or plant genetic resources for food and agriculture (PGRFA), see **Figure 1**. Researchers and breeders rely on a wide diversity of plant genetic materials to generate

[1] The definition of DSI could be much broader, but this brief uses the term 'digital sequence information' to refer to genetic sequence information.

reference genomes representing intra-species diversity to understand the relationship between DSI and traits across a wide array of different genotypes, and eventually to integrate the desired traits from promising genotypes into selected background varieties.

As climate, pests and diseases change, as well as consumer preferences, researchers and breeders need to constantly go back to PGRFA to decode the genetic instructions related to the traits that can respond to those changes, and use the related DSI, together with other information, in the generation of new varieties. Digital sequence information is not a final product, but rather an additional and increasingly important research and breeding tool. The ultimate goal is to generate novel PGRFA material, in the form of plant varieties that are made available to farmers as seeds and other planting materials.

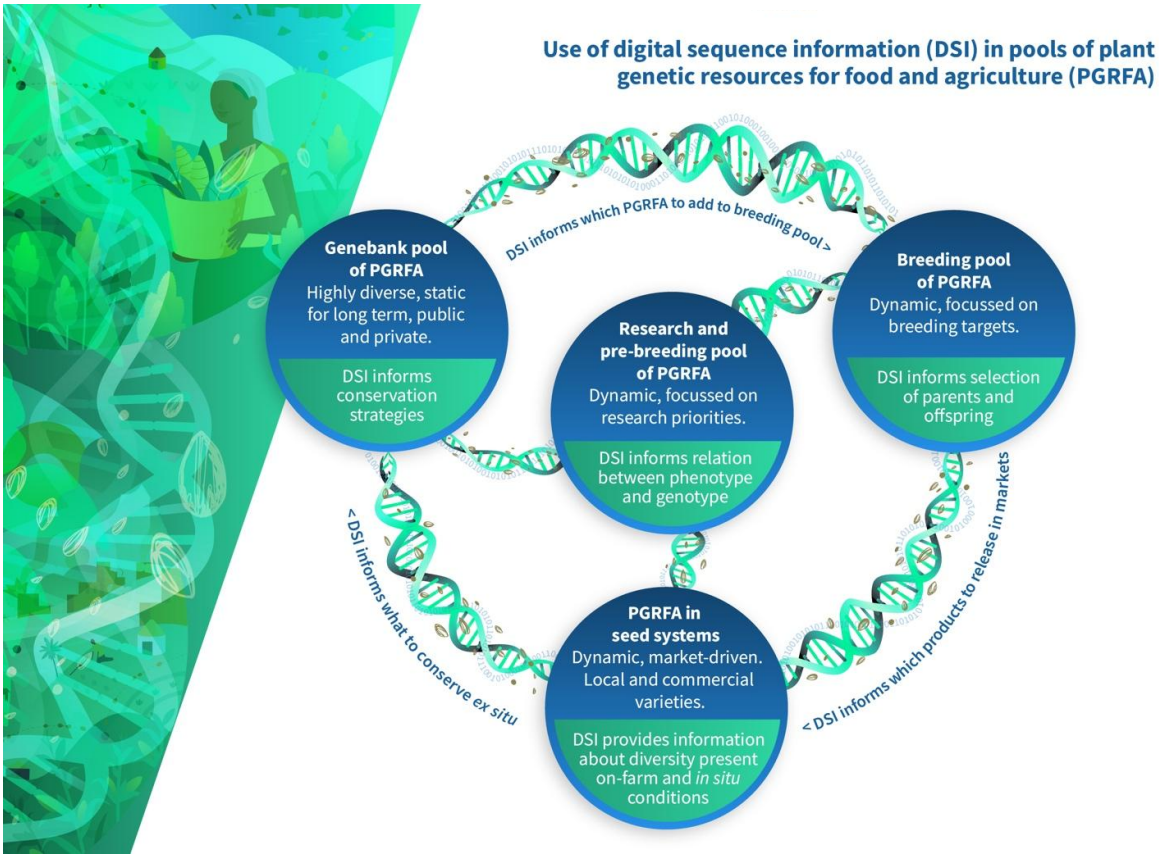

FIGURE 1: Use of digital sequence information in pools of plant genetic resources for food and agriculture
Source: authors' own elaboration, illustration by Dave Gray.

Digital sequence information is always used with other types of data. While genome sequencing and genetic fingerprinting may help to distinguish 'what is the same' and 'what is different' in genetic terms, a robust understanding of the species' physiology and good morphological data are needed to fully interpret how DNA sequences translate into different characteristics in plants. Most traits, particularly those related to abiotic stresses, are under complex genetic control involving multiple forms of multiple genes interacting in networks. For example, a crop's ability to tolerate drought depends on the anatomy and architecture

of roots, leaves and stems; the rate of progress through the life cycle in relation to the development of drought; and difficult-to-measure attributes of photosynthetic, respiratory and other biochemical and physiological capacities of the plant. Digital sequence information is therefore of limited use without other types of PGRFA information, in particular good phenotypic information – information about observable characteristics of the organisms, including appearance, development and behaviour in relevant environments. Phenotypic characterization and evaluation of PGRFA continue to be a fundamental step in assessing variability across individuals in relation to the traits of interest, to inform decisions on what to sequence and what for, and to interpret genetic information.

Digital sequence information of pathogens and soil microorganisms is also relevant for crop research and improvement, but it does not come within the scope of this paper.

HOW DSI IS USED IN CROP IMPROVEMENT

During more than ten millennia of farming, varieties and breeds have been improved based on phenotypic selection of parents and offspring that resulted in the alteration of their genomes. The process of developing superior new products in this way is slow, costly and inefficient vis-a-vis the requirements and demands of modern commercial agriculture. It is also potentially unreliable because phenotype varies with age, environment and other factors, and because phenotypic traits can be a function of multiple interacting genes, resulting in complex and potentially non-obvious inheritance.

By contrast, the genotype is fixed and can be determined at birth without having to wait months or years for the trait to be expressed in an appropriate organ and developmental stage of the organism grown in a given environment. Hence, given sufficient knowledge of the genetic control of the traits of interest, research organizations and commercial companies can develop new products more rapidly, more efficiently and more effectively by basing selection directly on genotype rather than on the resulting phenotype. However, this first requires determining the relationship between genotype and phenotype, which is an arduous cumulative process involving multiple steps over time.

PRE-BREEDING PHASE: UNDERSTANDING THE RELATIONSHIP BETWEEN GENOTYPE AND PHENOTYPE

Sequencing a whole species genome accurately from scratch is extremely time-consuming and expensive. Although time and costs have been reduced dramatically in recent years, it remains a significant investment to undertake on a large scale. While the cost of sequencing is relatively low today, the bioinformatic work involved in assembling the genome can be very expensive, particularly for species with complex genomes, such as wheat.

A single genome that has been sequenced with high accuracy and reliability may be used as a ‘reference genome’. High-quality reference genomes are near error-free, gapless sequences and include detailed information about the genome's structure. For most crops, given the huge magnitude of within-species diversity, more than one high-quality genome sequence selected from across the full spectrum of diversity may be necessary to provide a

more appropriate choice of reference genomes. For example, there are now 15 reference genomes for rice. Due to the time and resources required, the generation of whole genome sequences often represents the end goal of a research activity that takes many years, involving a consortium of research organizations. Once created, these reference genomes provide the basis for subsequent work – often by other research teams and frequently in other parts of the world. For example, the potato reference genome, first published in 2011 in *Nature*, has been referenced in 1 458 publications over the course of the past 13 years. Such work takes advantage of the reference genome as a basis for assembling and comparing a much higher number of genomes (but of lower quality) of the same species or closely related ones through a process known as resequencing ('next generation sequencing', or NGS).

Resequencing techniques include genotyping: the process of determining genetic variants or polymorphisms in an individual. A number of NGS-based genotyping approaches do not use any individual reference in the identification of specific variation within a group of samples, but compare sequences of the samples within the group.

Data on genome sequences are only of value once they have been 'decoded' to identify genetic variants, and their associations with phenotypic data. First, genome annotation seeks to find and designate locations of individual genes, non-coding and regulatory sequences on raw DNA sequences. Once the location of genes has been predicted, allele mining helps in tracing the evolution of alleles (alternative forms or versions of a gene), the identification of new haplotypes (groups of alleles that are inherited together) and the development of allele-specific markers (DNA sequences that are in proximity or tightly linked to a particular gene). Allele mining requires sophisticated bioinformatic tools that allow sequence alignment to compare the new sequences with the reference genome and/or among themselves.

Genome-wide association studies (GWAS) can quickly detect DNA markers for target traits. Furthermore, when genome sequences have been resolved into genes, this technique allows the preliminary identification of candidate genes' controlling traits. Genome-wide association studies rely on both genotypic and phenotypic data to identify associations between genotypes and phenotypes by testing for differences in the frequency of genetic variants among samples that are genetically similar and phenotypically different. Large numbers of samples are used in GWAS, preferably more than 100, and ideally more than 300. Facilitated access to crop germplasm from many gene banks has been crucial to assembling the wide and diverse sets of samples that are necessary for GWAS. GWAS can consider different types of sequence variation in the genome, although the most commonly studied genetic variants in these studies are single-nucleotide polymorphisms (SNPs) – a difference in a single nucleotide, the DNA building block. GWAS typically report blocks of correlated SNPs that show a statistically significant association with the trait of interest.

GWAS can make use of the same DSI dataset for multiple traits in multiple environments. For this reason, there has been a massive increase in demand for PGRFA whose DSI is available in public databases, when those PGRFA accessions and their sources are clearly identified in the databases. Researchers can conduct phenotypic trials with those same PGRFA for the traits and in the environments of interest to them, and then carry out the

GWAS, combining their own phenotypic data with the DSI obtained from public databases. This allows them to detect the DNA markers of relevance to them and eventually identify candidate genes associated with desirable traits. As an example, the Global Rice Phenotyping Network sought to facilitate exactly this *modus operandi*. The network was established under the CGIAR Research Program 'Global Rice Science Partnership' from 2011 to 2016 and enabled multi-partner and multi-country phenotyping of panels of accessions whose genomes had been previously genotyped by the 3000 Rice Genomes project. GWAS were carried out with the network's phenotypic data to shed light on the genetic architecture of complex traits, such as high night temperature tolerance and drought tolerance at the reproductive stage. In recent years, scientists have carried out hundreds of GWAS for a wide range of crops, most of them major crops. The results are available in journal articles and open databases.

Of course, researchers can add new genetic resources (for example, local farmers' varieties) that were not previously sequenced to the pool of materials they wish to study in the GWAS. To do so, they need to generate sequence data for those newly introduced varieties to add to the sequence dataset that they will use in the study to analyse associations between phenotype and genotype.

Finding which portion or portions of the genome actually determine a trait is a major challenge and typically involves seeking convergent evidence of multiple kinds. This can include, for example, studies of developmental genetics, physiology and biochemistry, analysis of genetic and environmental effects on gene expression (transcriptomics and proteomics), knowledge of gene function in related species, and mapping studies.

Once candidate genes have been identified, gene editing can be used to knock out (or disable) them and thus prove if these genes do indeed control the target trait. Gene editing entails the modification of a nucleotide using sequence-specific enzymes which act as 'molecular scissors' to create double-strand DNA breaks, which are then repaired by the cell's own DNA repair mechanism.

BREEDING PHASE: SELECTING GERMPLASM BASED ON ITS GENETIC MERIT

The application of genomic tools in plant breeding, generally termed genomics-assisted breeding, has progressed through various stages in the past few decades.

The simplest and oldest approach based on DNA analysis dates back to the 1980s, with the discovery of molecular markers and their application in marker-assisted selection. The first markers to be used, termed RFLPs (which stands for restriction fragment length polymorphism), enabled genetic variants to be revealed at random positions spread across the genome. RFLPs have since been replaced with other genotyping technologies, such as SSRs, AFLPs, RAPDs, SNPs, DArTs and GBS, but the principle remains the same. The genetic variants, while typically not part of a gene, may be physically close to, and thus tend to be inherited with, a gene controlling a trait of interest. Such a genetic variant can be an indicator, or marker, of the likely presence of the trait.

Markers are typically identified at the pre-breeding stage, as described above. They are widely used at various steps of the breeding process to select individuals with favourable genes. Early in the breeding cycle, markers make it possible to reduce the number of plants that are going to be evaluated in field trials, since breeders can select only those samples that have the superior alleles of genes controlling the traits of interest. This allows them to reduce the time and costs involved in phenotyping and evaluation.

The original approach to marker-assisted selection has major limitations. Developing and validating the markers is slow and expensive and has to be repeated for each trait of interest. The association between the marker and the trait is a correlation, which does not necessarily imply causation, and the correlation is prone to breakdown in other genetic resources. Even where the association is causal, the effect of a gene depends on the composition of the rest of the genome. For both reasons, such markers are typically effective only with the specific, small breeding populations for which they were developed, and must be redeveloped and validated for different breeding populations. The weaker the genetic association, the greater the functional interactions between genes and the narrower the range of efficacy. Further, since the approach targets a single gene-trait combination, markers are useful for identifying quantitative trait loci (QTLs) – genetic regions that influence phenotypic variation of a complex trait – that have major effects on high-value traits, but they are of limited use at the time of identifying QTLs with small effects on phenotypic variation. This hinders the complete understanding of the genetic architecture of many complex traits.

Most recently, genomic selection has combined whole-genome molecular marker data with phenotypic and pedigree data to estimate breeding value through genomic prediction, as shown in **Figure 2**. In contrast to conventional marker-assisted selection, genomic prediction employs a large number of genome-wide SNPs to quantify the comprehensive genetic merit of individual plants encompassing most contributing QTLs of a target trait, including QTLs that have small effects on the trait variation. Genomic selection reduces the cost per breeding cycle, increases selection intensity and accuracy, and significantly reduces the time required to develop a cultivar, compared with phenotypic-based selection. A critical step towards the implementation of genomic selection is the establishment of the training population set: the breeding lines that have been phenotyped for target traits and genotyped with genome-wide markers and that will be used to train a prediction model. Typically, this training population will be made of lines that are closely related to those used in the breeding programme. Once trained, this model is used to predict performance on a test population based solely on genotypic information by calculating genomic estimated breeding values. The prediction model will need to be updated when new germplasm is included in the training population. Developing statistical machine-learning models and training population optimization are the two main thematic areas actively explored in plant genomic prediction.

Genomic selection requires considerable bioinformatics capacities and high-performance computing. For this reason, only organizations that have sufficient resources currently apply this technique, and they do it mostly for crops that will compensate the necessary initial

investments. Markers continue to be widely used in the traditional way by many public and private organizations, and for a wide range of crops, increasingly assisted by GWAS.

Recently, a new approach called haplotype-assisted forward/backward breeding has emerged, where superior haplotypes are identified and combined to create tailor-made varieties for crop improvement programmes.

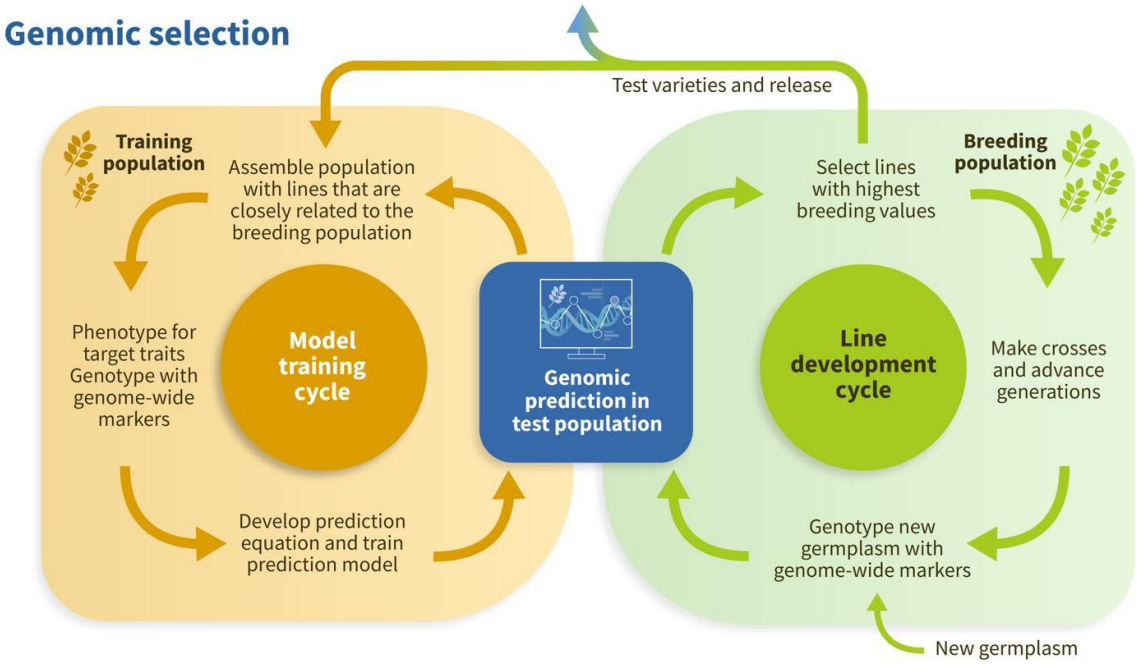

FIGURE 2: Genomic selection.
Source: Figure adapted by authors from Heffner, Sorrells and Jannick, 2009. Illustration by Dave Gray.

Thanks to all the knowledge that has been made available at early stages, breeders are much less dependent on PGRFA and phenotyping data at the time of assembling the breeding population and selecting the most suitable parents for developing new lines. The DSI that was generated from accessing thousands of accessions will allow them to select within a few parents with the highest breeding values, which will be then used to generate new plant varieties. Most plant genetic resources from which DSI and other information were derived is not utilized in the development of new plant varieties. They are used mostly for analysis and comparison purposes. Only a tiny fraction of DNA sequences that were decoded in previous steps will find its way into the new plants, but the value of that tiny fraction can only be realized as a result of all the previous work conducted. The benefits lie in the knowledge gained through the upstream comparison of DSI of different materials, and the subsequent development and use of genomic tools for more effective production of new and better cultivars.

Therefore, generating and utilizing DSI in crop improvement is a resource-intensive work that pays off at the breeding stage. Digital sequence information allows breeders to save time and money. For plant breeders, it is in these savings that the economic value of DSI lies.

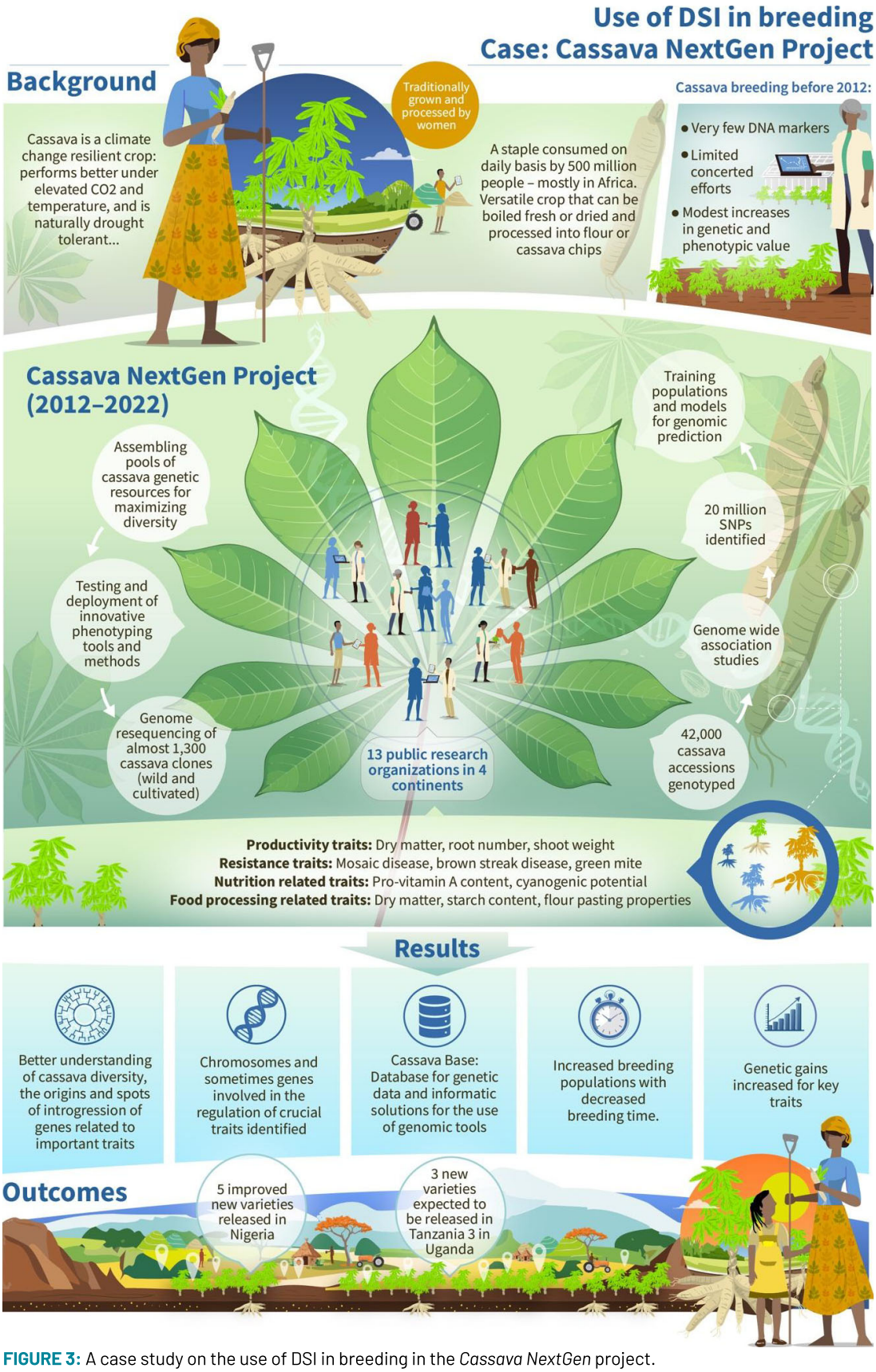

FIGURE 3: A case study on the use of DSI in breeding in the *Cassava NextGen* project.
Source: authors' own elaboration, illustration by Dave Gray.

WHO GENERATES AND USES DSI?

GENOME SEQUENCING CONSORTIA

Given the complexity and the resources involved in generating and assembling high-quality whole genome sequences, public and private organizations from different countries pool financial, human and technical resources in large genome sequencing consortia. The International Wheat Genome Sequencing Consortium, which is one of the largest consortia, involves 884 research institutes and private companies from 71 countries. Data from these consortia include full genome sequences of a small number of varieties that represent the species' diversity, genome annotations, and sometimes genomic tools. Most of these resources are made available in public databases. Consortia members, as well as other organizations, use these resources extensively to compare and interpret the data coming from their own resequencing or genotyping efforts (see case study in **Figure 4**).

PUBLIC ORGANIZATIONS AND PUBLIC-PRIVATE PARTNERSHIPS

In countries with a strong research and development base in plant breeding (i.e. most developed countries and some middle-income countries such as Brazil, China and India), public agricultural research organizations play a major role in pre-breeding work, including genotyping and phenotyping of PGRFA conserved in national gene banks for traits of interest. Following common scientific practice, these organizations usually publish their research results, including DSI, in journal articles and public databases. These organizations represent the major providers of data found in such sources.

Public-private partnerships for pre-breeding are common in developed countries and are often involved in genotyping, phenotyping and evaluating medium-to-large collections of PGRFA, with the objectives of identifying promising genotypes for breeding, generating easy-to-use markers and developing genomics-based breeding approaches and methods. These partnerships tend to focus on crops of commercial importance, such as major cereals, certain legumes and forages, horticultural crops and temperate fruits, and are less common for crops that generate smaller revenues in the seed market. Genotyping and phenotyping efforts focus on traits of potential commercial value, and such traits therefore determine the types and volumes of PGRFA that are sequenced through the partnerships.

There are multiple examples of such partnerships. Some are stable institutions that have been active for many years (such as the Center for BioSystems Genomics for potato, tomato, *Arabidopsis* and *Brassica* in the Netherlands); others are associated with time-bound projects (such as the CitruSeq-CitrusGenn consortium for citrus species and the EUCLEG project for forage and grain legumes). Public-private partnerships are also sought when the objective is to generate DSI and phenotypic data for traits, species or geographical areas whose commercial value is small or uncertain, and which will therefore attract limited private investment for upstream research. The Nordic Public-Private Partnership (PPP) for pre-breeding is a good example of this.

Although data management and sharing differs from partnership to partnership, typically, all partners have access to all the DSI generated within the consortium. However, not all DSI may be made accessible to the public in the short term, particularly genome annotations and markers. The partners' institutional policies and the conditions requested by the funding organizations largely determine the data-sharing modalities of each partnership. Sometimes the partnerships' terms and conditions request open access to the resulting DSI, but not before the members have published their work in a scientific journal. Sometimes the terms and conditions do not explicitly request the sharing of DSI in public databases, and this is left to the discretion of the partnership's members.

Generally, these public-private partnerships do not extend to the breeding phase: private companies develop and commercialize their own cultivars independently.

PRIVATE COMPANIES

Efforts invested in DSI generation and use of DSI in crop improvement in the breeding programmes of private companies vary from company to company, depending on the size and technical capacities of the company, the crops of interest and the target geographies.

Some of the biggest multinational seed companies are members of certain crops' genome sequencing consortia, but they rarely engage in public-private partnerships for genotyping and phenotyping. For their breeding programmes, they rely mostly on the phenotypic data and the genotypic data that they or other organizations contracted by them generate using the companies' own PGRFA. This is particularly true for major cereals such as wheat and maize. These seed companies occasionally obtain PGRFA from other sources. If in-house characterization shows valuable traits, the companies do the genotyping and apply the resulting DSI in their own breeding programmes. These companies may use journal articles and public databases to screen potential sources of desired traits, but in most cases they do not use the data as such in their breeding work. Instead, they access the PGRFA from external collections and do their own internal data generation for the development and application of genetic makers on the materials that they themselves manage.

Private companies maintain their own private databases, containing DSI and other information associated with their breeding programmes. In some cases, research results may be published in journals and patent applications, with associated DSI contributed to public databases.

INTERNATIONAL AGRICULTURAL RESEARCH CENTRES

International agricultural research centres partner with national agricultural research organizations for genetic and phenotypic characterization of PGRFA in the context of conservation and breeding initiatives funded by a wide range of donors in developing countries. In these partnerships, public research organizations contribute their own PGRFA, related information, human resources and research lands and facilities, and they receive DSI and other types of information, as well as technologies and know-how for the utilization of

DSI in crop improvement. Occasionally, private companies are also involved in these partnerships.

Based on CGIAR-wide policies, DSI that has been generated by CGIAR Centers, alone or in partnership with other organizations, must be made available in public databases. International centres use these and other public data in their breeding programmes, and the resulting improved lines are made available to public and private organizations for further breeding, or testing and release. Small- and medium-sized companies in developing countries play a major role in the commercialization of these varieties. Other organizations from the public and private sector also use these public data to develop their own varieties for commercialization in their target geographies.

DSI is also used for the rational, efficient and effective conservation and use of biological diversity in *ex situ* collections.[2]

[2] More information about the use of DSI for PGRFA conservation can be found in the following publication: Sackville Hamilton, R. *et al*. 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. https://hdl.handle.net/10568/125749

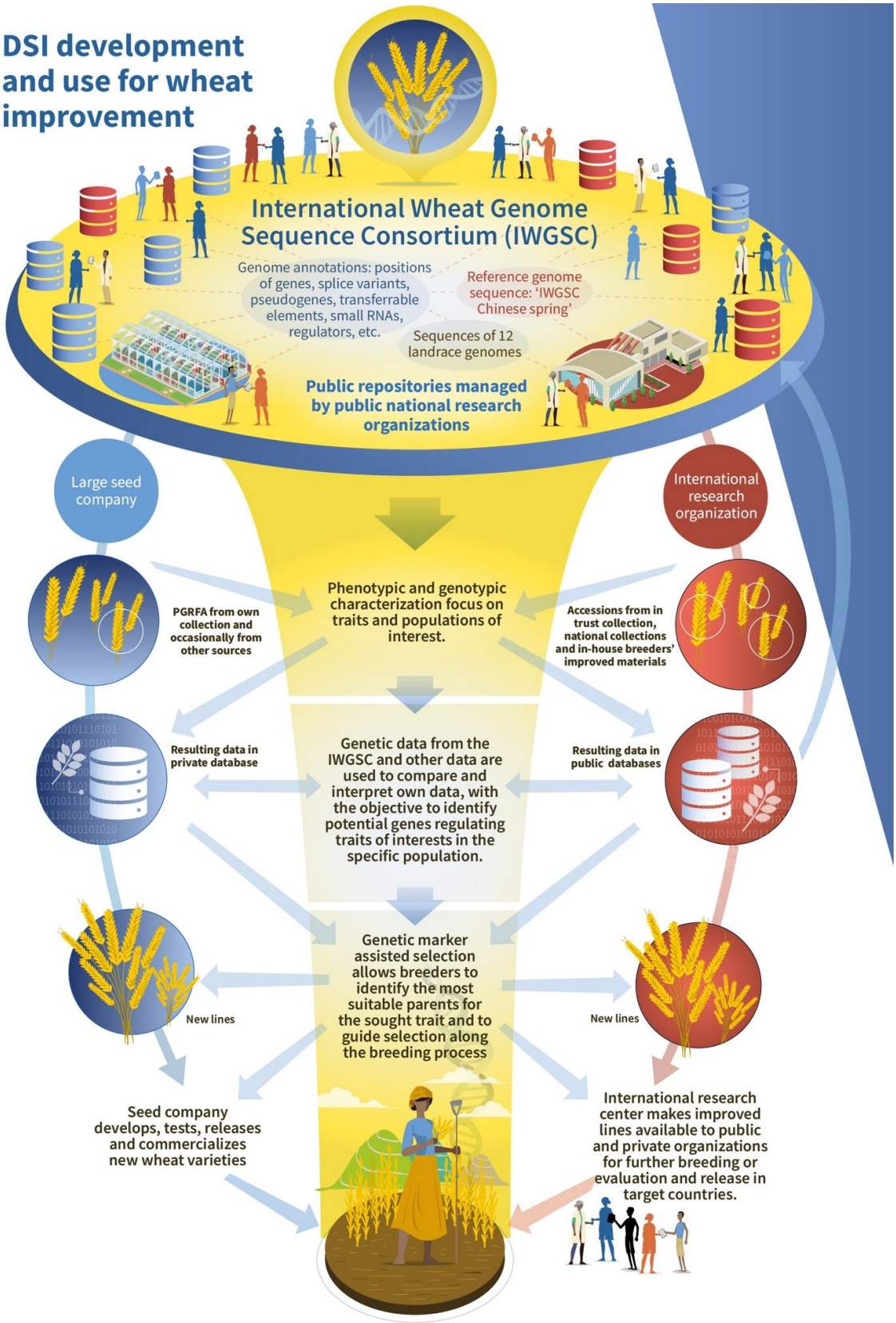

FIGURE 4: Case: DSI development and use for wheat improvement.
Source: authors' own elaboration, illustration by Dave Gray.

LANDSCAPE OF AGRICULTURE-RELATED DSI DATABASES

Massive amounts of genetic data are generated at all the stages described above, and they are stored and managed in databases. The current landscape of biological and scientific databases forms a complex ecosystem, and those dealing with DSI of crops and other edible plants are no exception (see **Figure 5**). At the core, generalist databases provide some level of stability, but there is significant variation among secondary databases. These secondary databases are highly specialized, building on raw data and focusing on specific datasets such as DNA, RNA, protein, or metabolomics for multiple purposes (such as genetic diversity, genome evolution, pre-breeding, breeding). Some are designed to be multispecies, leveraging the power of comparison between organisms across kingdoms. Others are crop-specific, such as crop hubs that aim to centralize as much data as possible for various uses. Many of these databases exhibit discontinuity in data storage and provision due to limited and unstable funding. Given the volume and rapid pace of DSI generation, even the most stable databases become overwhelmed and struggle to integrate all raw and intermediate datasets promptly.

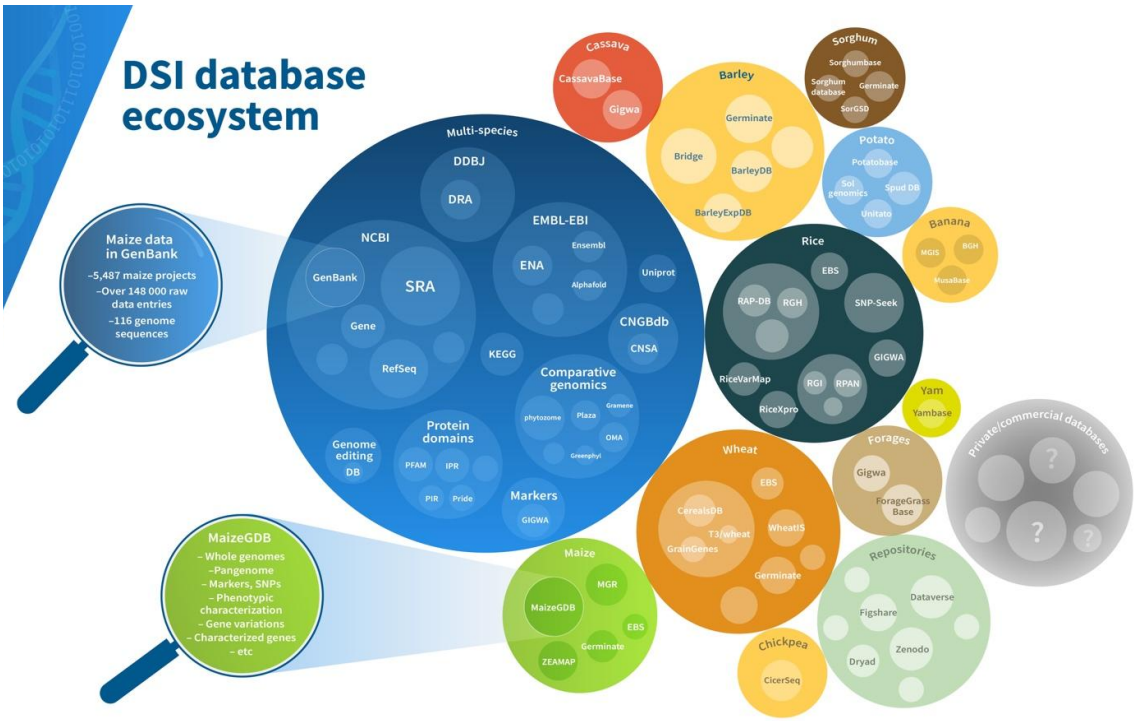

FIGURE 5: DSI database ecosystem.
Source: authors' own elaboration, illustration by Dave Gray.

Besides, many datasets, often supporting peer-reviewed publications, are deposited in repositories such as Zenodo, Dataverse, FigShare and Dryad, without additional curation. The lack of standardization and integration of files in these repositories further adds to the complexity, preventing seamless incorporation into the broader ecosystem. As a result, data are scattered across multiple locations and countries (mostly in high-income countries –

the United States of America, the United Kingdom, the European Union, Japan – and China), tailored to different communities and uses. With advances in genotyping, data can easily become obsolete and be superseded by new data that is of better quality, more precise or more reliable.

In most cases, raw sequence data are made available through public, collective databases, primarily those of the International Nucleotide Sequence Databases Collaboration (INSDC), which includes the National Center for Biotechnology Information (NCBI) of the United States National Institutes of Health. Gene annotations, markers, and other essential tools for breeding and genetic diversity conservation can be found in secondary, specialized databases.

If we take maize (*Zea mays*) as an example, as of August 2024, around 5 487 maize BioProjects were deposited in NCBI, with over 148 000 raw data entries associated with 121 654 samples. In NCBI terminology, a BioProject is the entire set of data associated with a particular project or study. In addition, NCBI provides access to the raw sequences of high-quality assemblies of 116 maize genomes, including the B73 inbred maize line's genome, the first maize genome that is still used as reference for the crop.

The specialized platform MaizeGDB hosts sequences of high-quality assemblies of 103 maize genomes along with related gene annotations, including the 25 founder lines for the maize Nested Association Mapping (NAM) population, which were chosen to represent a broad cross-section of the maize lines used in modern breeding. Additionally, MaizeGDB offers dedicated interfaces and curated data that are particularly useful for maize breeding, including:

- Variant data from a diverse set of 1 498 inbred lines
- SNP markers associated with traits
- Phenotypic characterization of 1 120 phenotypes
- Multi-genomic gene expression analysis
- Newly characterized genes from biological experiments
- Maize pan-genome – a pan-genome represents the entire set of genes within a species, consisting of a core genome (i.e. sequences shared between all individuals of the species) and the 'dispensable' genome (i.e. genes that are shared by some but not all individuals in the same species).
- Gene effects from to 2.3 million natural variations
- Maize PeptideAtlas – a compendium of proteomic data

Some organizations, including CGIAR Centers, have put in place their own data platforms to facilitate access to data of a particular species, or groups of species. Having their own data platforms, for example for bananas (see **Figure 6**), allows these organizations to arrange and provide the data, particularly the genotyping data, in an optimal way for their target audiences.

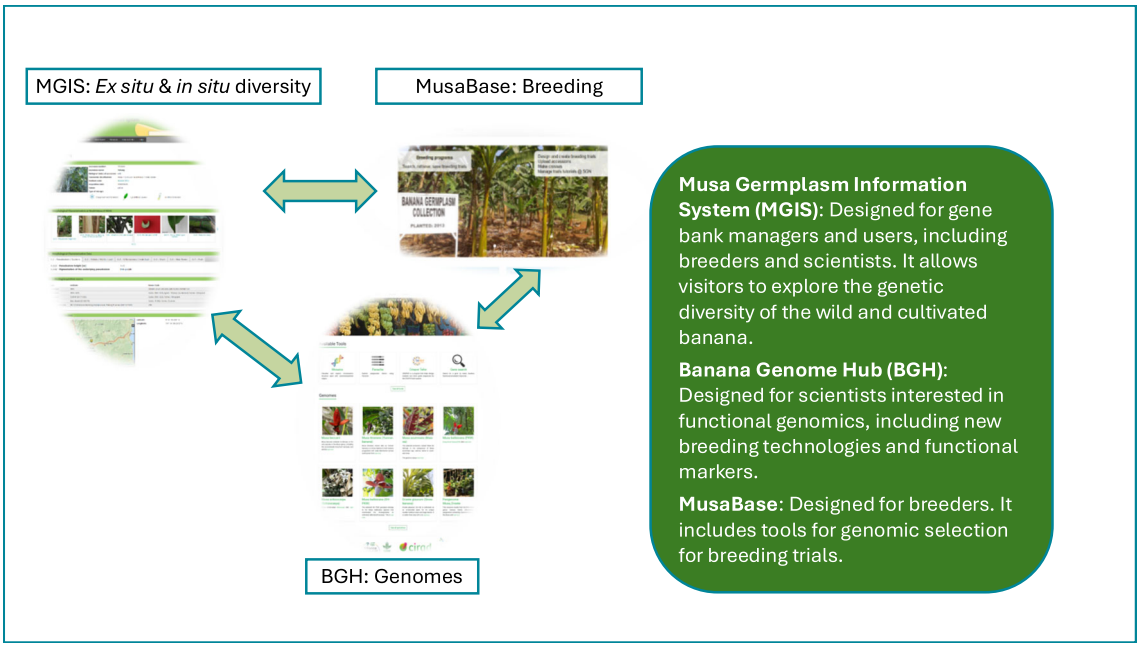

FIGURE 6: Different databases for different uses and users – an example on banana databases.
Source: authors' own elaboration.

Open and easy access to DSI is critically important. As previously explained, researchers rely heavily on knowledge and data that have been generated by other researchers, including from genetic resources belonging to other species, genera and kingdoms. If the data remain private, difficult to access or are only accessible in a raw, unprocessed form, they lose their value for crop breeding and PGRFA conservation. Statistics on the use of public databases are evidence of their usefulness to researchers around the world. Some examples: more than 5 000 different users from more than 180 countries use MaizeGDB per month. In the period July 2023–July 2024, 11 200 users from 106 countries consulted the Rice SNP Seek portal, which provides access to genotype, phenotype and variety information for rice, with China (4 555 users), India, the Philippines, the United States of America and Germany heading the list in that order. In the case of the Banana Genome Hub, Viet Nam appears as the country with the largest number of visitors in the last 12 months (5 000 users), followed by the Philippines (1 700), China (1 500), Brazil and France.

Although international efforts and guidelines to promote open access and enhance the FAIR (Findable, Accessible, Interoperable, Reusable) aspects of these resources are progressing, fragmentation still poses significant challenges for users navigating the complex web of information. This complexity also affects policymakers aiming to regulate such an ecosystem.

Minimum standards and essential principles for the curation and sharing of DSI and related metadata in databases would be highly useful. Finding funding mechanisms to ensure their sustainability is also crucial, given their importance in the sharing of non-monetary benefits for the conservation and use of PGRFA.

SUMMARY OF RELEVANT POINTS

The generation, analysis and use of DSI are inextricably linked to plant genetic material. The more PGRFA samples are studied to generate DSI, the more accurate, reliable and valuable the data become.

Digital sequence information is not a final product, but a research and breeding tool, and is effective in combination with related data such as phenotypic characterization and morphological data.

Digital sequence information is increasingly used to inform conservation, research and development of PGRFA by researchers and breeders in both the public and the private sector. It allows breeders to reduce the time and costs involved in the development of new crop varieties. In the crop improvement sector, this is where the economic value of DSI lies.

Public research organizations and private companies join forces at the early stages of DSI generation, in particular for the sequencing of reference genomes. Once generated, this information becomes relevant for subsequent research and activities, often by other research teams in other parts of the world.

Public organizations play a major role in the generation of DSI of PGRFA held in national and international gene banks, sometimes working in partnership with seed companies.

The bulk of data available in public databases comes from public organizations. The data are scattered across different databases, including crop-specific ones. The amount, value and use of DSI that is available in public databases varies greatly from crop to crop; major cereals and other major crops have the largest number of associated records and datasets, while for some minor crops DSI is virtually inexistent. Researchers and breeders look for different types of genetic information in different databases, depending on their research objectives.

Private companies utilize DSI that they obtain from partnerships and open databases to develop their own genomic tools for breeding. Rarely, information from public databases can be applied as such in the companies' breeding populations. Companies need to invest in genotyping, phenotyping, generating in-house DSI, and DNA marker generation and validation in support of their breeding programmes.

Large private companies almost exclusively use DSI of the PGRFA from their own collections. Prompted by the need for new traits, they occasionally seek additional PGRFA from external sources, for genotypic and phenotypic characterization in house. They maintain their private DSI databases.

POLICY IMPLICATIONS FOR THE DESIGN OF BENEFIT-SHARING NORMS FOR PGRFA AND DSI

Different actors working across the globe in the continuum of PGRFA conservation and use generate resources – including DSI – that contribute to the commercial value of new plant varieties. The design of monetary benefit-sharing obligations must take this fact into consideration.

Benefit-sharing on DSI should recognize the value creation enabled through *all* the PGRFA that were used to create the eventual knowledge and tools that are used at the breeding stage.

The International Treaty has the advantage of having the Multilateral System of Access and Benefit-sharing (Multilateral System) in place, catering for both monetary and non-monetary benefits. Pools of PGRFA and related information (including DSI) and benefit-sharing for these pools can be integrated into the Multilateral System, and this is currently being considered in the context of deliberations on the enhancement of its functioning.

Multilateral benefit-sharing schemes that require tracking and tracing of PGRFA and DSI being used in particular products (such as the current default benefit-sharing approach under the Multilateral System) do not fully reflect the contribution of all the upstream PGRFA and DSI. Given the way that DSI has changed the utilization of genetic resources and information in crop improvement, a new approach to benefit-sharing is justified: an approach that does not imply the tracking of PGRFA material and information along the research and development process of particular products, but that is based on payments linked to the sales of specified classes of products (namely, seeds and other relevant propagating materials), regardless of whether or not specific samples of PGRFA and derived DSI have been used directly for the development of such products.

The multilateral reality regarding DSI shows that there is already considerable *non-monetary benefit-sharing*, albeit largely unmeasured, through extensive public sharing and use of knowledge and data.

Authorship and contributions

This policy brief was written by Isabel López Noriega, Mathieu Rouard, Michael Halewood, Ruaraidh Sackville-Hamilton and Claudio Chiarolla, from the Alliance of Bioversity International and the International Center for Tropical Agriculture (CIAT). Álvaro Toledo, Daniele Manzella and Tobias Kiene from the Secretariat of the International Treaty on Plant Genetic Resources for Food and Agriculture, provided general guidance to the process. Hedwig de Coo, from the Secretariat of the International Treaty, coordinated the design of the infographics, which were developed by Dave Gray. Inputs were gathered from representatives from various private seed companies thanks to the support from Szonja Csörgő, from the International Seed Federation (ISF), and Jasmina Muminović Rilak from Bayer Crop Science and the ISF Coordination Group on Genetic Resources.