E

联合国粮食及农业组织 | Food and Agriculture Organization of the United Nations | Organisation des Nations Unies pour l'alimentation et l'agriculture | Продовольственная и сельскохозяйственная организация Объединенных Наций | Organización de las Naciones Unidas para la Alimentación y la Agricultura | منظمة الأغذية والزراعة للأمم المتحدة

COMMISSION ON GENETIC RESOURCES FOR FOOD AND AGRICULTURE

Item 7 of the Provisional Agenda

INTERGOVERNMENTAL TECHNICAL WORKING GROUP ON AQUATIC GENETIC RESOURCES FOR FOOD AND AGRICULTURE

Fourth Session

Rome, 21–23 February 2023

THE ROLE OF DIGITAL SEQUENCE INFORMATION IN THE CONSERVATION AND SUSTAINABLE USE OF GENETIC RESOURCES FOR FOOD AND AGRICULTURE: OPPORTUNITIES AND CHALLENGES

1. The Commission on Genetic Resources for Food and Agriculture (Commission), at its Eighteenth Regular Session, took note of actual and potential applications of digital sequence information (DSI) to the conservation and sustainable use of genetic resources for food and agriculture (GRFA). It stressed the innovation opportunities DSI offers for research and development related to GRFA as well as the challenges many countries face in developing the technical, institutional and human capacity necessary to use DSI for research and development.[1]

2. The Commission requested the Secretary to prepare a document reflecting key practices and experiences on how DSI is generated, stored, accessed and used for research and development related to GRFA, for review by the Working Groups at their next sessions.

3. The document *Digital sequence information and genetic resources for food and agriculture* (CGRFA/WG-AqGR-4/23/7) provides information on the generation, storage, access to and use of DSI for research and development related to GRFA. Further information is provided in the draft study contained in this document-

[1] CGRFA-18/21/Report, paragraph 32.

The Role of Digital Sequence Information in the Conservation and Sustainable Use of Genetic Resources for Food and Agriculture: Opportunities and Challenges

Draft study

David Smith, Matthew J. Ryan and Alan G. Buddie

Centre for Agricultural Bioscience International, CABI

This draft study has been prepared at the request of the Secretariat of the FAO Commission on Genetic Resources for Food and Agriculture with a view to facilitate consideration by the Commission of the role of digital sequence information for the sustainable use and conservation of genetic resources for food and agriculture. The content of this draft study is entirely the responsibility of the authors and does not necessarily represent the views of FAO or its Members.

Abstract

This study discusses applications of digital sequence information (DSI) that are relevant to genetic resources for food and agriculture (GRFA), including DSI that is not derived from GRFA but nevertheless contributes to their identification, characterization, use, improvement and conservation. Applications of DSI are also fundamental to the characterization of other components of biodiversity for food and agriculture (BFA) and are important tools in efforts to make agriculture more sustainable.

Searches of CABI's literature database, CAB Abstracts, which contains 10.9 million records, revealed many examples of publications that demonstrate the important contribution of DSI to improvement of crop production, control of emerging diseases and adaptation to climate change. The database searches revealed a rise in the number of publications on DSI from 20 000 in 2002 to 1 180 915 in 2022 (almost 12 percent of the records).

Scientific literature focusing on climate change adaptation and on improving the yields of the major global crops wheat, rice, maize, soybean, potato and chickpea was explored. Examples found included publications that addressed the following topics: discovery of candidate genes for improved abiotic stress tolerance in wheat; the contribution of DSI to progress on drought and heat tolerance in rice; use of DSI-based technologies to increase grain yield and starch content in maize; and DSI-assisted development of disease resistance and drought and salt tolerance in chickpea. These are clear examples of DSI playing an increasingly important role in research on climate change adaptation, crop production and plant health.

The increasing significance of DSI is further confirmed by the fact that the quantity of DSI available in public databases is growing exponentially: the content of the International Nucleotide Sequence Database Consortium (INSDC) database exceeded 9 petabytes in 2020. Analysis of the Science-based Approaches for Digital Sequence Information (WiLDSI) Data Portal demonstrates that data on biodiversity are generated globally and are being used extensively to help characterize it and to create innovative solutions to growing problems and threats.

However, making DSI available through public databases does not mean that it is accessible to everyone in the same way. Many countries face serious obstacles both in terms of access to DSI and in terms of its use. CABI received feedback from several of its member countries via its regional centres. The Bahamas, Brazil, China, Ghana, India, Kenya, Malaysia, Pakistan, Trinidad and Tobago, the United Kingdom and Zambia confirmed that they were using DSI, but most indicated that the capacities needed to generate it and make optimum use of it were not in place. However, the feedback from China confirmed that the country was in a good position with respect to the generation, storage, management and use of DSI.

There are currently several options on the table for how access to and use of DSI can be guaranteed while at the same time equitably sharing benefits associated with this use, especially with countries in need of capacity building and support in the field of GRFA conservation. There is some convergence towards a global, multilateral solution, while some countries are anticipating hybrid approaches that will incorporate both bilateral and multilateral systems for benefit-sharing. However, there is currently insufficient information available to carry out a cost–benefit analysis on these options, and discussions will continue on them during COP 15 in December 2022 (upcoming at the time of writing). The key messages of this paper are:

1. There are many different existing and potential applications of DSI that are highly relevant to GRFA, including applications of DSI that is not itself derived from GRFA.
2. The current and potential applications of DSI show that its generation, storage, accessibility and use are fundamental to the characterization of BFA and are important to efforts to make agriculture more sustainable.
3. Access to and use of DSI face serious obstacles in many countries. There is an urgent need to address the root causes of these problems, which include lack of technical infrastructure, financial and human resources, educational and training opportunities, scientific

collaboration, computing infrastructure, reliable electricity and high-speed internet, and may in the future possibly include prohibitive charges for database use.

4. There is a need for a regulatory environment that facilitates access to DSI and the fair and equitable sharing of benefits arising from its use.

Table of Contents

1. Introduction

This study contributes to the work stream on digital sequence information (DSI) of FAO's Commission on Genetic Resources for Food and Agriculture (the Commission). It presents key practices and experiences related to the ways in which DSI is generated, stored, accessed and used for research and development (R&D) related to genetic resources for food and agriculture (GRFA). It also explores the availability and accessibility of DSI to the research community and the private sector in all parts of the world, and presents solutions currently being discussed for access to and use of DSI and the sharing of benefits arising from its use. The term "DSI" was first used by the Convention on Biological Diversity (CBD)[2] in discussions of the issue of data on genetic resources being accessible without precipitating the benefit-sharing measures anticipated by the CBD and the Nagoya Protocol for the genetic resources themselves.[3]

The debate began at the Conference of the Parties (COP) to the CBD and spread to other major fora such as the United Nations Convention on the Law of the Sea (UNCLOS), the World Intellectual Property Organization (WIPO), the World Health Organization (WHO) and the Commission, each of which has explored how DSI affects its fields of responsibility. From these discussions, it is clear that the use of digital information relating to genetic resources delivers benefits: the current focus is on how these benefits should be shared. There has been some convergence on what DSI could cover, but the route to a solution for equitable benefit-sharing remains unclear. Given the importance of DSI to research and development, the ongoing debate is of considerable significance to the food and agriculture sector.

One argument is that the conservation of biological resources under the CBD could be supported without controlling access to DSI and sharing monetary benefits from its use. The alternative view is that, in view of the direct link between DSI and genetic resources more broadly, there is an obligation to share the benefits arising from the use of DSI.

Genomics is revealing previously undiscovered biodiversity that has an important role in ecosystems and is responsible for functions that are essential to biological cycles. It also provides a deeper understanding of how organisms function and thereby enables constructive manipulation and utilization of genetic resources. These developments provide opportunities for important advances in food and agriculture. The present study provides examples of the impact of DSI and demonstrates that developments and innovations in genomics are not bound to particular species or sectors. Work on species considered to be outside the remit of GRFA may have relevance to the field (and vice versa). The study presents examples of how the generation and use of DSI-nucleotide sequence data (NSD) enables advances in agriculture, food production and food security.

1.1 The term Digital Sequence Information (DSI)

The term DSI was originally developed in the context of the CBD and the Nagoya Protocol, although with the caveat that it "may not be the most appropriate term and … is used as a placeholder until an alternative term is agreed." (CBD Decision 14/20). Although still not clearly defined, DSI in its narrowest sense refers to genomic data (digitally recorded DNA and RNA sequences). However, in many cases the term is also used to refer to data on proteomics (protein sequences) and sometimes also to data on metabolomics (relating to primary and secondary metabolites, and other chemical entities). So-called "omics"-based techniques provide genomic blueprints of microorganisms, allowing their function, and their roles in water, carbon, nitrogen, phosphorus and sulphur cycles to be elucidated (Zhou *et al.*, 2022).

There is a pressing need for an agreed definition of DSI that can encompass potential future discoveries and new technologies, but this is proving difficult to achieve. It has been suggested that the term could be taken to encompass "the kind of information in, or that might be added to, databases

[2] https://www.cbd.int/dsi-gr/

[3] Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization.

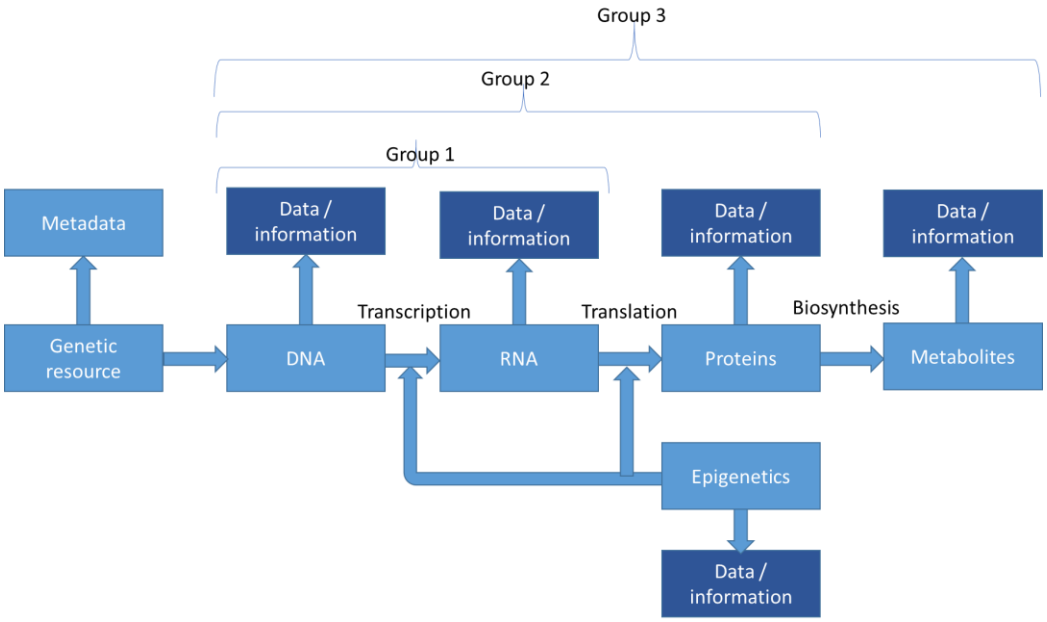
of the kind currently in use and collated by the scientific journal *Nucleic Acids Research*" (Heinemann, Coray and Thaler, 2018). The authors that made this suggestion cited the 2017 Database Issue of *Nucleic Acids Research* (Nucleic Acids Research, 2017), which documented 54 new databases added since the previous review. Subsequent reassessments have been made annually, with the latest in 2022 (Rigden and Fernández, 2022). This definition is associated with, but goes beyond, DNA sequences: it also encompasses proteomics and metabolomics, as these are included in the *Nucleic Acids Research* database lists.

The CBD Ad Hoc Technical Expert Group (AHTEG) on DSI and the Open-ended Working Group on the Post-2020 Global Biodiversity Framework (OEWG) did not attempt to define DSI: their approach was to compartmentalize the scope of DSI into three subgroups of information (Table 1) (AHTEG, 2020; POEWG, 2021). Group 1 includes DNA and RNA. Group 2 includes Group 1 and adds proteins and epigenetic modifications. Group 3 includes Groups 1 and 2 and adds metabolites and other macromolecules. However, it was not agreed whether Groups 2 and/or 3 should be considered DSI.

Data/information flows linking GR and related NSD generated by research are summarized in Figure 1. According to Lyal, 2022, "the main basis for accepting DSI as coming under the CBD and Nagoya Protocol is the (disputed) proposition that DSI is the 'intangible equivalent' of a physical genetic resource and as such falls under the sovereign rights of the country from which the original genetic resource was accessed." Lyal describes DNA or RNA sequences (NSD) as "the closest functional analogy between a genetic resource and an intangible equivalent" and notes that "a number of countries have apparently adopted this concept. 'GSD' (genetic sequence data) is used in the World Health Organisation pandemic influenza preparedness framework and has the same meaning. This is the Group 1 of the latest AHTEG" (Lyal, 2022).

Ruiz Muller, 2018 introduced the term "natural information" to the debate and defined it as follows: "any non-uniformity, difference, or distinction not intentionally produced by *H. sapiens* which derives from thermodynamically open systems to dissipate energy gradients and create copies of itself", also putting forward the concept of "bounded openness for natural information", which includes sequence data and all "natural information" (Ruiz Muller, 2018) This would include the "associated information" mentioned in Table 1. Vogel *et al.* (2022) note that a "more colloquial and maybe legal definition could also be 'any non-uniform expression, difference or distinction produced by nature." They further note "conceptualization of alternative terms to replace DSI" and conclude that "natural information (biotic) captures what should fall within the scope of the CBD while excluding information that is artificial or natural but abiotic." Vogel *et al*. (2022) include "*in silico* utilization" (ISU) of GR, genetic information (GI), GSD and NSD of the biotic natural information within the natural information category. They believe that "once artificial or natural information is interpreted as the object of access in R&D, a multilateral system can be constructed in a way that all the international agreements that concern ABS become harmonious. The optimal modality is bounded openness" (Vogel *et al.,* 2022). They define "bounded openness", in turn, as "legal enclosures which default to, yet depart, from *res nullius* [property of no one] to the extent the departures enhance efficiency and equity, which must be balanced when in conflict." They go on to say that it "satisfies … three criteria: genetic resources flow freely for R&D …; royalties are due only on the value added through intellectual property and distributed proportional to custodianship …; and transaction costs are minimized …."(Vogel *et al.,* 2022).

Figure 1. The concept and scope of DSI and its generation in the metabolic pathway



Source: Lyal, C.H.C. 2022. Digital sequence information on genetic resources and the convention on biological diversity. In: E. Chege Kamau, ed. *Global transformations in the use of biodiversity for research and development*. Ius Gentium: Comparative Perspectives on Law and Justice, 95, pp. 589–619. Cham, Switzerland, Springer. after AHTEG 2020. *Report of the Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources, Montreal, Canada, 17–20 March 2020*. CBD/DSI/AHTEG/2020/1/7. Montreal, Canada. https://www.cbd.int/doc/c/ba60/7272/3260b5e396821d42bc21035a/dsi-ahteg-2020-01-07-en.pdf

The various concepts described here would result in different outcomes if used to define the scope of access and benefit-sharing (ABS) regimes and would result in different levels of complexity in traceability and monitoring. The ultimate outcome depends on how far along the metabolic pathway (the flow of information) the scope extends and is justifiable; human (research) intervention is required at several stages (see Figure 1). Lyal (2022) discusses the elements along the metabolic pathway that "reflect the degree of biological processing and the proximity to the underlying genetic resource." Human interventions include those related to the further analysis of the raw nucleotide sequence, the technical aspects of sequencing, the "associated data" to which the AHTEG referred and metadata from the collection of the genetic resource. They also include those related to aligned nucleotide sequences, information on sequence assembly, structural annotation of genomic elements, biochemical and biological function, behavioural observations, the structure of organisms, the molecular structures of gene products and derivatives (cell metabolites, etc.), and patentable discoveries and inventions.

Table 1. Grouping and scope of DSI

	Information related to a genetic resource			
	Genetic and biochemical information			Associated information
Group reference	Group 1	Group 2	Group 3	
High-level description of each group	DNA and RNA	Group 1 + proteins + epigenetic modifications	Group 2 + metabolites and other macromolecules	

Examples of granular subject matter	• Nucleic acid sequence reads • Associated data to nucleic acid reads • Non-coding nucleic acid sequences • Genetic mapping (e.g. genotyping, microsatellite analysis, single nucleotide polymorphisms [SNPs], etc.) • Structural annotation	• Amino acid sequences • Information on gene expression • Functional annotation • Epigenetic modifications (e.g. methylation patterns and acetylation) • Molecular structures of proteins • Molecular interaction networks.	• Information on the biochemical composition of a genetic resource • Macromolecules (other than DNA, RNA and proteins) • Cellular metabolites (molecular structures)	• Traditional knowledge associated with genetic resources • Information associated with DSI Groups 1, 2 and 3 (e.g. biotic and abiotic factors in the environment or associated with the organism) • Other types of information associated with a genetic resource or its utilization

Source: AHTEG. 2020. *Report of the Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources, Montreal, Canada, 17–20 March 2020*. CBD/DSI/AHTEG/2020/1/7. Montreal, Canada. https://www.cbd.int/doc/c/ba60/7272/3260b5e396821d42bc21035a/dsi-ahteg-2020-01-07-en.pdf

Table 2 shows the number of hits from the CAB Abstract database for DSI Groups 1 to 3 as defined by the AHTEG. If the terms chosen by the AHTEG to characterize these groups are used to search the database for DSI, the numbers of publications found is relatively low for Group 1 (4 965 hits). When the search is expanded to include Group 2, an additional 23 246 hits are obtained, and for Group 3 an additional 320 hits are obtained. This is well short of the 1 180 915 hits obtained in the CAB Abstracts database using the comprehensive set of search terms listed in Table 3, where the numbers were 10 to 100-fold larger (see Section 1.2).

Table 2. Numbers of records in the CAB Abstracts citing DSI (the data pool) for the three groups

	Information related to a genetic resource			
	Genetic and biochemical information			Associated information
Group reference	Group 1	Group 2	Group 3	
High-level description of each group	DNA and RNA	Group 1 + proteins + epigenetic modifications	Group 2 + metabolites and other macromolecules	
CABI DSI data pool hits	4 965	23 246 (additional) i.e. 28 211 (Groups 1 and 2) in total	320 (additional) i.e. 28 531 (all groups) in total	All 1.8 million records in the DSI data pool

These figures reflect the research covered in the CAB Abstracts database, which in turn reflects current research in agriculture. The database contains fewer publications specifically on RNA and DNA sequences, but when searches include their direct products, proteins and epigenetic modifications (i.e. Group 2), there is an almost five-fold increase in hits. The creation of the DSI data pool required the use of all relevant terms for Groups 1 to 3. The high-level description of each group, as given in Table 1, fails to find all publications citing DSI. This reflects the complexity of the subject and potentially raises issues around our ability to monitor and trace DSI usage.

Table 3 presents the list of terms put forward in the AHTEG report related to activities and processes in the generation of DSI. They represent elements of the genetic information on an organism that may be utilized to generate products. It is important to note that the coding of sequences could potentially produce functional information without needing to go to the protein level, specifically coding nucleic acid sequences and information on gene expression at the transcript level (e.g. messenger RNA and complementary DNA) can result in products for exploitation. When the terms presented in Table 3 were used to create the CABI DSI data pool, significantly higher hit rates were obtained, with a total of 1.18 million records found.

Table 3. The AHTEG options for terminology to describe DSI on genetic resources

Group reference	Group 1	Group 2	Group 3	Associated information
Category/term	Nucleotide sequence data (NSD) Genomic sequence information Genomics information Nucleotide sequence information (NSI) Genetic Resource Sequence Data (GRSD) Digital sequence data (DSD) Data on the genomic DNA (or RNA) of a sample genetic resource	Genomic and proteomic sequence information Nucleotide sequence information (NSI) Genetic information (GI) Sequence data Nucleotide and amino acid sequence data (NASD) Nucleotide and amino acid sequence and structural information (NASSI) Nucleotide and amino acid sequence, structural and functional information (NASSFI) Functional digital information of NSD Proteomic data Genomic and proteomic sequence information Data on the macromolecular composition of a sample genetic resource	Genomic, proteomic and metabolomic information Genetic and "omics" information Metabolomic data "Omics" information Genomic, proteomic and metabolomic information Data on the biochemical and genetic composition of a sample genetic resource.	Associated information Contextual Information Subsidiary Information.

Source: AHTEG 2020. *Report of the Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources, Montreal, Canada, 17-20 March 2020*. CBD/DSI/AHTEG/2020/1/7. Montreal, Canada. https://www.cbd.int/doc/c/ba60/7272/3260b5e396821d42bc21035a/dsi-ahteg-2020-01-07-en.pdf

Aside from genetic/biochemical factors, there is a need to address the data side: published research results (see Figure 1). The process of data generation, storage and management is complex, data may be generated – or even computationally derived from other data. For example, sequence databases contain the sequence of "bases" (components of DNA), with associated metadata describing the source of the sequence and related information. It is not just "data": in each case there has been some processing, and thus more "information" is included, i.e. the process includes a human cognitive element that adds to its ultimate value. This is particularly relevant to the development of the end product, and ultimately the generation of benefits, in that the process is one-way: it is possible to deduce a protein's composition from the gene but not the gene from the protein (or from a given metabolite). This is because more than one gene can code for the same protein (and more than one codon may code for the same amino acid).

Table 4 provides further details of the methodologies used to create DSI on biodiversity for food and agriculture (BFA), briefly describing the uses of the technologies, the type of data produced and the

number of hits obtained in the CAB Abstract database. They are presented under the AHTEG groups, and the number of papers in the CAB Abstracts database citing these technologies gives an indication of the extent of their use. There were 86 655 hits for genome sequencing number (Group 1), 81 528 hits for proteins and epigenetic modifications (Group 2 additions) and 6 208 hits for Group 3 additions. Thus Groups 1 and 2 had in the greatest number of hits.

Table 4. Methods for analysing DSI in food and agriculture

Method	Description	Uses	Data type	CABI Search hits
Genome Sequencing (AHTEG Group 1 DSI)				
De novo sequencing	A step towards understanding the genetic component of an plant or animal's traits and interactions with the environment	Assigning map positions and stacking breed information for subsequent resequencing to discover single nucleotide polymorphisms (SNPs) and other genetic variations	DNA sequence data free of any constraints or assumptions	9 877
Whole-genome resequencing	A comprehensive method for analysing the entire genome when a species' reference genome is available	Identifying genes, SNPs and structural variants while simultaneously determining genotypes	Sequence differences from the reference genome	8 469
Transcriptome sequencing	A method that provides novel insights into changing gene expression levels that occur in development and during disease and under conditions of stress	Elucidating gene and protein function and interactions, identifying tissue-specific lists of RNA transcripts and discovering new SNPs	Messenger RNA expressed in the tissue sample at the time of extraction	32 945
Epigenetics	Adaptive responses to changes in the environment (such as food availability or drought conditions) can trigger phenotypic changes in plants and animals that affect their viability and reproductive fitness	Using sequencing to identify changes in DNA methylation, chromatin structure and small RNA expression to better understand how epigenetic factors contribute to controlling these and other traits in a species of interest	Generally, changes in cytosine methylation at position 5 when sequencing methods are used	10 992
Targeted resequencing	A method for sequencing predetermined areas of genetic variation over many samples	Identifying common and rare variants – such as SNPs and copy number variants (CNVs) – to help inform breeding decisions or characterize disease susceptibility	Sequence variants compared to a reference sequence at each locus	2 558
Genotyping by Sequencing (GBS)	A low-cost genetic screening method for discovering novel plant and animal SNPs and performing genotyping studies, often simultaneously in many specimens	Include genetic mapping, screening backcross lines, purity testing, constructing haplotype maps, and performing association and genomic evaluation for plant genome studies	SNPs in genotyping studies, such as genomic wide association studies (GWAS); GBS uses restriction enzymes to reduce genome complexity and genotype multiple DNA samples	16 213

Method	Description	Uses	Data type	CABI Search hits
Environmental DNA sequencing	An effective biomonitoring tool that allows characterization of both bacterial and eukaryotic species in aquatic, soil and other samples	Include port monitoring, biodiversity surveys, ballast water testing and soil testing	eDNA from environmental samples, rather than directly sampled from an individual organism	483
Genome editing	CRISPR (clustered regularly interspaced short palindromic repeats) genome editing holds great potential for agriculture, food science, environmental science and a broad range of other applications	Confirming gene knockouts, analysing on- and off-target effects and assessing the functional impact of gene edits	Edited DNA from a cell expressing specific properties	5 118
Proteins and epigenetic modifications (AHTEG Group 2 additions)				
Mass spectrometry	Determines mass-to-charge ratio of ions (e.g. in gas or liquid chromatography)	Identifying and characterizing small molecules and proteins (proteomics)	Data that allow protein identification and annotation of secondary modifications, and determination of the abundance of individual proteins	13 887
ELISA	Enzyme-linked immunosorbent assay	Detection of antigen	Antigen presence, and possibly quantitation	24 332
Gel electrophoresis	A method in which an electric current is applied to samples, creating fragments that can be used for comparison between samples	Separation of DNA, RNA and protein samples	Separate bands of DNA, RNA or protein molecules based on their size and electrical charge	23 546
Chromatography	A method in which a mixture is dissolved in a fluid solvent (gas or liquid) that is carried through a system; components move at different velocities and are thus separated	Separation of a mixture into its components	Fractionation position and abundance	19 665
Protein microarrays	Application of small amounts of sample to a “chip” for analysis	Detection of protein–protein interactions	Data on the presence of antigens	98
Metabolomic information (AHTEG Group 3 additions)				
MALDI-ToF-MS	Matrix-assisted laser desorption/ionization coupled to time-of-flight mass spectrometry	Bacterial identification	Molecular weight and abundance profiles for acid-soluble proteins, with identification from comparison of unknowns with a database of knowns	2 813
Nuclear magnetic resonance (NMR) spectroscopy	A method for determining the structures of complex molecules by measuring the interaction of nuclear spins when placed in a powerful magnetic field	Determining the molecular structure at the atomic level of a sample; Can be used to generate metabolic fingerprints from biological fluids	Data on the structure, dynamics, reaction state, and chemical environment of molecules	3 130

Method	Description	Uses	Data type	CABI Search hits
High-resolution mass spectrometry (HRMS)	Modern separation techniques such as liquid chromatography, gas chromatography or capillary electrophoresis, are often coupled with HRMS	Fractionation of molecules	Fractionation position and abundance	265

1.2 Resources used in this study (material and methods)

For tracking the use of DSI, the outputs of the Wissenschaftsbasierte Lösungsansätze für Digitale Sequenzinformation (Science-based Approaches for Digital Sequence Information) (WiLDSI) Data Portal[4] were consulted. This portal enables discovery of data from 198 countries for biogeographical studies, exploration of collaborative networks, and profiling of access and use of sequence data (DNA and RNA grouped as NSD). An analysis by Lange *et al.* (2021) linked NSD to scientific publication citations to enhance understanding of NSD provenance, scientific use and reuse. The analysis connected publications with NSD records, geographical information and author contribution based on their country of origin to infer trends in scientific knowledge gain at the global level. These data were further analysed to determine whether there was an imbalance in generation and use of the data from a sample of the 198 countries.

The present study used the CAB databases to analyse growth in the generation and the application of DSI. CABI has been gathering data on agriculture for over 100 years, and much of this is presented via CAB Direct.[5] The CAB Abstracts bibliographic database is part of this resource and covers applied life sciences, including agriculture, plant sciences, animal sciences and related subjects. It contains over 10.9 million records dating from 1973 to the present (and an archive covering the period from 1912 to 1973). It is searchable on several platforms including CAB Direct, which was used in this study. The search strategy was to filter out the DSI-related studies from CAB Abstracts and to group the "hits" obtained into various categories such as "dominant crops in FAO regions", various terminologies for describing DSI, and examples of actual and potential applications of DSI in food and agriculture. The terms used to identify the DSI-related records in CAB Abstracts were in accordance with the three DSI groups identified by the AHTEG. The analysis is described in more detail in Appendix 1.

Search terms	Database
"digital sequenc*" or "genetic engineer*" or "genetic sequenc*" or "dna sequenc*" or "nucleotide sequenc*" or "RNA sequenc*" or "genomic*" or "gene expression" or "recombinant dna" or "ngs" or "mRNA" or "transcription" or "genetically engineer*" or metagenomic or "next generation sequenc*" or "genome*" or "genetic manipulation" or "molecular genetic*" or "polymerase chain*"	CAB Abstracts

Comparative searches were carried out in PubMed for comparison:

Search terms	Database
"digital sequenc*" or "genetic engineer*" or "genetic sequenc*" or "dna sequenc*" or "nucleotide sequenc*" or "RNA sequenc*" or "genomic*" or "gene expression" or "recombinant dna" or "ngs" or "mRNA" or "transcription" or "genetically engineer*" or	PubMed

[4] https://apex.ipk-gatersleben.de/apex/wildsi/r/wildsi/home
[5] https://www.cabdirect.org/

metagenomic or "next generation sequenc*" or "genome*" or "genetic manipulation" or "molecular genetic*" or "polymerase chain*"	

A direct comparison with Google Scholar was not possible, because its 256-character limit for searches precluded the inclusion of all the search terms. A search for DSI on Google Scholar resulted in only 876 hits.

Search results and parameters used to filter papers citing microorganism and invertebrate DSI in CAB Abstracts:

- Microorganisms = 375 509 records [using ("digital sequenc*" or "genetic engineer*" or "genetic sequenc*" or "dna sequenc*" or "nucleotide sequenc*" or "RNA sequenc*" or "genomic*" or "gene expression" or "recombinant dna" or "ngs" or “mRNA” or "transcription" or "genetically engineer*" or metagenomic or "next generation sequenc*" or "genome*" or "genetic manipulation" or "molecular genetic*" or "polymerase chain*" AND ("bacteri*" OR "fung*" OR "microbe*" OR "microorganism*" OR "micro-organism*" OR "protist*")]
- Invertebrates = 165 440 records [using ("digital sequenc*" or "genetic engineer*" or "genetic sequenc*" or "dna sequenc*" or "nucleotide sequenc*" or "RNA sequenc*" or "genomic*" or "gene expression" or "recombinant dna" or "ngs" or “mRNA” or "transcription" or "genetically engineer*" or metagenomic or "next generation sequenc*" or "genome*" or "genetic manipulation" or "molecular genetic*" or "polymerase chain*" AND ("invertebrate*" OR "arthropod*" OR "insect*" OR "mollusc*" OR "annelid*" OR "tardigrad*" OR "echinoderm*" OR "bryozo*")[6]

To assess the availability/accessibility of DSI to the research community and the private sector in all parts of the world, key managers and researchers in CABI’s regional and national centres carried out local enquiries with contacts, national authorities and project partners to determine the extent to which they generated, accessed and utilized DSI/NSD (see Appendix 2).

2. Relevance of DSI in general

DSI underpins much current research in the life sciences, contributing to advances in medicine, conservation, agriculture and other fields. Since the first complete bacterial genome was sequenced in 1995 (Fleischmann *et al.*, 1995), over 200 000 bacterial and archaeal complete or draft genomes have been uploaded to public databases, and this has been happening at an increasing rate thanks to advances in sequencing technology and associated decreases in “per base” costs (Land *et al.*, 2015). The rapid increase in the rate of sequence data acquisition since the end of the 1990s has been driven by matched advances in the fields of high-throughput (massively parallel) nucleic acid sequencing and computing power for data analysis. This has enabled increases in per-machine output from about 5 kilobase pairs per day in the mid-1980s (early automated Sanger sequencing machine) to up to 180 gigabase pairs per day from a current Illumina NextSeq. The total amount of sequence data maintained by the International Nucleotide Sequence Database Consortium (INSDC), comprising the DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (ENA) and GenBank, exceeded 9 petabytes in 2020 (Arita, Karsch-Mizrachi and Cochrane, 2021).

The DSI that is now available is thus a product of recent unprecedented improvements in genetic sequencing technology. This vast amount of data can be utilized by scientists to obtain an understanding of an organism’s DNA, the genes it possesses and ultimately the function of these genes. The data can help identify changes in particular genes and, through comparison with other examples from the species, provide information on the role of inheritance in susceptibility to disease and response to environmental influences. This can provide vast potential for diagnostics and

[6] The terms were selected in consultation with the CABI Abstracts database managers with knowledge of the database content.

therapies (see the National Human Genome Institute Fact Sheet (NIH, 2020) for further information on the latter).

In non-medical applications, DSI is mainly generated for the purpose of identifying and characterizing biodiversity. When a DNA sequence (partial [barcode] or whole genome) is generated from a genetic resource, it is compared to existing data to determine its similarity to previously identified sequences. If a close match is found, the genetic resource can be identified. The information obtained can be utilized in multiple sectors for research and product development.

GRFA can be relevant for uses in other sectors, for example where sequence data from GRFA are utilized to characterize sequence data from other organisms; data can also lead to the discovery of enzymes and metabolites for use in industry and healthcare. Once a resource has been sequenced, it can be compared with existing sequences derived from other sectors. Such cross-over between sectors presents a potential challenge for ABS instruments and schemes that address the use of DSI for specific uses such as food and agriculture.

Scientists use DSI in various ways to inform their research and provide the baseline for solutions to challenges such as those involved in the identification of organisms or the selection of the most appropriate production and application strains. Analysis of specific regions of DNA can provide insight into genes that are essential to regulatory mechanisms and the switching "on" or "off" of their activity. Increasingly, such comparisons can tell us which genes cause or increase susceptibility to disease, taking the influence of both inheritance and the environment into account.

Zhou *et al*. (2022) considered the effect of current technologies such as metagenomics and single cell genomics on the reconstruction of genomes from mixed microbial communities. They concluded that such approaches allow scientists to "read genomic blueprints of microorganisms, decipher their functional capacities and activities, and reconstruct their roles in biogeochemical processes."

DSI can help to explain the molecular basis and evolutionary theory of life and provide new methods for the conservation and sustainable use of biodiversity (Li- and Xue, 2014). It is now possible to design and build products from DSI, removing the need to access the genetic resource itself. Biofoundries (see Box 1) can build such products using automation and high-throughput equipment, including process scale-up, computer-aided design software, and other workflows and tools.

Box 1 Biofoundries

The generation and characterization of DSI has allowed the establishment of biofoundries, highly automated facilities that enable products and discoveries to be obtained from DSI (Hillson *et al.*, 2019; Richardson, *et al.*, 2017 and Si *et al.*, 2017). The Global Biofoundry Alliance (GBA)[7] develops, promotes and collaborates on biological engineering to allow researchers to test large-scale genetic designs and apply artificial intelligence machine learning to enhance the design process. It provides an infrastructure that enables the rapid design, construction and testing of genetically reprogrammed organisms for biotechnology applications and research (Hillson *et al.*, 2019). The GBA website states that:

"the overall objective of GBA is to:
1. Develop, promote, and support non-commercial biofoundries established around the world,
2. Intensify collaboration and communication among biofoundries,
3. Collectively develop responses to technological, operational, and other types of common challenges,
4. Enhance visibility, impact and sustainability of non-commercial biofoundries, and
5. Explore globally relevant and societally impactful grand challenge collaborative projects."

It also demonstrates the huge potential of such infrastructures and the innovation and benefits that can arise from them, including the development of new products for the market.

[7] https://biofoundries.org/

Source: Hillson, N., Caddick, M., Cai, Y. Carrasco, J.A., Chang, M.W., Curach, N.C., Bell, D.J. et al. 2019. Building a global alliance of biofoundries. Nature Communications, 10: 2040. https://doi.org/10.1038/s41467-019-10079-2; Richardson, S.M., Mitchell, L.A., Stracquadanio, G., Yang, K., Dymond, J.S., Dicarlo, J.E., Lee, D. et al. 2017. Design of a synthetic yeast genome. Science, 355(53223): 1040–1044. https://www.science.org/doi/10.1126/science.aaf4557; Si, T., Chao, R., Min, Y., Wu, Y., Ren, W. & Zhao. H. 2017. Automated multiplex genome-scale engineering in yeast. Nature Communications, 8: 15187. https://doi.org/10.1038/ncomms15187Sayers *et al.* (2022a) reported that, as of 2021, GenBank®,[8] a comprehensive public database, contained over 15.3 trillion base pairs from over 2.5 billion nucleotide sequences from 504 000 formally described species. Daily data exchange with the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDBJ) ensures worldwide coverage (Benson *et al.*, 2011). The database has 21 main divisions (Sayers *et al.*, 2022a). A snapshot of the figures presented on the growth of the database (Table 5) shows huge increases in the number of base pairs (bp) of sequence added in the year from August 2020, with the largest increases being in invertebrate and virus sequences (the latter unsurprisingly in view of the worldwide pandemic during this period).

Table 5. A snapshot of the growth in the number of nucleotide base pairs in selected GenBank Divisions, 2020–2021

	GenBank Division[9]	**Number of bp of sequence in GenBank in August 2021**	**Growth in sequence databases in the year to August 2021**
	Plants	350 590 744 188	30.12%
	Phages	935 884 237	19.59%
	Viruses	39 351 597 469	575.68%
	Bacteria	130 518 385 589	32.07%
	Primates	15 165 437 356	72.97%
	Rodents	23 336 550 435	93.02%
	Other mammals	28 568 850 588	37.06%
	Other vertebrates	85 320 979 451	34.22%
	Invertebrates	108 680 334 593	450%
Total increase (08/2020-08/2021)	Above sets plus 12 other GenBank divisions	15 309 209 714 374	54.79%

Source: Sayers, E.W., Cavanaugh, M., Clark, K., Pruitt, K.D., Schoch, C.L., Sherry, S.T. & Karsch-Mizrachi, I., 2022a. GenBank. *Nucleic Acids Research*, 50: D161–D164. https://doi.org/10.1093/nar/gkab1135

Over the past 20 years, the sequences of over 1 000 plant genomes have been published, representing 788 highly diverse species (Sun *et al.*, 2022). However, this remains a small proportion of the more than 390 000 plant species known to science. Although the cell and taxon proportions of genome-sequenced bacteria or archaea on earth remain unknown, 155 810 prokaryotic genomes can be found in public databases (Zhang *et al.*, 2020). These studies reveal the current situation of prokaryotic genome sequencing for Earth biomes, where only 2.1 percent of prokaryotes are represented by sequenced genomes, and the large number of taxa with unknown genomes. Despite the increasing amount of data generated, understanding of this genome information remains limited. Continuing such work with further analysis will accelerate advances towards a comprehensive understanding of microbial ecological functions in different environments (Zhang *et al.*, 2020). Forin *et al.* (2018) note that recent studies have reported only approximately 35 000 correctly identified fungal species

[8] www.ncbi.nlm.nih.gov/genbank/
[9] https://www.ncbi.nlm.nih.gov/genbank/samplerecord/

represented by DNA sequences in public databases. This is far short of current estimates that there are 3.8–5.1 million extant fungal species (Blackwell, 2011; Hawksworth and Lücking, 2017) and up to 1 million species of prokaryotes (Louca *et al.*, 2019).

3. Generation and storage of DSI

Generating DSI is expensive, although the cost is falling as methods improve and become more efficient. The Earth BioGenome Project, which aims to sequence the DNA of all 1.5 million known eukaryotic species, is projected to cost USD 4.7 billion. The AfricaBP effort to sequence the genomes of 105 000 eukaryotic species in Africa will cost a total of USD 1 billion over ten years.

These estimates of sequencing costs are based on the plateauing cost per megabase of sequence. The National Human Genome Research Institute estimates that the cost of DNA sequencing at its sequencing centres[10] has come down to around USD 0.10 per megabase of raw sequence (though their subsequent cleaning up or annotation takes considerable additional investment in computing resources and trained bioinformaticists). Given that costs have plateaued, and the (likely continuing) increase in energy costs and reagent costs – especially single-use plastics – we consider that the cost per megabase is unlikely to drop below USD 0.075 in the near future.

If such a data resource were to be built elsewhere, or if it needed to be recreated, for example under a single standard, on the basis of a cost of USD 0.075/megabase, the 6.25 trillion bp of sequence data held by the European Bioinformatics Institute (EBI) would cost USD 468 750. However, if we take a more pragmatic view that we would need to (re)sequence the 1.6 billion individually deposited sequences in EBI at USD 0.075 per sequence, the cost would be USD 120 million. (Obviously, many sequences will be barcode sequences of < 1 kilobase pairs, but many of these will have been obtained by Sanger sequencing, which costs a lot more than USD 0.075 per megabase.)

The costs are not negligible. One way in which this issue can be addressed is to look to distributing the effort (e.g. with data-brokering arrangements) so that globally distributed expertise can contribute to the building of ENA/INSDC and the costs can be shared. The rewards to countries relate initially to growth of the science sector but later translate into societal benefits from new-found national influence within the global scientific effort. Building these kinds of distribution arrangements is an ongoing priority for ENA (and INSDC as a whole) and will address INSDC operational issues as well as the global imbalance of DSI generation, storage and analysis.

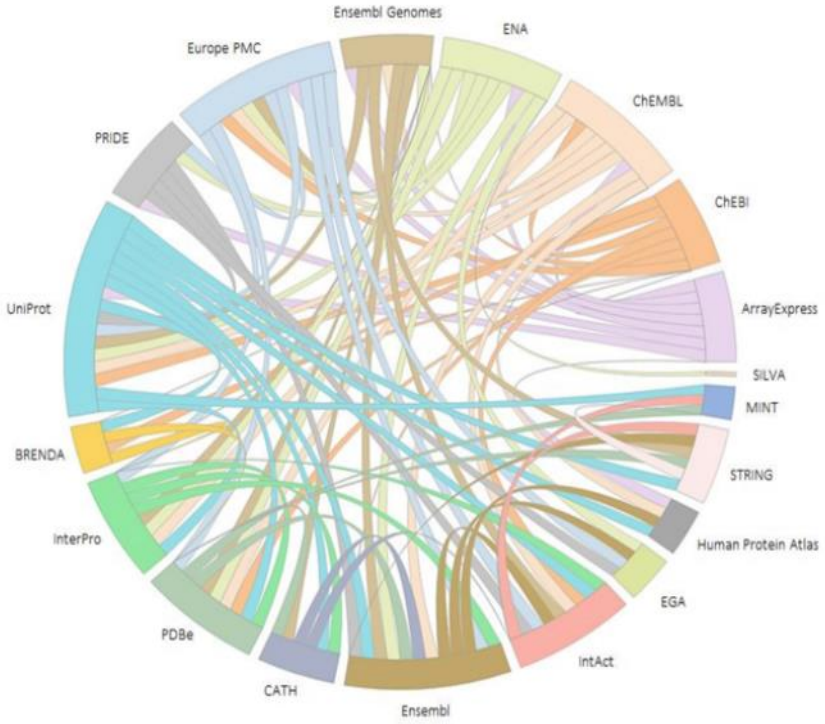
Data are held in many places, in databases managed by different organizations. The resources are often termed “archival data repositories”, with their data not just stored but actively curated and managed. Examples include the INSDC, Worldwide Protein Data Bank[11] (3D structure of macromolecules) and the ProteomeXchange[12] collaboration. Knowledge bases integrate information from multiple sources, often using computational approaches, for example the Universal Protein Resource[13] (protein sequences and function). Many, but not all, such repositories are public. INSDC content is transmitted to more than 1 700 other public databases with specialized content, processes and uses; many databases use and repurpose content from other databases. Figure 2 illustrates the data movements between some of these databases.

DSI generated by publicly funded research or published in scientific literature usually has to be made freely available; this is usually achieved through open-access public databases. A mechanism for making public databases more inclusive and ensuring balanced global coverage is needed. There is a broad range of opinions on this topic, and a full discussion of them goes beyond the scope of this report. One example is the suggestion that it may help to bring in key a person or persons from

[10] https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data
[11] http://www.wwpdb.org/
[12] http://www.proteomexchange.org/
[13] https://www.uniprot.org/

regional bioinformatics networks such as the Pan African Bioinformatics Network for the Human Heredity and Health in Africa (H3Africa[14]) consortium (H3BioNet) (Ebenezer *et al*., 2022).[15]

Figure 2. A representation of data flows between a small sample (19) of databases



Source: Lyal, C.H.C. 2022. Digital sequence information on genetic resources and the convention on biological diversity. In: E. Chege Kamau, eds. *Global transformations in the use of biodiversity for research and development*. Ius Gentium: Comparative Perspectives on Law and Justice, 95, pp. 589–619. Cham, Switzerland, Springer.

3.1 Where is DSI generated and used?

The vast majority of countries use and provide access to GR from which DSI has been generated, albeit to varying extents. Scholtz *et al*. (2021) reported in 2021 that the total output of publications from scientists in low- and middle-income countries (LMICs) is 40 percent less than that of their counterparts in high-income countries. However, scientists are using DSI in almost every country in the world – and thus DSI is truly being generated and used on a global scale (Scholtz *et al*., 2021).

The WiLDSI Data Portal[16] enables discovery of data for biogeographical studies, exploration of collaborative networks, and profiling of the flow of access and benefit relating to sequence data (DNA and RNA, grouped as NSD). To explore NSD provenance and scientific use and reuse in the community, Lange *et al*. (2021) in 2021 extracted NSD records from ENA and linked them to citations in open-access publications aggregated at Europe PubMed Central. By connecting publications with NSD records, NSD geographical provenance information and author geographical information, Lange *et al*. were able to assess the contribution of NSD to scientific knowledge, and to infer global trends. A total of 8 464 292 ENA accessions with geographical provenance information were found to be associated with publications, and the authors concluded that global provision and use of NSD enable scientists worldwide to join literature and sequence databases in a multidimensional fashion. Analysis of WiLDSI shows that most countries (to varying extents) use and provide DSI for basic and applied research in both the public and private sectors. For example, DSI

[14]https://h3africa.org/
[15]https://www.h3abionet.org/
[16] https://wildsi.ipk-gatersleben.de/apex/wildsi/r/wildsi/home

from Kenya is being used by 79 countries worldwide, while scientists in Kenya use DSI from 83 countries; DSI from Brazil is used by 111 countries, while scientists in Brazil use DSI from 153 countries.

The WiLDSI Data Portal currently presents data on 198 countries. The United States of America (7 726 083 sequences) and China (5 399 676 sequences) are the biggest providers of the genetic resources used to generate DSI. The next biggest providers are the United Kingdom (2 518 762) and Canada (2 158 026): these four countries are also among the greatest users of DSI. Only six countries produced more than 1 million sequences, the other two being Japan and Germany; 157 countries produced fewer than 100 000 sequences, and 27 countries each produced fewer than 2 000 sequences.[17] These data are based on the metadata/country tag used by the three INSDC databases. They refer to the **country of origin** of the genetic material that was sequenced and **not** to the location where sequencing took place. The 27 lowest-producing countries have provided access to genetic resources that have produced less than 0.1 percent of the total sequence information. Genetic and genomic analysis is often done in collaboration, and usually uses DSI that is from the local region. Nevertheless, China provides 26 percent of the DSI used globally and is responsible for the use of 23 percent of the sequence data analysed on the WiLDSI Data Portal. Thirty-six countries qualify for the table of top user and provider countries. These include countries in five of the six FAO regions, as listed below.

Africa: South Africa, United Republic of Tanzania.

Asia and the Pacific: Australia, China, India, Iran (Islamic Republic of), Japan, New Zealand, Republic of Korea, Thailand.

Europe and Central Asia: Austria, Belgium, Czechia, Denmark, Finland, France, Germany, Italy, Netherlands, Poland, Portugal, Russian Federation, Spain, Sweden, Switzerland, United Kingdom.

Latin America and the Caribbean: Argentina, Brazil, Costa Rica, Mexico, Panama, Peru.

North America: Canada, United States of America.

Scholz *et al*.[27] explored the range of countries from which DSI has originated and the range of countries accessing and utilizing it. Despite DSI being utilized quite broadly, the level of use does not reflect the traditional rhetoric around the provider–user relationship, which assumes that provision (access) to GR happens in the Global South and that the subsequent use occurs in the Global North. Instead, the situation is much more complex, with use **and** provision happening in both directions (see below). Scholz *et al*. (2021) rejected the traditional hypothesis of the provider–user relationship. Their study included the 17 816 729 sequences in the INSDC database that carry a tag showing the country of origin (of the GR from which the DSI was generated), representing around 16 percent of the global database holdings. For each of these sequences, a publication listed within the INSDC database was counted as a "primary" publication, and any publication in the Europe PubMed Central database that listed the sequence was counted as a "secondary" publication. A total of 117 483 publications were included in the analysis.

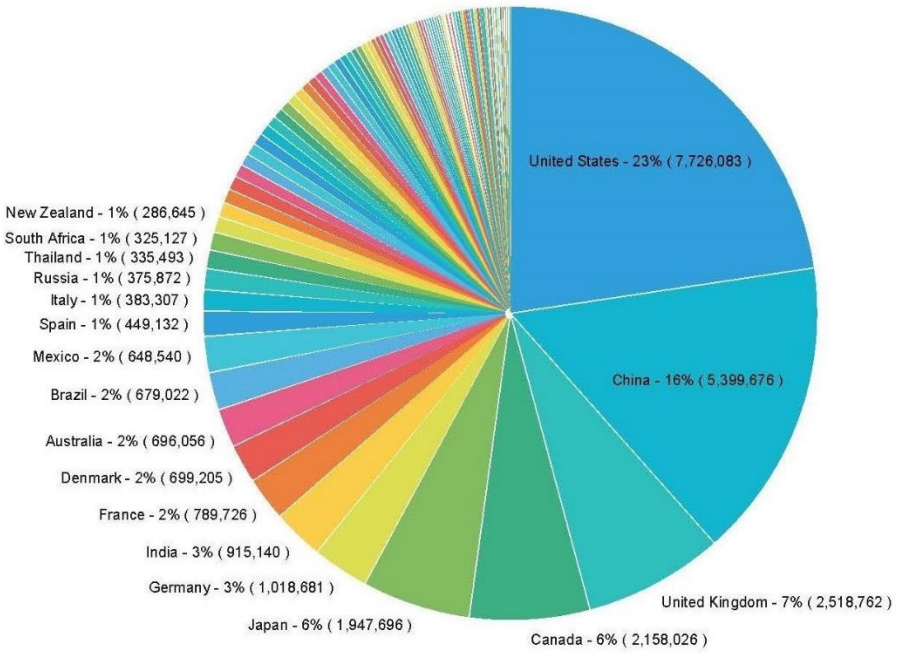
The study found that DSI is both provided and used by more than 99 percent of countries and that DSI use and provision often occur in roughly similar proportions. The figures from the WiLDSI Data Portal also show North–South collaboration (Europe and North America collaborating with South America, Africa and Asia) and South–South collaboration. The same study compared the "users" (as counted by a publication, not an individual) of nationally generated sequence information for lower-income (Group of 77 [G77]), middle-income (BRICS: Brazil, Russian Federation, India, China, South Africa) and high-income countries (Organisation for Economic Co-operation and Development – OECD). They found that in each case the biggest users of a given economic group's DSI were users located in the same economic group (i.e. users appear to be mostly using their own region's DSI).

[17] Andorra, Antigua and Barbuda, Barbados, Democratic People's Republic of Korea, Dominica, Eritrea, Eswatini, Grenada, Kiribati, Lesotho, Libya, Liechtenstein, Maldives, Marshall Islands, Mauritania, Monaco, Nauru, Palestine, Saint Kitts and Nevis, Saint Vincent and the Grenadines, Saint Lucia, Somalia, San Marino, South Sudan, Timor-Leste, Turkmenistan and Tuvalu.

However, there were significant differences in the numbers of publications in the different regions, with a total of 64 178 publications in low-income countries, 52 039 in BRICS countries and 82 971 in OECD countries.

Some countries that are heavy providers (see Figure 3) are also heavy users of DSI, which in turn drives innovation. To an extent, the use of DSI data correlates to their provision. Th the United States of America, for example, provides a great deal of access to DSI while also being a major user of DSI data. Other countries, especially in Africa, provide less DSI and are also relatively light users. Given that much of the DSI data in public databases does not have "country of origin" tags, the scope for tracing DSI provision and use by country is limited. As the data in the global resources cannot be split into sectors, it is not possible to determine easily whether the situation reflects any use of DSI on GRFA. However, searches of the agriculture-specific CABI database revealed the same pattern (see Appendix 1). A significant amount of DSI data generated from African species is housed in European and North America museums (e.g. the Natural History Museum in the United Kingdom, the National Museum of Denmark and the Smithsonian Museum in the United States of America), as are many specimens of those species. In many cases, more museum-held samples are located outside Africa than in Africa, although many of the non-African institutions have active outreach, project and visiting-scientist programmes. Again, analysis of such data must be based on the source of the genetic resources and not the generator and depositor.

Figure 3. Number and percentage of sequences provided, by country



Source: WiLDSI Portal (https://wildsi.ipk-gatersleben.de/apex/wildsi/r/wildsi/12)

The WiLDSI Data Portal also presents "country use of DSI", which makes it possible to explore how each country's scientists use their own DSI. LMICs both provide and use less DSI than, for instance, the United States of America, where almost 49 000 of the country's authors are using the country's data. Although the usage level varies greatly, the pattern of authors using their own country's DSI remains similar elsewhere: for example, the majority of authors (2 348) from South Africa use South African data. When it comes to authors using DSI from countries other than their own, the WiLDSI Data Portal reports that by far the greatest number, 118 980, are authors affiliated to the United States of America. As mentioned above, much work is done in, and for, LMICs through collaborations with higher-income countries, as demonstrated by the presence of more than twice as many North–South collaborations as South–South collaborations in the WiLDSI Data Portal.

3.2 Public databases

Currently, a vast amount of DSI is openly accessible, meaning that it is freely and easily available for anyone to access and analyse using public data resources, such as those managed by EMBL/EBI (Cochrane, 2022).[18] Taking AHTEG's DSI Groups 1 to 3 together, the main resources for storing and distributing sequence data are: the National Institutes of Health (NIH) Sequence Read Archive (maintained by National Center for Biotechnology Information – NCBI);[19] ENA; [20] and DDBJ.[21]

As of August 2021, there were 15 309 209 714 374 sequences stored. At the same point in time, the NIH Sequence Read Archive comprised 11.5 petabytes of publicly accessible data (Sayers *et al.*, 2022b).[29] The above "mirror" databases make DNA, RNA and protein sequence data available for free. They exchange data on a regular basis and therefore contain essentially the same data – all available within a single unified accession number scheme. Importantly, these data are available to all, with evidence for use, for example, by approximately 1 700 other databases that curate, organize, integrate, annotate or add some further value to the holdings.

The 2022 database issue of the journal *Nucleic Acids Research* (NAR) (Rigden and Fernández, 2022)[4] contains 185 papers spanning a wide range of biological fields and types of investigation. It includes 87 papers reporting on new databases and 85 covering recent changes to resources previously published, starting with reports from the major database providers NCBI, ENA-EBI and the National Genomics Data Centre (NGDC) in China (Sayer *et al.*, 2022a; Cantelli *et al.*, 2022; CNCB-NGDC Members and Partners, 2022). The NAR database issue also reports on data created and stored on (i) nucleic acid sequence and structure, transcriptional regulation, (ii) protein sequence and structure, (iii) metabolic and signalling pathways, enzymes and networks, (iv) genomics of viruses, bacteria, protozoa and fungi, (v) genomics of human and model organisms plus comparative genomics, (vi) human genomic variation, diseases and drugs, (vii) plants and (viii) other topics, such as proteomics databases.

Vast quantities of sequence data are generated annually, and these are stored with their associated metadata. INSDC has announced that, from 2022, it will make spatio-temporal metadata mandatory for new submissions.[22]

The Global Biodata Coalition (GBC)[23] brings together research funders to coordinate and organize the rapid growth of biodata. There is a growing need to share approaches that allow the efficient management of this process, to address associated challenges such as fragmentation and duplication of effort and to develop a strategy for long-term sustainability. The direct interest of GBC, as a coalition of research funders, is the sustainability of biodata resources that comprise an essential infrastructure for research. However, to achieve this, GBC must also address global inclusion, as the life sciences require samples/data from all parts of the world, and biologists in all countries have expertise to add to global science. The GBC will be trying to engage with parties and bring them together to push for global distribution of effort and benefit in the operation and use of biodata resources such as DSI databases.

The data in the databases covered here originate from genetic resources from all parts of the world, as demonstrated by the WiLDSI Data Portal. Not all DSI has full associated metadata that identifies the country of origin of the genetic resources from which it was derived. Public databases are addressing this issue. Without information on the country of origin, it is difficult to ensure that benefits arising from the use of the DSI are shared with the country that provided the genetic resources. Public databases, including INSDC, store information on patented sequences, which are often deposited in

[18] https://www.ebi.ac.uk/
[19] www.ncbi.nlm.nih.gov/
[20] https://www.ebi.ac.uk/ena/browser/home
[21] www.ddbj.nig.ac.jp/
[22] https://www.ebi.ac.uk/about/news/technology-and-innovation/ena-new-metadata/
[23] https://globalbiodata.org/

publicly accessible databases in order to properly disclose inventions, in line with the disclosure requirement of patent law.

3.3 Private databases

By their very nature, little information is available about privately held databases. It is clear that companies and organizations are generating sequence data and harvesting what they need from publicly available databases in order to undertake their research, which they then wish, and need, to keep confidential. In 2020, a study on DSI in public and private databases showed that the DSI stored is very diverse and that such databases are often distributed internally according to the end uses and types of data stored, which include data on proteins- (Rohden *et al.*, 2020) (see Table 6). The study goes on to say that in general, it seems that at least half of the biological data stored in private databases originated from public databases, although this is only an estimate. The authors of the study noted that privately generated DSI can also be fed back into the public domain, primarily in the form of publications or the registration of patents. The study also found that there are large quantities of unpublished private DSI that do not become part of patents and would not necessarily need to be kept private. However, there are few incentives for companies to publish these NSD. The fact that huge quantities of DSI are freely available to the private sector has not so far convinced the private sector to make its DSI similarly available (Rohden *et al.*, 2020). Unfortunately, for the purpose of the present and other studies, DSI in private databases cannot be analysed to quantify the volume of data or their uses, users or biological scope.

Table 6. Overview of private database case studies

CASE STUDY	EMPLOYEES	FOCUS FOR NSD + SI	% OF PUBLIC DATA	SUBMIT DATA TO PUBLIC	P&P PARTNERSHIPS	USE PATENT DATABASES
1: Novozymes	> 6,000	Enzymes	~ 50%	yes	yes	yes
2: Company X	> 20,000	Health, materials and nutrition	~ 95%	yes	yes	yes
3: Company Y	> 2,000	Plant breeding and seed prodution	~ 50–80%	yes	yes	yes
4: TraitGenetics	> 20	Molecular markers and genotypes in plants	?	yes	yes	no
5: BASF SE	> 122,000	Various areas	~ 50–90%	yes	yes	yes
6: Company Z	> 350	Enzymes for DNA handling	?	yes	yes	yes

Source: Rohden, F., Huang, S., Dröge, G. & Scholz, A.H. 2020. *Combined study on digital sequence information in public and private databases and traceability*. Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources, Montreal, Canada, 17–20 March 2020. CBD/DSI/AHTEG/2020/1/4. Montreal Canada. https://www.cbd.int/doc/c/1f8f/d793/57cb114ca40cb6468f479584/dsi-ahteg-2020-01-04-en.pdf

The study (Rohden *et al.*, 2020) concluded that there are likely to be thousands of companies that use the public DSI available from the INSDC and integrate it into their in-house databases, noting that some of these companies do add data to public databases, especially in the context of collaborations with public institutions. The study further concluded that backtracking to the original genetic resources by the company itself works in general for DSI generated in-house but not for all data obtained from the public databases. It also found that companies use patent NSD databases (commercial databases) to check for already existing patents but that other commercial NSD databases are uncommon. The authors note that this finding may suggest that such databases represent a challenging business model, as NSD is freely available at the INSDC and many downstream NSD and sequence information databases.

4. The state of management of DSI

This section describes the state of management of DSI. It starts by discussing the genetic sequencing technologies that generate DSI. It then describes how DSI on GRFA is used for research and development. This is followed by a discussion of where and how DSI is stored, and finally an analysis, based on DSI data flows and a literature review, of where and by whom DSI is used.

4.1 Use of DSI for food and agricultural research and development

A FAO document (FAO, 2021) has described the opportunities, challenges and implications of DSI for GRFA. The present study adds to this information through the survey of literature in the CAB Abstracts database (Appendix 1) and several case studies of CABI member countries, which confirmed that many countries are not yet in a position to make full use of DSI (Appendix 2).

4.1.1 Characterization

Elements of DSI are now used in taxonomy to characterize unknown organisms and partial sequences of genomes by comparing taxonomically informative genetic sequences with those available in databases. It is possible to extract such sequences from dead, dried specimens in museums and botanical gardens (Kates *et al.*, 2021). Comparison of new sequences against this massive, authenticated resource enables identification of the source organism.

Identifying pests (*sensu lato*) and their natural enemies is of high and direct relevance to GRFA. Using "DNA barcodes" (and more sophisticated DSI) that are compared against the Barcode of Life Data System (BOLD),[24] GenBank[25] and other major repositories, Morand (2018) reported on advances and challenges in barcoding microbes, parasites, and their vectors and reservoirs. He stated that BOLD[26] held more than 6 million barcodes from over 270 000 species (including animals, plants and fungi). Where reference barcodes are not available for the species sampled, the most similar barcodes (which may not be generated from GRFA) often give clues to the identity of the test sample. This approach can be applied to all animals, plants and microorganisms. Just a short genetic sequence is enough to identify most species (Whitfield, 2003). The Barcoding Table of Animal Species, for example, provides a new tool for selecting appropriate methods for identifying animal species using DNA barcoding (Matthes *et al.*, 2020). Hotalinga *et al*. (2021) reported 3 278 unique animal species in GenBank. There is still a lot of work to do to sequence all organisms, but huge projects have been initiated to produce the overall "Tree of Life", including the Earth BioGenome project, a global enterprise with ambitions to sequence genomes for all of Earth's eukaryotic diversity (Lewin *et al.*, 2018, 2022).

4.1.2 Use and development

Substantial advances in DNA sequencing bring the potential to enhance food security and the sustainable use of global biodiversity, and hence to benefit the world's poorest people (Cowell *et al.*, 2022). Sharing sequence information has been critical in allowing evolutionary relationships, population dynamics and gene function to be inferred from the comparison of multiple sequences. DSI, in combination with other data, such as predictive climate models, provides the foundation for finding nature-based solutions to current global challenges (Antonelli *et al.*, 2020).

Sequencing genes to discover their products is providing a rich foundation for further discovery, which in turn could generate products for the market and lead to the generation of monetary benefits. There are numerous publications demonstrating the impact of DSI studies on R&D in the field of GRFA. "Omics" technologies can drive plant engineering, ecosystem surveillance, and human and animal health. Hurgobin and Lewsey (2022) provide a collection of reviews that introduce readers to current and future use of "omics" technologies to solve real-world problems. The authors describe "omics" as a "collection of research tools and techniques that enable researchers to collect data about biological systems at a very large, or near-complete, scale. These include sequencing individual and community genomes (genomics, metagenomics), characterisation and quantification of gene expression (transcriptomics, meta-transcriptomics), metabolite abundance (metabolomics), protein content (proteomics) and phosphorylation (phospho-proteomics). Though initially exploited as tools for fundamental discovery, 'omics techniques are now used extensively in applied and translational research, e.g. in plant and animal breeding, biomarker development and drug discovery."

Numerous benefits can arise from DSI activities. Table 7 gives some examples of these.

[24] https://www.boldsystems.org/
[25] https://www.ncbi.nlm.nih.gov/genbank/
[26] http://v4.boldsystems.org/

Table 7. Benefits of DSI studies

Activity	DSI component of research	Benefits
Identification of biodiversity and its characterization	DNA barcoding and whole genome sequencing (WGS) to name the organism; annotated sequences can reveal potential traits and properties	Contributing to biodiversity inventories and data comparison, allowing biodiversity to be monitored and conservation to be improved
Diagnosis and identification of pests and diseases	Identification of causative organisms using barcoding or WGS	Enabling appropriate management recommendations and improving yields; combatting threats to livelihoods, agriculture and the environment from pests and diseases
Rapid identification of newly introduced (invasive) alien species	Sequencing and automated identification systems based on proteomics and metabolomics accelerate the identification process	Early warning that facilitates containment and management, reducing losses
Assessment of the impact of land use and climate change on biodiversity and ecosystem services, which often involves finding species new to science	Microbiome studies are often carried out, as many microorganisms can, as yet, not be grown and are only detected by chemical signatures and sequence data	Improving biodiversity inventory, climate change mitigation and soil health monitoring; identifying interventions that can result in improved agricultural production
Development of microbial solutions to improve health and nutrition security	DSI observations can lead to identification of traits and properties that can be utilized	Improving yields and reducing losses of biodiversity for food and agriculture; monetary benefits arise when products are placed on the market
Development of biological control agents (BCAs)	Identification and characterization of suitable BCAs	Management of invasive species; reduction of crop losses and minimization of unnecessary pesticide use
Increasing and improving access to agricultural and environmental scientific knowledge	Sequencing DNA, RNA, proteins and the metabolome chemical profiles of species	Improvement of the knowledge base that all can use for innovation and discovery

Animal genetic resources

DSI leads to a better understanding of the genetic basis of an animal's traits for use in breeding programmes, for example how adaptive responses to environmental changes (such as feed availability or drought conditions) can trigger phenotypic changes that affect viability and reproductive fitness.

DSI allows the genetic variability within populations to be maintained and hence contributes to the sustainable use of animal genetic resources. It can advance the discovery and development of new livestock breeds, with enhanced outcomes for sustainable and resilient livestock systems and food security. DSI can also improve understanding of traits for adaptation to new breeding conditions, particularly in the context of climate change.

Whole genome sequencing (WGS) data from the 1000 Bull Genomes Project is aiding the discovery of positive and negative traits and thus benefiting the cattle industry globally (Illumina, 2022).

Aquatic genetic resources

DSI can enable the characterization of genes and identification of genetic sequences for use in population genetics and stock assessment, and molecular markers for disease diagnosis and prevention, and for pedigree assignment in breeding programmes. DSI contributes to reproductive technologies, detection of hybrids and disease diagnosis. It can improve market access and consumer confidence in supply chains by improving traceability and identifying product substitution, and can support product labelling and certification schemes (OEWG, 2021).

DSI is used in the breeding of pikeperch (*Sander lucioperca*), a fish species of growing economic significance (de los Ríos-Pérez *et al.*, 2020). The quality of the meat of the pikeperch, which has low fat content and high protein content, gives it high commercial value, and it has become a candidate for intensive inland aquaculture. Knowledge of the pikeperch's genome enables selection of the best candidates for breeding.

Marine biotechnology is a growing and globally significant economic sector that can help address global economic challenges (Rotter *et al.*, 2021).

Forest genetic resources

DSI is contributing to the assembly of breeding populations in newly developed and advanced breeding programmes, and to the selection of genetic material for storage or micropropagation. It can potentially be used as a powerful tool for breeding forest trees as well as for enhancing the productivity of plantation forests and enabling judicious control of pest infestation. Using predictive genomics may help in the conservation of trees by identifying the environment most suited to the particular genotype and by providing information for assisted migration. Accumulated DSI enables comparison of large numbers of individuals and populations of the same and related species to identify their current distribution areas and project changes due to climate change.

Microbial and invertebrate genetic resources

DSI is commonly used in microbiology, as microorganisms, by definition, cannot be seen with the naked eye and require sequence-based technologies to detect their presence and describe them. It is now routine to use barcoding and WGS to identify microorganisms. Automated identification tools that depend on comparisons with reliable and complete databases are proving extremely useful, especially for the detection of organisms of regulatory importance. It is also possible to express properties through genome engineering and even produce chemically synthesized genomes (see Box 2).

The development of high-throughput molecular methods utilizing specific gene regions (barcodes) provides massive amounts of DSI on microorganisms. Although a large fraction (more than 50 percent) of the detected genes have no assigned function to date, the use of functional metagenomics applications and tools has led to novel enzyme discovery. Importantly, shotgun metagenome analysis allows the sequencing of the genomes of all microbial species in a sample, for example a sample from the marine environment, where the majority of microorganisms remain to be discovered. This enables enzymes of interest to be linked directly to organisms, thus allowing the properties of unculturable microbes to be accessed (Rotter *et al.*, 2021). Improved technologies have led to the discovery of many novel bioactive compounds through the sequencing of complete microbial genomes from selected niches, enabling bioprospecting for marine microorganisms. Open-access knowledge bases containing tandem mass spectrometry (MS/MS) data or structures of microbial natural products have been greatly enhancing dereplication processes (excluding those already discovered), leading to the identification of new molecules and natural products (Rotter *et al.*, 2021).

Box 2 *Saccharomyces cerevisiae* as an example organism

Genome-scale engineering is employed in bacterial systems, and an automated platform for multiplex genome-scale engineering is being used for the yeast *Saccharomyces cerevisiae* (Si *et al.*, 2017). This

yeast is an important eukaryotic model and a widely used microbial cell factory. Standardized portions of the genome are created in a single step from a full-length complementary DNA library with the aid of CRISPR-Cas technology (removing, adding or altering sections of a sequence) (Pickar-Oliver and Gersbach, 2019). These genetic parts are iteratively integrated into the repetitive genomic sequences in a modular manner using robotic automation. This allows expression of diverse phenotypes, including cellulase expression, isobutanol production, glycerol utilization and acetic acid tolerance, and may accelerate genome-scale engineering endeavours in yeast.[16] A complete synthetic version of a highly modified *Saccharomyces cerevisiae* genome, reduced in size by nearly 8 percent, has been designed. Chemically synthesised genomes like this are customizable and allow scientists to ask questions about chromosome structure, function and evolution with a bottom-up design strategy (Richardson *et al.*, 2017).

Sources: Pickar-Oliver, A. & Gersbach, C.A. 2019. The next generation of CRISPR–Cas technologies and applications. *Nature Reviews Molecular Cell Biology*, 20: 490–507. https://www.nature.com/articles/s41580-019-0131-5; Richardson, S.M., Mitchell, L.A., Stracquadanio, G., Yang, K., Dymond, J.S., Dicarlo, J.E., Lee, D. *et al*. 2017. Design of a synthetic yeast genome. *Science*, 355(53223): 1040–1044. https://www.science.org/doi/10.1126/science.aaf4557; Si, T., Chao, R., Min, Y., Wu, Y., Ren, W. & Zhao. H. 2017. Automated multiplex genome-scale engineering in yeast. *Nature Communications*, 8: 15187. https://doi.org/10.1038/ncomms15187

DSI is used extensively to identify and characterize microorganisms. Fungi are often employed as biological control agents (BCAs), termed biopesticides, and their conidiation capacity (capacity to produce spores – propagules) and conidial quality are critical to their production and application, for example in the mass production of fungal insect pathogens, such as *Metarhizium acridum* (Zhang, Peng and Xia, 2010).

Metabolomics is an emerging tool for studying plant–microbe interactions (Gupta, Schillaci and Roessner, 2022). It enables access to the cellular metabolites, which belong to the AHTEG's Group 3 (AHTEG, 2020). Gupta, Schillaci and Roessner, in 2022 report that in natural environments, interaction between microorganisms and plant roots, surfaces and metabolic processes are common. Metabolomics research based on mass spectrometric techniques underpins systems biology and relies on precision instrument analysis, providing a qualitative and quantitative approach to determining the mechanisms of multitrophic relationships between bacteria, fungi and plants. This also helps to elucidate the tolerance mechanisms of host plants against various abiotic stresses. Metabolomics is a data-driven, hypothesis-generating approach that detects and quantifies thousands of compounds per analysis, enabling the study of complex biological interactions in the rhizosphere and reciprocal responses between plants and organisms. The use of metabolomics to study plant–microbe interactions involves challenges, such as identifying the origin of the metabolites analysed, uncovering the metabolic complexity of multiple interacting organisms, and linking metabolome information with other "omics" data such as transcriptomics, proteomics or phenomics (Gupta, Schillaci and Roessner, 2022).

It is now possible to explore microbial communities that are present in the environment and surround plants, animals and humans. These communities comprise bacteria, archaea and fungi, among other microscopic organisms that have potential for use to improve plant growth and crop yield as well as human and animal health. The composition and differentiation of microbial communities are now being explored, but understanding the microbiome remains challenging. For example, there is a need to understand the microbial community's interactions with other organisms and its environment and to identify ways of optimizing the taxonomic composition of microbiomes to improve the overall health and fitness of the plant and the soil, and hence to help make agriculture more sustainable (Singh and Goodwin, 2022). With regard to the maize microbiome, Singh and Goodwin point out the huge opportunities that genomics may offer in terms of improving yields and facilitating adaptation to climate change. It is clear that generation of DSI will only increase and that its importance in the food and agriculture sector is still becoming apparent.

The use of locally acclimatized rhizobial strains can replace the use of nitrogen-based fertilizers for the cultivation of soybean. Chibeba *et al*. (2017) used DSI-generating technologies (BOX-PCR fingerprinting) to select ten isolates from Mozambique that outperformed a commercially available strain, providing a possible strategy for increasing soybean yields.

Crop rotation can improve soil properties and is an important way of preventing soil-borne diseases. A study (Zhang *et al.*, 2022a) that used different preceding crops and combinations of soil microorganisms for the control of clubroot disease in Chinese cabbage showed that growth and disease resistance could be improved. Different combinations of preceding crops, including soybeans, potato, onions and wheat, were used, and metagenomic sequencing demonstrated the differences they induced in the abundance and diversity of the bacteria and fungi. The study showed that the preceding crops changed the structure of soil microbial communities, reduced clubroot disease in Chinese cabbage, promoted growth and suppressed disease.

Another study (Bonanomi *et al.*, 2020) showed how intensive agricultural practices negatively affect soil fertility and soil microbial communities and compromise the crop quality and yield of rocket (*Eruca sativa)*. High-throughput sequencing was used to monitor changes in populations of bacteria and fungi after chemical applications. This demonstrated that synthetic fertilizer and fumigation induced soil acidification and increased soil salinity, with a detrimental impact on microbial diversity, activity and function, and resulting negative effects on crop yield. The application of organic amendments significantly improved crop yield, especially when alfalfa and glucose were applied as a single dose.

Insects

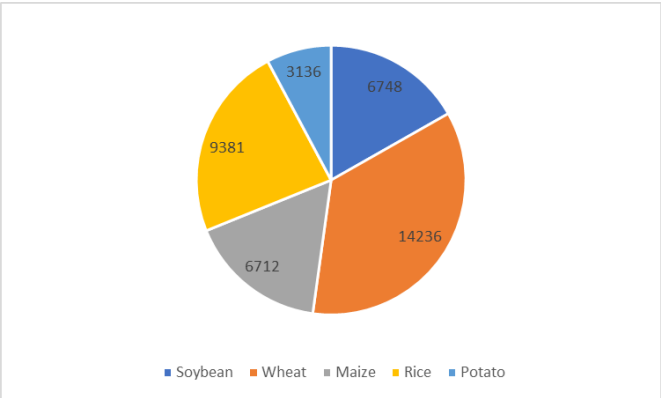
DNA barcoding has been used successfully for biosurveillance of forest and agricultural pests (Javal *et al.*, 2021). The authors of the study cited point out the urgent need to develop advanced tools for the early detection and accurate identification of new or emerging insect pests, and note that the online database BOLD allowed them to automatically identify species based on their DNA barcodes.

Although barcode databases are far from comprehensive for insects, DNA barcoding can complement existing morphological identification tools and will facilitate the identification of the most common species, irrespective of the taxonomic skills of the observer or the developmental stage of the insect Javal *et al.*, 2021).

Plant genetic resources

The literature survey carried out on CAB Abstracts (Appendix 2) provided many examples of the use of DSI to improve crops. Figure 4 summarizes the number of literature records for soybean, wheat, maize, rice and potato.

Figure 4. The number of literature records for soybean, wheat, maize, rice and potato referencing DSI in the CAB Abstracts database



A deeper look into these papers reveals the technologies used and the advances made. DSI is used in several ways to improve crop yields and resistance to disease, and to address threats to biodiversity

such as invasive species and climate change. The technologies in question include epigenome-guided crop improvement. The epigenome is described as a "multi-modal layer of information superimposed on DNA sequences" that informs gene expression and can improve crop performance (Zhang *et al.*, 2022b). Some consider that epigenetic data do not qualify as DSI and should be outside the scope of any benefit-sharing regime. However, epigenetics is considered part of the digital flow of information from biodiversity and is included under the AHTEG groupings of the scope of DSI, (AHTEG, 2020) (see Group 2 in Table 1). Epigenome-guided crop improvement can "identify and select for heritable epialleles that control crop traits independent of underlying genotype" (Zhang *et al.*, 2022b). In doing so, specific opportunities and challenges for grain and horticultural crops have been identified, for example the genomic profile of a leaf in fruit and grain crops could potentially be extrapolated to predict genes and guide genome modification in other tissues or organs (Zhang *et al.*, 2022b).

The hits in the CAB Abstracts database are summarized for wheat, rice, maize and soybean in Appendix 1. Examples from these publications of how DSI is being used in crop improvement and sustainability are presented in Box 4. In addition to the use of DSI from specific crops to improve production of the respective species, there are also examples in which it is being used to enhance other crops: this is discussed in Section 4.1.4.

Box 4. DSI from food crops

Wheat is the most widely grown crop globally, providing 20 percent of all human calories and protein (Cakderini *et al.*, 2021). Wheat yields globally are affected by climate change, for instance by drought or heat stress. For example, losses of between 17 percent and 45 percent of total production (about 8 million tonnes) were recorded in France in 2016. Candidate genes for improved abiotic stress tolerance have been discovered (Schmidt *et al.*, 2020). Wheat characters such as leaf chlorophyll content, leaf greenness, cell-membrane thermostability and canopy temperature have been proposed as candidate traits for improving adaptation and yield potential under high temperatures (Pradhan *et al*., 2020).

A genome-wide association study (GWAS) discovered marker-trait associations (MTAs) that affect grain yield and yield-related traits, some of which were found in genes encoding different types of proteins associated with heat stress. These MTAs can be used in marker-assisted selection and breeding to develop varieties with high stability for grain yield under high temperatures (Pradhan *et al*., 2020). Similarly, GWAS (Hutter, 2022) has been used to evaluate grain yield-related traits, including fruiting efficiency in wheat, and 44 significant MTAs have been identified, some of which are in novel loci (Gerard *et al.*, 2019). These markers are linked to fruiting efficiency, grain number and spikelet weight, and can potentially be used as selection criteria to increase yield potential in wheat breeding programmes. In the case of multigene complex traits, such as yield, there is a need for improved strategies for research into which genes are responsible and how new varieties can be generated (Skraly *et al.*, 2018). Predictive models facilitate the identification of gene targets, but these must provide a metabolic perspective, for example by addressing the transcriptome to ensure such properties are expressed (Skraly *et al.*, 2018). The method for trait engineering can affect the commercialization cost and timeline (Skraly *et al.*, 2018).

Rice is a staple food crop for more than half the world's population and has more than 110 000 cultivated varieties. It is grown in more than 100 countries, with 90 percent total global production occurring in Asia (Fukagawa and Ziska, 2019). Studies using DSI have been carried out on rice in relation to drought and heat tolerance, where the overexpression of the Rab7 gene improves tolerance and increases grain yield. This works by modulating expression of osmolytes, antioxidants and abiotic stress-responsive genes. It could also be used to improve grain yield and stress tolerance (El-Esawi and Alayafi, 2019). A combination of stress tolerance (e.g. to salinity and drought) and enhanced grain yield is a major focus of rice-breeding strategies (Faisal *et al.*, 2017). Modifying the function of the drought and salt tolerance gene by downregulating it using artificial microRNA technology produced transgenic plants that had higher stress tolerance and better yields.

Maize is one of the most important food crops in the world after wheat and rice (Shiferaw *et al.*, 2011). Erenstein *et al.* (2022) report that "together, the three big global staple cereals – wheat, rice, maize – comprise a major component of the human diet, accounting for an estimated 42 percent of the world's food calories and 37 percent of protein intake." Cell wall invertase genes increase maize grain yield and starch content by up to 145.3 percent in transgenic maize plants compared to the wild-type plants, as confirmed in two-year field trials at different locations (Li *et al.*, 2013). The dramatically increased grain yield is due to enlarged ears that have more and larger grains along with increased total starch content (up to 20 percent) in the transgenic kernels. The results suggest that the cell wall invertase gene can be genetically engineered to improve both grain yield and grain quality (Li *et al.*, 2013).

Simmons *et al*. in 2020 described how DSI studies could identify genes that could be harnessed to enhance yield in commercial strains of maize. Hannah *et al.* (2012) found that the maize shrunken-2 (*Sh2*) gene encodes the large subunit of the rate-limiting starch biosynthetic enzyme ADP-glucose pyrophosphorylase. Expression of a transgenic form of the enzyme with enhanced heat stability and reduced phosphate inhibition increased maize yield by up to 64 percent. These DSI studies have enabled a greater understanding of maize production and helped identify interventions that can improve production and contribute to climate change adaptation.

Chickpea is a highly nutritious grain legume crop that is widely consumed, especially in the Indian subcontinent. The major constraints to chickpea production are biotic stresses (*Helicoverpa*, bruchid beetles, aphids, *Ascochyta*) and abiotic stresses (drought, heat, salt, cold), which reduce the yield by up to 90 percent. Recent advances in gene technologies have enabled the development of genetically modified chickpeas, including ones that are resistant to *Helicoverpa armigera*, *Callosobruchus maculatus* and *Aphis craccivora* and to drought and salt stress (Kumar *et al.*, 2018).

Soybean is a major animal feed crop that is being negatively affected by climate change and for which future yield losses are predicted (Fodor *et al.*, 2017). DSI studies in soybean have explored the roles of a transcription factor in regulating fatty acid biosynthesis and have demonstrated that it could influence many aspects of plant structure and growth. Overexpression of the gene GmWRI1b improves yields and increases total seed oil production under field conditions (Guo *et al.*, 2020).

Sources: Calderini, D.F., Castillo, F.M., Arenas-M, A., Molero, G., Reynolds, M.P., Craze, M., Bowden, S. et al. 2021. Overcoming the trade-off between grain weight and number in wheat by the ectopic expression of expansin in developing seeds leads to increased yield potential. *New Phytologist*, 230(2): 629–640. https://doi.org/10.1111/nph.17048; Faisal, A., Biswas, S., Zerin, T., Rahman, T. & Seraj, Z. 2017. Downregulation of the DST transcription factor using artificial microRNA to increase yield, salt and drought tolerance in rice. *American Journal of Plant Sciences,* 8(9): 2219–2237. https://doi.org/10.4236/ajps.2017.89149; Fodor, N., Challinor, A., Droutsas, I., Ramirez-Villegas, J., Zabel, F., Koehler, A-K. & Foyer, C.H. 2017. Integrating plant science and crop modeling: Assessment of the impact of climate change on soybean and maize production. *Plant and Cell Physiology*, 58(11): 1833–1847. https://doi.org/10.1093/pcp/pcx141; Fukagawa, N.K. & Ziska, L.H. 2019. Rice: importance for global nutrition. *Journal of Nutritional Science and Vitaminology*, 65(Supplement): S2–S3. https://doi.org/10.3177/jnsv.65.S2; El-Esawi, M.A. & Alayafi, A.A. 2019. Overexpression of rice Rab7 gene improves drought and heat tolerance and increases grain yield in rice (*Oryza sativa* L.). *Genes*, 10(1): 56. https://doi.org/10.3390/genes10010056; Erenstein, O., Jaleta, M., Sonder, K., Mottaleb, K. & Prasanna, B.M. 2022. Global maize production, consumption and trade: trends and R&D implications. *Food Security*, https://doi.org/10.1007/s12571-022-01288-7; Gerard, G.S., Alqudah, A., Lohwasser, U., Börner, A. & Simón, M.R. 2019. Uncovering the genetic architecture of fruiting efficiency in bread wheat: a viable alternative to increase yield potential. *Crop Breeding & Genetics*, 59(5): 1853–1869. https://doi.org/10.2135/cropsci2018.10.0639; Guo, W., Chen, L., Chen, H., Yang, H., You, Q., Bao, A., Chen, S. *et al*. 2020. Overexpression of GmWRI1b in soybean stably improves plant architecture and associated yield parameters, and increases total seed oil production under field conditions. *Plant Biotechnology Journal*, 18(8): 1639–1641. https://doi.org/10.1111/pbi.13324; Hannah, L.C., Futch, B., Bing, J., Shaw, J.R., Boehlein, S.,

Stewart, J.D., Beiriger, R., Georgelis, N. & Greene, T. 2012 A shrunken-2 transgene increases maize yield by acting in maternal tissues to increase the frequency of seed development. *The Plant Cell,* 24(6): 2352–2363. https://doi.org/10.1105/tpc.112.100602; Hutter, C. 2022. Genome-wide association studies. In: *National Human Genome Research Institute*. Bethesda USA. Cited 13 December 2022. https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies#:~:text=A%20genome%2Dwide%20association%20study,disease%20or%20a%20particular%20trait; Li, B., Liu, H., Zhang, Y., Kang, T., Zhang, L., Tong, J., Xiao, L. & Zhang, H. 2013. Constitutive expression of cell wall invertase genes increases grain yield and starch content in maize. *Plant Biotechnology Journal,* 11(9): 1080–1091. https://doi.org/10.1111/pbi.12102; Kumar, M., Yusuf, M.A., Nigam, M. & Kumar, M. 2018. An update on genetic modification of chickpea for increased yield and stress tolerance. *Molecular Biotechnology,* 60: 651–663. https://doi.org/10.1007/s12033-018-0096-1; Pradhan, S., Babar, M.A., Bai, G., Khan, J., Shahi, D., Avci, M., Guo, J. *et al*. 2020. Genetic dissection of heat-responsive physiological traits to improve adaptation and increase yield potential in soft winter wheat. *BMC Genomics,* 21: 315. https://doi.org/10.1186/s12864-020-6717-7; Schmidt, J., Garcia, M., Brien, C., Kalambettu, P., Garnett, T., Fleury, D. & Tricker, P.J. 2020. Transcripts of wheat at a target locus on chromosome 6B associated with increased yield, leaf mass and chlorophyll index under combined drought and heat stress. *PLoS ONE,* 15(11): e0241966. https://doi.org/10.1371/journal.pone.0241966; Shiferaw, B., Prasanna, B.M., Hellin, J. & Bänziger, M. 2011. Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. *Food Security,* 3: 307. https://doi.org/10.1007/s12571-011-0140-5; Simmons, C.R., Weers, B.P., Reimann, K.S., Abbitt, S.E., Frank, M.J., Wang, W., Wu, J., Shen, B. & Habben, J.E. 2020. Maize BIG GRAIN1 homolog overexpression increases maize grain yield. *Plant Biotechnology Journal,* 18(11): 2304–2315. https://doi.org/10.1111/pbi.13392; Skraly, F.A., Ambavaram, M.M.R., Peoples, O. & Snell, K.D. 2018. Metabolic engineering to increase crop yield: from concept to execution. *Plant Science,* 273: 23–32. https://doi.org/10.1016/j.plantsci.2018.03.011

4.1.4 Cross-species knowledge transfer and research on metabolic pathways

Plants produce many specialized metabolites with distinct biological activities and potential applications outside agriculture (Muhich *et al.*, 2022). Despite this potential, most biosynthetic pathways, including for metabolite production, remain poorly understood. However, advances in genomic technologies have enabled the identification of some specialized metabolic pathways, including those involved in improving abiotic and biotic stress resistance and in boosting nutritional content. Muhich *et al.* (2022) reviewed the potential and limitations of (1) identifying these metabolic pathways and (2) using the discovered enzymes in synthetic biology or crop engineering. This technology demonstrates that the use of DSI in plants is not restricted to one sector, with DSI enabling discoveries that can be utilized in other sectors, for example industrial enzymes.

Cell- and tissue-specific "omics" techniques can be used to improve plant productivity. Hurgobin and Lewsey (2022) report that plants have unique patterns of gene expression and protein and metabolite content, enabling specific patterns of growth, development and physiology. They explain that plants normally considered as resources for agriculture can provide properties for use in medicines, textiles and construction materials. There are many instances of cross-species knowledge transfer.

Capacity to assimilate carbon and nitrogen and to transport and convert incoming sugars and amino acids into storage compounds is a key determinant of crop yield (Vallarino *et al.*, 2020). Vallarino *et al*. in 2020 demonstrated that genes artificially introduced into tomato from various sources, including potato and *Arabidopsis,* increase carbon and nitrogen flows and boost fruit yield by up to 23 percent. Lack of potassium in soil limits crop yield, and under such circumstances crops require improved potassium-use efficiency (KUE). Many genes influence KUE in plants. A pyrophosphatase gene that was induced by low potassium stress was identified by Zhou *et al.*, (2020) who report that overexpression of this gene in two wheat varieties resulted in increases in yield, grain number per spike, plant height and potassium uptake in four transgenic lines over several years.

Understanding of the mechanisms that can be used to increase crop yield can be explored and improved by studying relevant DSI. For instance, a study found that overexpression of transcription

factors regulating photosynthesis and related metabolism in switchgrass gave rise to an increase of 160 percent in above-ground biomass (Ambavaram *et al.*, 2018).

DSI has been used to improve plant breeding programmes for cereal varieties with improved salinity tolerance and more profitable grain yields in saline soils. The *Arabidopsis* gene encoding a vacuolar proton-pumping pyrophosphatase has been shown to improve the salinity tolerance of transgenic barley plants in greenhouse conditions (Schilling *et al.*, 2013). The transgenic barley was found to have higher grain yield per plant in field studies under saline conditions.

Studies on DSI relating to cell proteins can also be carried out with the aim of improving crop resilience and making discoveries that can be used in breeding programmes. Targeted overexpression of an α-expansion in early developing wheat seeds led to an increase in grain yield of 11.3 percent in field experiments (Calderini *et al.*, 2021). In photosynthetic organisms, the photosystem II complex is the system most vulnerable to thermal damage, and it is therefore an obvious target for efforts to improve a crop's resilience to climate change. Expression of the chloroplast-based gene driven by a heat-responsive promoter protects transgenic rice plants from severe loss of protein and dramatically enhances their biomass and grain yield under heat. These findings represented a breakthrough in bioengineering plants to achieve efficient photosynthesis and increase crop productivity under both normal and heat-stress conditions (Chen *et al.*, 2020).

A study of regulated expression of isopentenyltransferase, a critical enzyme in the cytokinin biosynthetic pathway, demonstrated how to significantly improve the drought tolerance of groundnuts in both laboratory and field conditions (Qin *et al*., 2011). To understand the role of jasmonate plant hormones in tuberization in potato, the *Arabidopsis* jasmonic acid carboxyl methyltransferase gene was constitutively overexpressed in transgenic potato plants (Sohn *et al.*, 2011). Increases in tuber yield and size, as well as in *in vitro* tuberization frequency, were observed in the transgenic plants.

In summary, plant genome sequencing for crop improvement enables the discovery of genes and molecular markers associated with diverse agronomic traits. This, in turn, creates new opportunities for crop improvement (Edwards and Batley, 2009). However, converting these data into knowledge that can be applied in crop breeding programmes remains a challenge.

Genomic sequence information, coupled with phenotypic and other data, may also identify genotypes that are adapted to different, and changing, agroecological conditions. When integrated into crop breeding programmes, genomic sequence information is increasingly useful in efforts to achieve targeted, efficient uses of genetic diversity in sustainable agriculture.

4.2 The role of DSI in the conservation of genetic resources for food and agriculture

DSI is an important tool in conservation in that it enables the identification and characterization of genetic diversity. It is helping us understand life and evolutionary processes and enabling the discovery of new approaches to the conservation of endangered species. The DNA unique digital barcode is increasingly used to identify species, describe the composition of communities, and combat poaching and illegal wildlife trade (Mongabay, 2017a) by identifying closely related animal species when trade in some but not all of them is illegal, for example among sharks and rays (Mongabay, 2017b). Efforts are also ongoing to apply the technology to plant products, such as timber.[27] Inclusion of molecular data in biodiversity inventories allows changes over time to be tracked, as required for countries' biodiversity monitoring under the CBD (Cowell *et al.*, 2022). Genomic analysis provides an alternative method for evaluating long-term *in situ* conservation programmes.

Digital genomic sequence data are used to assess the genetic diversity of *ex situ* collections and to identify unique germplasm in farmers' fields that is not included in collections. This baseline information is essential for developing more effective *ex situ* and *in situ* conservation strategies (Halewood *et al.*, 2017).

Supple and Shapiro (2018) discuss how genome-scale data can inform species delineation in the face of admixture (the mix of diverged or isolated genetic lineages**)** and identify adaptive alleles, and thus

[27] https://wildtech.mongabay.com/2016/09/experts-hack-away-portable-dna-barcode-scanner-fight-timber-wildlife-trafficking/

enhance evolutionary rescue based on genomic patterns of inbreeding. “Conservation genomics” (Supple and Shapiro, 2018) encompasses the idea that genome-scale data will improve the capacity of resource managers to protect species. It has only recently become possible to generate genome-wide data at a scale that is useful for conservation, and in future this will have a positive impact on policy and management.

DSI is important in conservation and for the use of genetic resources, its generation, storage, management and availability are critical if we are to take advantage of it and harness it to improve output of GRFA.

5. Obstacles to access and use of DSI, and the need for capacity-building

All sequencing methods require access to data and to information on the prior use of the methods, i.e. resequencing a crop variety needs a reference genome, while *de novo* sequencing needs gene models for annotation, expression and association studies. It is therefore important for data to be openly accessible and held in a single, uncomplicated system. Any sequence is only useful for research or development if it can be compared with sequences from other studies, countries and species. For this reason, data need to be *findable, accessible, interoperable* and *reusable*, i.e. comply with the so-called FAIR principles (Wilkinson *et al.*, 2016). Any ABS system must consider such principles for data accessibility and use. The INSDC collaboration of publicly accessible, open-access databases enables this to occur, but the process would benefit from the implementation of standard methodologies for data collection, recording and sharing to maintain the consistency and reproducibility of results. Currently, data are often deposited by researchers to meet the policy requirements of scientific journals when publishing their findings. It is important that sample (i.e. DSI) metadata not only cite the country where the genetic resource originated but also provide details of the methodology used to produce it. Views from a recent report on open access are presented in Box 3.

Box 3 A perspective on the debate on benefit-sharing and digital sequence information

Frictionless data sharing and use of DSI in public databases minimize transaction costs in accessing and using data but this has implications for the design of benefit-sharing from DSI. First, to maintain the high degree of interoperability that characterizes the status quo, a multilateral access model applied as universally as possible (i.e. across all DSI and all international benefit-sharing fora) is needed. This could take the form of uniform terms of access for DSI across public databases. Benefit-sharing obligations that apply to the entire DSI dataset globally will best protect the open system. These types of benefit-sharing obligations are decoupled from access, which remains open. In comparison, options that require accounting of DSI access, movement, and use (bilateral mechanisms) appear more likely to impair interoperability and to generate high transactions costs and frictions to data flow that will significantly hinder research.

A historical context for open access and open scientific research data can serve as a starting point for characterizing and defining open access in the context of DSI. This may assist in the development of a working definition tailored to DSI. However, a consensus definition might not necessarily need to be the primary focus. Instead, it is advisable that policy discussions should pay closer attention to whether, and to what extent, scientific research and innovation would be significantly hindered by changes to the current “open and unrestricted” access and use of DSI in public databases. This lens appears better suited to guiding discussions on the design of the access pillar for any potential benefit-sharing solution.

Taking inspiration from the Scholarly Publishing and Academic Resources Coalition (SPARC)’s approach, it is useful to look beyond the question “what is open access?” towards a more nuanced approach to the design and evaluation of ABS policy solutions based on the question “is it as open as possible?” Efforts should be made to ensure that any necessary changes made to the status quo are proportionate and justified. Applying this nuanced approach to benefit-sharing objectives suggests that a multilateral and universal mechanism for DSI should be “as open as possible” provided it can be designed to deliver benefits that are deemed acceptable by the Parties to the Nagoya Protocol. Certainly, the scientific community’s quest to ensure that open access to DSI will continue to be guaranteed, and that biological data will be publishable, available, linkable, downloadable and

continue to flow into the downstream databases and software that they use, is strongly aligned with international, regional and national policies concerning science and innovation. Globally, there is a move towards greater openness to promote research, innovation and technological development, with the ultimate goal of sustainable economic development. The outstanding questions are whether the CBD and the broader benefit-sharing community will follow or go against this trend, and what the consequences of that choice will be.

Source: adapted from Rodrigo, S., Hufton, A.L., Sett, S. & Scholz, A.H. 2022. *A technical assessment for the debate on benefit-sharing and digital sequence information*. Zenodo. DOI: 10.5281/zenodo.5849643.

DSI databases offer a crucial research infrastructure for the global research community. Beagrie and Houghton, in 2021, estimated that the annual return on investment in research and development depending on EMBL-EBI managed data is GBP 1.3 billion. More than 4 900 researchers participated in the study. The most direct measure of the value is the time researchers spend using EMBL-EBI data resources. This added up to more than 140 million hours during 2020, equivalent to an estimated GBP 5.5 billion.

In some areas of research, tracing a sample to its country of origin could be crucial. Adding this information will enrich the scientific value of the data, especially for scientists working on infectious disease, biodiversity and ecology. Cochrane (2022) discussed how to ensure that countries rich in biodiversity can benefit from research on this biodiversity and discoveries resulting from such research. He reported that megaprojects such as the Darwin Tree of Life,[28] African BioGenome Project[29] and the Earth BioGenome Project[30] are daily sequencing hundreds of new species, and discussed how the data are made available to the scientific community. The INSDC currently names only those who have submitted samples or sequence data and not the primary owners or custodians of the sample (Ebenezer *et al.*, 2022). The new proposals to include improved metadata with submitted samples[31] may need to be extended to enable links to the locality of the genetic resources sequenced and to those who manage it, for example to the Indigenous Peoples or local communities providing the material for sequencing, rather than only to the person/entity that submitted the samples (Ebenezer *et al.*, 2022).

Databases are constantly exchanging data to ensure all are up to date (see Figure 2). Currently, sequence data are available to all from public databases such as INSDC, and free online training is available.[32] GenBank also provides instructions and tools for accessing and utilizing the data.[33] Theoretically, therefore, DSI data are available to all given a little knowledge and computer capacity. However, full analysis requires more specialist knowledge and bioinformatic skills. The ENA and INSDC have a uniform policy[34] of granting free and unrestricted access to all their data records, giving scientists worldwide access these data and the freedom to publish any resulting analysis as long as the original submission is cited. This practice enables traceability to source and follows the accepted practices of scientists utilizing published scientific literature. The journal *Nature* provides data repository guidance[35] to facilitate access to data. In the health sciences, some repositories have datasets requiring restricted data access, for example where there is a need for participant anonymity in clinical datasets. There is overwhelming support for keeping DSI data open access provided that

[28] https://www.darwintreeoflife.org/
[29] https://africanbiogenome.org/
[30] https://www.earthbiogenome.org/
[31] https://www.ebi.ac.uk/about/news/technology-and-innovation/ena-new-metadata/
[32] https://hbctraining.github.io/Accessing_public_genomic_data/lessons/accessing_public_experimental_data_odyssey.html
[33] https://www.ncbi.nlm.nih.gov/genbank/
[34] https://www.ebi.ac.uk/ena/browser/about/policies
[35] https://www.nature.com/sdata/policies/repositories

the source metadata are published with the DSI, thus enabling traceability back to the provider country.

The outcomes of the analysis of the WiLDSI Data Portal data suggest that the global goal should be to increase the scientific output and generation of DSI from LMIC (G77) and BRICS countries to levels similar to those observed in the OECD (Scholz *et al.*, 2021). Increased research capacity in LMICs would have global benefits and would allow global biodiversity knowledge gaps to be filled more effectively. To do this, any DSI policy mechanism should recognize the existing divide between richer and poorer countries, and encourage DSI use, publication and collaboration, perhaps explicitly dedicating significant capacity-building to scientifically levelling the DSI playing field.

Access to and use of DSI in many countries are still constrained by serious obstacles. There is an urgent need to address deficits caused, for example, by a lack of technical infrastructure, financial and human resources, educational and training opportunities, scientific collaboration, computing infrastructure, reliable electricity and high-speed internet, and potentially in the future by prohibitive charges for database use. The feedback received by CABI centres from their own researchers and their partner institutions supports this finding (Table 8). CABI has centres[36] in Brazil, China, Ghana, India, Kenya, Malaysia, Pakistan, Switzerland, Trinidad and Tobago, the United Kingdom and Zambia that use GRFA in their research and development activities. All run projects with numerous partners[37] under the following seven main themes[38] (numbers of projects given in brackets): climate change (2), crop health (57), development communication and extension (40), digital development (23), invasive species (81), publishing (4) and value chains and trade (40). These are carried out in 85 countries[36] in the following regions: Africa (51), Asia (49), Central America and the Caribbean (8), Europe (25), North America (29), Oceania (6) and South America (3). Feedback was obtained from CABI Centres and partners in Brazil, the Caribbean (including the Bahamas and Trinidad and Tobago), China, Ghana, Kenya, Pakistan and Zambia.

All those consulted reported (Table 8) that they use DSI in the course of their work, mainly for the identification and characterization of organisms. However, they noted that this DSI is often generated through collaboration, for instance with Australia, the United States of America, European countries or South Africa. Pakistan, for example, identified microorganisms through the use of DSI in collaboration with China and Saudi Arabia. The CABI Centre in India has been working in the region since 1948 and has witnessed rapid growth in the generation and use of DSI. It reported that "what started as a trickle in early 2000 is now a flood." The resulting data are mostly published in public databases. Researchers in all the countries consulted have accessed DSI from other countries for genetic improvement programmes. In India, some institutions have online databases, while others do not but are working to create them. However, much of the DSI generated is for local use and it is often not published or shared.

Globally, the generation and use of DSI is expanding rapidly, but for the majority it is often still hindered by a lack of resources, including funding and expertise. Even in China, recognized by the DSI network analysis of the WiLDSI Data Portal as one of the world's largest generators and user of DSI, the ability to use DSI fully is not shared evenly by all researchers across the country.

Country responses to questions on accessibility and ability to use DSI are summarized in Table 8, where a number of constraints to the generation of DSI, access to DSI, analysis of DSI and utilization of DSI to its full potential are reported. Many respondents mentioned a lack of investment in sequencing infrastructure, problems with internet access, shortages of trained staff, financial constraints to paying fees for access, and a major shortage of people to carry out data analysis. Redressing the imbalance in the ability to access and fully use DSI will require capacity building.

[36] https://www.cabi.org/what-we-do/cabi-centres/
[37] https://www.cabi.org/what-we-do/how-we-work/
[38] https://www.cabi.org/what-we-do/cabi-projects/

Table 8. Commonalities from CABI Centre feedback on capacity and ability to access and use DSI

	Bahamas	Brazil	China	Ghana	India	Kenya	Malaysia	Pakistan	Trinidad and Tobago	United Kingdom	Zambia
1. Does the centre operate across a region?	-	Y	-	Y	-	Y	Y	-	Y	Y	Y
2. Does the country use DSI?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
2.1 Private sector use of DSI	-	Y	-	-	Y	-	-	-	-	Y	-
2.2 Public sector use of DSI	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3. Does the country generate DSI	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3.1. In-country	-	Y	Y	N	Y	-	Y	Y	N	Y	Y/N
3.2 By out-sourcing	Y	Y	-	Y	N	Y	N	Y	Y	Y	Y
4. Is there country legislation covering use of genetic resources/DSI?	-	Y	-	-	Y	-	Y	-	-	N	-
4.1. Require benefit sharing at access	-	-	-	-		-	-	-	-	N	-
4.2 Benefits triggered by product on the market	-	Y	-	-		-	-	-	-	N	-
5. Are there constraints to generation or access of DSI?	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
5.1. Sequencing Infrastructure	Y	Y	N	Y	N	Y	N	N	Y	Y	Y
5.2 Internet access	Y	N	Y	Y	-	Y	Y	N	-	N	Y
5.3 Trained staff	-	Y	-	Y	Y	Y	Y	N	Y	Y	Y
5.4 Financial	-	Y	-	Y	Y	Y	Y	N	Y	Y	Y

5.5 Other IT infrastructure	Y	Y	Y	Y	-	Y	Y	N	Y	Y	Y
5.6 Data analysis	Y	Y	Y	Y	-	Y	-	N	Y	Y	Y
6. Are data easily accessible?	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y

Notes: Bahamas and Trinidad and Tobago are covered by the same CABI centre.
Y = Yes; N = No; - = no stated view.

Capacity building

Coordinated and targeted capacity-building activities are crucial and could include:

(a) on-site and/or virtual courses;

(b) case studies, exchange of information and experiences, and sharing of lessons learned;

(c) joint scientific research, technology transfer, scientific visits, partnerships and collaborations, including through regional networks;

(d) support for the development of scientific infrastructure, including through regional approaches (e.g. CGIAR centres);

(e) intercultural dialogue through face-to-face meetings for Indigenous Peoples and local communities using culturally appropriate tools and methodologies in indigenous languages, which could include dialogue between scientists and holders of traditional knowledge;

(f) integration in academic curricula; and

(g) integration in regional and international development agendas.

6. Access and benefit sharing for DSI

Bagley *et al.* in 2020 explored how domestic measures address benefit-sharing arising from commercial and non-commercial use of DSI. The study identified 16 countries and one subnational jurisdiction as having domestic ABS measures addressing DSI. The study reported that 18 countries indicated that ABS measures on DSI were under preparation. Most domestic ABS measures allow for the inclusion of specific obligations, including benefit-sharing obligations relating to DSI, as part of mutually agreed terms (MAT) for genetic resources.

The study found that countries take different approaches in their domestic measures for addressing DSI. While some consider DSI only in the context of the utilization of the tangible resource, others require prior informed consent (PIC) and MAT for DSI. Others do not require PIC for access to DSI for research and development but require the sharing of benefits derived from DSI that has been generated from their GR. A fourth group of countries has taken a conscious decision not to address DSI in their ABS measures, and a fifth group addresses DSI in some other way (Bagley *et al.*, 2020).

6.1 Benefit-sharing practices

Even where there are no legal obligations to share the benefits resulting from DSI, there may be voluntary benefit-sharing arrangements involving DSI. There are cases where the parties involved did not explicitly address DSI and benefit-sharing but in fact collaborated on DSI and shared benefits, even if these benefits may not have been monetary. This has been the case for CABI United Kingdom Centre projects, where the benefit-sharing arising from 116 projects active in 2019 that involved access to GR was analysed by Smith *et al*. (2021) The majority of these CABI projects were carried out jointly with partners in the provider countries and often with joint funding. However, fewer than 20 were in countries with Nagoya legislation that required compliance with ABS, and most of the projects were outside the scope of Nagoya legislation. CABI shares benefits with all its partner countries, including in the following ways: sharing results; collaboration in education, training and research; joint authorship of publications; joint ownership of intellectual property rights; and provision of access to CABI facilities and databases. The projects also invariably result in knowledge and technology transfer and in institutional capacity development to help build or maintain local collections. For instance, the projects include discovery, formulation and application of BCAs such as *Metarhizium* and *Beauveria* spp. to tackle resistance to pesticides used against insect crop pests in Brazil. The biopesticide product will be owned by the Brazilian partners, who also benefit from CABI's know-how and technology. Another biological control project involves a partnership with India to address an invasive weed problem caused by Himalayan balsam in the United Kingdom. An Indian strain of the rust fungus *Puccinia komarovii* var. *glanduliferae* was approved for release in

England and Wales in 2014 (Ellison, Pollard and Varia, 2020). A second strain, from Pakistan, was released in 2017 to control a different cohort of Himalayan balsam. However, there are several weed genotypes in the British Isles that are not susceptible to either rust strain, and further collaborative surveys with scientists from the India are now taking place. Benefits shared to date include joint papers and training activities with Indian and Pakistani collaborators, plus collaborative payments. Training was also provided for an MSc student from Kenya and a PhD student from Malaysia. The project has also been extended to include control of Himalayan balsam in Canada.

CABI shares benefits with provider countries regardless of their status under the Nagoya Protocol.[39] The 27 projects covered in the CABI United Kingdom and Nagoya Protocol benefit-sharing report (Smith *et al.*, 2021) involved the use of genetic resources, all characterized with DSI, from 22 countries. Of these, nine countries are parties to Nagoya Protocol with implementing law, eight are parties to the Nagoya Protocol with no implementing law (as indicated as of 8 July 2021 on the Access and Benefit Sharing Clearing House), and five are not parties to the Nagoya Protocol. Yet non-monetary benefits were shared with all the countries.

6.2 Examples of triggered benefit sharing

While most countries addressing DSI expect monetary benefit-sharing arising from its use, to date no country has reported receiving such benefits (Bagley *et al.*, 2020). Those countries that omit DSI from domestic benefit-sharing measures because they consider it out of the scope of the CBD and the Nagoya Protocol nonetheless facilitate scientific advancement through open access to DSI and regard it as a form of non-monetary benefit-sharing. The WiLDSI white paper (2020) offers five monetary benefit-sharing open-access policy options for DSI. The details are given in the white paper and include a microlevy, membership fee, cloud-based fees, commons licenses and blockchain metadata for open access to DSI. The idea is that open access does not equal "free of any obligations" and models can be implemented that make DSI "visible to all" but subject to conditions. The authors of the white paper conclude that "benefit-sharing is most likely to materialise when free exchange can happen" and that any system should avoid "attempts at monitoring/tracing/controlling this highly complex, dynamic ecosystem", as this would require huge investment and be unlikely to result in cost-effective benefit-sharing.

CABI has published a report on its ABS policy and practices in which it addresses DSI (Smith *et al.*, 2021) and argues that amendments to the Nagoya Protocol are not necessary with respect to DSI and that the issue should be treated at country level. It notes that DSI is akin to derivatives, naturally occurring biochemical compounds resulting from a cell's metabolism, and that it is clear that if it is accessed with the genetic resource on which it is based or generated DSI may be covered by MAT. It argues, however, that this would mean that each country would take its own position, potentially making international collaboration and usage difficult. It further argues that to avoid this problem, it would be beneficial to have a common agreement on the generation of DSI and how it can be used in a way that facilitates innovation in the life sciences.

Both monetary and non-monetary benefits are possible for DSI, as is the case for the genetic resources themselves, and the decision on which type of benefit is most appropriate might potentially be determined by the type of use.[40] For example, where generating and publishing sequence data produces descriptive information on the organism and is not utilization, this could trigger non-monetary benefits. These might include access to the data and elements of capacity building, as with the publishing of the sequence as electronic data. DSI can be used at many non-exploitative levels: for example, its use to confirm organism identification is an observation rather than research; in most cases the resulting sequence data are published in public databases. There could be similar non-monetary arrangements that could revolve around uses delivering public good, such as addressing the Sustainable Development Goals (SDGs). However, if DSI is used for financial benefit then this should

[39] https://www.cabi.org/wp-content/uploads/PDFs/AboutCABI/Cabi-Abs-Policy-Draft-For-Website-May2018.pdf

[40] This paragraph draws on Smith, D., Ryan, M.J., Luke, B., Djeddour, D., Seier, M.K., Varia, S., Pollard, K.M. *et al.* 2021. *CABI UK and Nagoya Protocol triggered benefit sharing*. CABI Working Paper 25. Egham, UK, CABI.

be considered utilization and could trigger monetary benefit-sharing. The full benefit-sharing arrangement could be negotiated with the provider country, as would be done for access to the organism itself, or managed in a similar way in a multilateral system. The latter could reduce transaction costs and provide a global system that might not require tracking and tracing the source of the DSI. Such use and its implications should be made clear in the terms and conditions for the use of public databases containing DSI. Currently, for some countries, the generation and use of DSI must be considered when negotiating access – i.e. be expressed in the MAT and presented in any material transfer agreement to make it clear what can and cannot be done regarding DSI (at least until clarification is given by the COP in guidance or regulation).

6.3 Resolving the common approach to DSI use and benefit-sharing

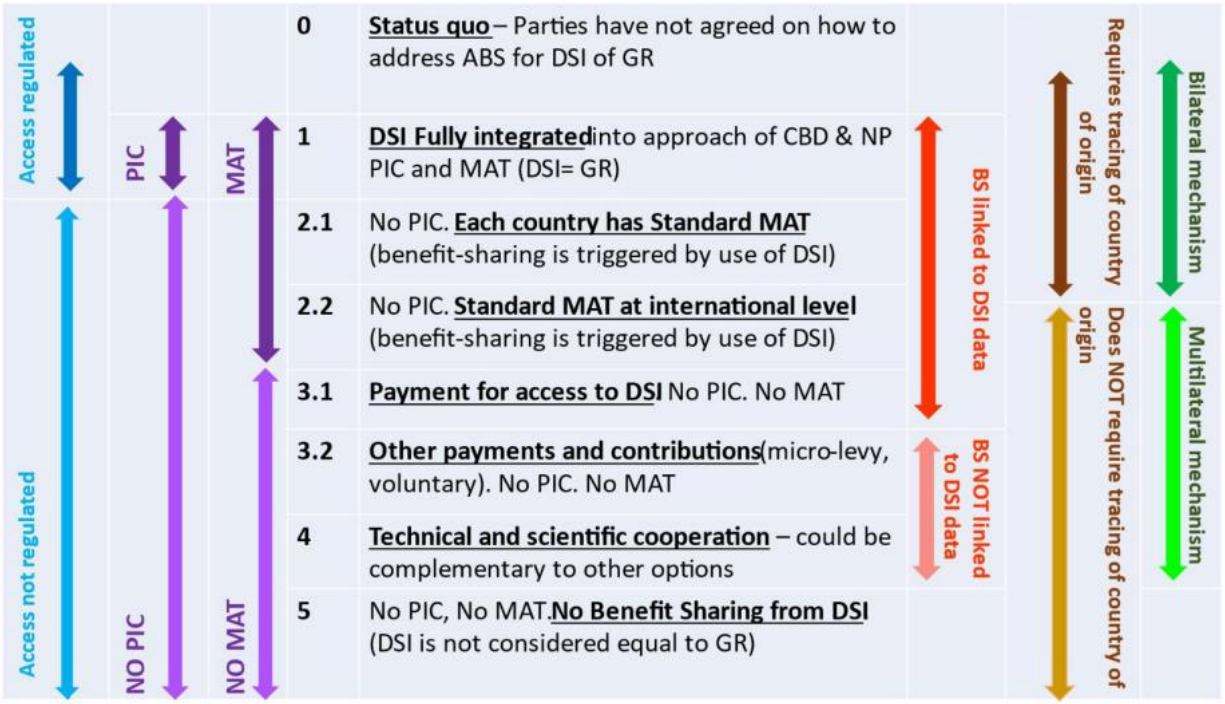
There are several different options on the table in the CBD discussions for a common approach to DSI and benefit-sharing (Figure 5). At the time of writing, these were still under discussion in the OEWG and the Informal Advisory Group (IAG) on DSI. It is clear that not all Parties agree, and that there is still some way to go. However, there is convergence on some issues, such as open access to DSI. The hope is that some resolution may emerge from the UN Biodiversity Conference (COP 15) in December 2022. These discussions will precede the Nineteenth Regular Session of the Commission on Genetic Resources for Food and Agriculture in 2023, which will be able to take their outcomes into consideration.

The monetary benefit-sharing options for DSI under discussion by the CBD are (a) a bilateral system similar to the Nagoya Protocol, where access is coupled to benefit-sharing, and (b) multilateral mechanisms with standardized rules for benefit-sharing. These policy options for ABS and DSI have been published by the Secretariat of the CBD (CBD, 2021). The options are not exhaustive, not mutually exclusive, and are presented without judgement as to their viability, cost efficiency, enforceability or capacity requirements. They do not address operational aspects of distributing benefits, and the traditional knowledge associated with GR is not covered.

The options include retaining the status quo, which allows each country to decide its approach, resulting in many different approaches, as described above (Bagley *et al.*, 2020; Smith *et al.*, 2021). Alternatively, DSI fully could be integrated into the CBD and the Nagoya Protocol, which would imply that access to DSI is subject to PIC and MAT. In this case, researchers would have to comply with national ABS requirements when accessing DSI through a database, possibly trace each element of DSI back to the country of origin of the genetic resource it came from, and negotiate ABS arrangements, possibly with multiple countries, each potentially having different ABS measures in place (CBD, 2021).

Another option is for access to DSI to continue to be open but for benefits to be shared once DSI has been utilized and benefits have been generated. Benefit-sharing would be triggered by milestones along the value chain, with unrestricted access, and benefit-sharing determined by standard MAT, license or terms and conditions. This would avoid individual negotiation of contracts for each DSI utilization, as they would be covered by an umbrella contract or agreement. This option would require downstream monitoring of DSI use for enforcement and monitoring. There are two suboptions, with MAT being dealt with either at national or international level. In the first instance, users could choose countries with favourable conditions. An internationally agreed MAT would avoid such competition among potential providers. This option (open access, and benefit-sharing based on standard – national or international – MAT) would ensure open access to DSI and facilitate science and the generation and public storage of DSI while providing for the sharing of benefits once DSI is being utilized. However, the benefit-sharing obligation provided under this option would still require some sort of compliance mechanism, which could make traceability necessary, leading to potentially high transaction costs.

Figure 5. High-level classification of policy options



Source: OEWG. 2021. *Digital sequence information on genetic resources. Note by the Executive Secretary*. Open-Ended Working Group (OEWG) on the Post-2020 Global Biodiversity Framework, third meeting (online), 23 August – 3 September 2021. CBD/WG2020/3/4. Montreal, Canada. https://www.cbd.int/doc/c/afd4/4df3/d2d62f5f6a1bfe367c7448f4/wg2020-03-04-en.pdf. See also OEWG. 2021. *Digital sequence information on genetic resources. Addendum: Note by the Executive Secretary*. Open-Ended Working Group (OEWG) on the Post-2020 Global Biodiversity Framework, third meeting (resumed), Geneva, Switzerland, 12–28 January 2022. CBD/WG2020/3/4/Add.1. Montreal, Canada. https://www.cbd.int/doc/c/1081/7ad0/05a4577d6c756e8d2f6cb22f/wg2020-03-04-add1-en.pdf

Another group of options currently under debate would require every user of DSI databases to make a payment or contribution into a multilateral fund. This option would avoid the need to trace the source of the genetic resources that the DSI is generated from or to monitor the use of products from the DSI. Access would not be regulated, and the payment would be decoupled from benefits (WiLDSI, 2020). Discussions on ABS for DSI may have progressed too far to consider, as an alternative to regulating DSI, a multilateral mechanism that would neither restrict access to DSI nor include specific benefit-sharing obligations for users of DSI. This would require a clear commitment by governments and relevant stakeholders, in particular governments of countries and stakeholders that play leading roles in the generation and use of DSI, to make a binding commitment to provide substantial technical and scientific support, and substantial funding, for capacity-building to enable countries and researchers in the Global South to fully benefit from DSI.

There would be various ways of delivering this, including research collaboration, training, knowledge and technology transfer, and technology co-development. While this would perhaps be the most beneficial and certainly the simplest option, it would require a clear commitment by governments and stakeholders, and mutual trust between the providers and the users of DSI and of the genetic resources from which the DSI has been generated.

Work ahead of COP 15 by the OEWG for the CBD has resulted in a co-leads' report on the work of the IAG on DSI on genetic resources.[41] Members of the IAG presented additional policy options, one on the African proposal for a multilateral mechanism, and a further two on hybrid approaches (involving both bilateral agreement between country and provider and a multilateral system). Of relevance to this study was the IAG's assessment of the proposed policy options using the criteria

[41] https://www.cbd.int/doc/c/383f/a3b3/589ff1552cd75c9841f08d33/wg2020-05-inf-01-en.pdf

from a matrix designed for this purpose.[42] The assessment showed that carrying out a full cost–benefit analysis is difficult because of the lack of detail on the mode of operation and data on the options and deficiencies in practical and technical knowledge. The aim of the assessment was to identify areas of convergence and divergence of the policy options shown in Figure 5, with some adjustments and additions for the sake of clarity, to be provided to COP 15 as background for their discussions.

Members of the IAG considered that Option 0 (status quo), Option 1 (DSI fully integrated) and Option 2.1 (no PIC, each country with standard MAT) should not be considered further. Option 2.2 (no PIC, standard MAT at international level) was not favoured by all, but some members thought it should be further considered as part of a hybrid solution. The IAG thought Option 3.1 (payment for access to DSI) should not be considered further. Option 3.2 (other payments for benefits and contributions) received mixed views, and further information was considered to be needed. Option 4 (technical and scientific cooperation) was the most favourably received, and it was suggested that it should be considered further, potentially in combination with another option or options. Option 5 (no PIC, no MAT, no benefit-sharing) had mixed reviews, as few of the assessment criteria were applicable, and most members of the IAG suggested that it should not be considered further, as it did not achieve the objective of benefit-sharing.

6.4 Addressing utilization for the public good

A number of interested parties have argued that activities that result in delivery of public goods, such as contributing to the SDGs, should be exempt. However, as mentioned earlier, rather than exemption, an alternative might be for activities only to generate non-monetary benefits, particularly with regard to reducing losses and improving yields from food and agricultural biodiversity. Identification of emerging plant and animal diseases and research on how to minimize their impact would be of high relevance here, and such an approach would be similar to the WHO position on emerging human diseases (WHO, 2017). Research and development of agents for classical biocontrol is a case in point: here benefits could include sharing the knowledge base and giving facilitated access to BCAs (Smith *et al.*, 2018). This might be achieved through an intergovernmental agreement (binding or non-binding) through which governments commit themselves, on the basis of reciprocity, not to limit access to, or use of, plant pests and BCAs within their territories. This fits well with Option 4 (technical and scientific cooperation). A country giving access to its genetic resources for such activities would benefit from solutions developed by other countries. This meets the intention of the CBD and the Nagoya Protocol and provides the appropriate level of benefit-sharing that countries are seeking (Smith *et al.*, 2018). Silvestri *et al*. (2019) concluded that it is important to raise awareness among policymakers of the key role that classical weed biocontrol could play in different sectors, and to persuade them to develop ABS legal frameworks tailored to this. Classical biological control studies almost exclusively result in the release of the BCA and may not result in a product being released onto the market, although the results, including formulation and the genetic resource itself, may be published. Clearly, where a biopesticide product is developed this could trigger monetary benefit-sharing via the relevant agreements, and this fits well with the IAG suggestion of combining Option 4 (technical and scientific cooperation) with other options.

7. Discussion and conclusions

The definition of DSI remains controversial. The scope of DSI can range from only covering DNA and RNA sequences (NSD) to also including protein sequences, metabolites and other macromolecules; it may also include associated information and traditional knowledge. This study does not take a position on this issue.

DSI is highly and increasingly relevant to R&D in all sectors of GRFA. Multiple reports and studies as well as literature surveys carried out for this study indicate that DSI is used extensively to identify, characterize and monitor GRFA. Furthermore, DSI on BFA is enabling crops and breeds to be enhanced, improving yields, and – by providing resistance to pests and diseases and tolerance to drought and heat – enhancing robustness against climate change. Additionally, DSI from biodiversity

[42] https://www.abs-biotrade.info/fileadmin/Downloads/EVENT%20REPORTS/2022/20220608_Webinar_DSI_MCA_Report.pdf

outside the food and agriculture sector contributes to and is used for research on GRFA. Identifying sequences and their properties utilizes biodiversity data from all sectors, within and beyond food and agriculture, and often genes from organisms considered to be outside food and agriculture are used to improve agricultural productivity and disease resistance. The CABI Abstracts database survey found most hits for technologies that generate NSD (86 655), although work on proteins and epigenetic modifications gave only slightly fewer hits (81 528). However, hits for the metabolome were significantly fewer (6 208).

It is clear from our study that the scope of DSI is complex, and as more information is gathered the complexity increases further and discussions on potential solutions expand. Scientifically, DSI represents a continuum: scientists may use both genomic sequence and systems biology data from metabolomics and proteomics to explore the context for their research questions.

In science, the primary benefit that is generated and shared is information, which in turn enables discoveries and allows public goods to be generated, as discussed in this report and by others in the working on GRFA. To this end, DSI is deposited freely, openly and transparently into public databases by its generators, also meeting the requirement of scientific journals that DSI (e.g. DNA and RNA sequences) be deposited with a unique identifier that is quoted in the publication to enable tracking. When a sequence is generated, it has little value until it is compared with other known reference sequences held in databases. This allows the source genetic resource to be identified and gives some indication of its characters and properties.

The generation of DSI should be encouraged for the benefit of science and discovery. A system that only recognized the source country of the genetic resources for benefit-sharing may disincentivise DSI producers, who may not reside in the country from which the genetic resources originate. It is proposed that access to DSI be “decoupled” from benefit-sharing: this could be done by establishing mechanisms that do not limit access to DSI but enable countries to receive appropriate benefits from a possible global fund. However, it is very clear that the ability to generate, manage and utilize DSI is not shared equally across the world. LMICs such as the CABI member countries the Bahamas, Ghana, Kenya, Malaysia, Trinidad and Tobago, and Zambia, often do not have the requisite facilities and capacity. DSI from these countries are often generated in one of the following ways: by foreign scientists who access the genetic resources; through collaboration between local scientists and partners in other countries that have access to facilities; or by buying in the necessary services. Feedback via the CABI centre in China also indicated that not all scientists in China have equal access to DSI and ability to utilize it. Action is needed to resolve these inequalities. Countries, and even industry, might help with this through multilateral processes. However, levying fees or royalties from products incorporating DSI is unlikely to generate sufficient funding. This issue needs to be explored, and appropriate measures put in place to maximize output from science that generates DSI for the public good but equally to find mechanisms and resources to fund the delivery of the CBD objectives of conservation, sustainable use and the fair and equitable sharing of benefits that arise from use.

Regarding current discussions within the CBD to find a common approach to DSI with respect to fair and equitable benefit-sharing, a number of requirements need to be addressed. First, any common approach must provide certainty and legal clarity for providers and users of DSI so as not to hinder the research and innovation that improves our ability to feed the global population and meet the SDGs. It must retain open access to data, available through a single gateway, to enable ready comparison of new NSD with known sequences. The system must be compatible with international legal obligations and recognize that the monetary and non-monetary benefits arising from the use of DSI should be used to support the conservation and sustainable use of biodiversity. The continuing discussions require further information and analysis of the options, including assessment of the potential consequences of different policy approaches. Areas to be explored include options or modalities for benefit-sharing, legal feasibility in implementation, and options for addressing the challenges of tracking and tracing. Countries already implementing their own approaches need ways to integrate them into a global system, such as hybrid approaches. A global system should consider a multilateral benefit-sharing mechanism for monetary benefits.

The key messages that arise from this study are:

1. There are many different existing and potential applications of DSI that are highly relevant to GRFA, including applications of DSI that is not itself derived from GRFA.
2. The current and potential applications of DSI show that its generation, storage, accessibility and use are fundamental to the characterization BFA and important to efforts to make agriculture more sustainable.
3. Access to and use of DSI face serious obstacles in many countries. There is an urgent need to address the root causes of these problems, which include lack of technical infrastructure, financial and human resources, educational and training opportunities, scientific collaboration, computing infrastructure, reliable electricity and high-speed internet, and may in the future possibly include prohibitive charges for database use.
4. There is a need for a regulatory environment that facilitates access to DSI and the fair and equitable sharing of benefits arising from its use.

Acknowledgements

Several individuals have helped review this document and to ensure the accuracy and adequate coverage of the content. These include the following:

Amber H. Scholz, DSI Network, WILDSI Project, Deputy to the Director, Leibniz-Institut DSMZ German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany;

Christopher Lyal, Scientific Associate, Natural History Museum, London, United Kingdom;

Dan Leskien, Senior Liaison Officer, Commission on Genetic Resources for Food and Agriculture, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy;

Devin Arbuthnott, Policy Advisor, Agriculture and Agri-Food Canada (AAFC), Ottawa, Canada;

Emmanuel Hala Kwon-Ndung, African BioGenome Project (AfricaBP), Professor of Plant Breeding and Genetics, Federal University of Lafia (FULafia), Nasarawa State, Nigeria;

Guy Cochrane, Data Coordination and Archiving Team Leader, Head of European Nucleotide Archive, European Bioinformatics Institute, European Molecular Biology Laboratory, Cambridge, United Kingdom;

Irene Hoffmann, Secretary, Commission on Genetic Resources for Food and Agriculture, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy;

Justin Eze Ideozu, Co-Chair, Ethics Legal and Social Issues Subcommittee, African BioGenome Project (AfricaBP), Senior Scientist Pharmacogenomics, Genomic Medicine, Genetic Research Center, North Chicago, United States of America;

Manuela da Silva, General Manager of Fiocruz COVID-19 Biobank, Fiocruz, Brazil;

Peter Mason, Research Scientist, Agriculture and Agri-Food Canada (AAFC), Ottawa, Canada;

Sally Mueni Katee, Chair, Ethics, Legal and Social Issues Subcommittee, Africa BioGenome Project (AfricaBP), ABS Legal Specialist/Officer, Livestock Genetics Program, International Livestock Research Institute, Nairobi, Kenya;

ThankGod Echezona Ebenezer, Founder and Co-Chair, African BioGenome Project (AfricaBP), Bioinformatician, EMBL's European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, United Kingdom.

CABI centre[43] contributions specifically coordinating, seeking and compiling country feedback.

Brazil, Centre: coordinated and compiled by Yelitza Colmenarez, Country Director Brazil.

China Centre: Feng Zhang coordinated the survey, Min Wan, Hongmei Li and Jinping Zhang conducted the interviews with Chinese experts. Xin translated the feedback and compiled relevant literature.

Ghana Centre: coordinated and compiled by Victor Clottey, Regional Representative, West Africa.

India Centre: coordinated and compiled by Gopi Ramasamy, Regional Director South Asia

Kenya Centre: coordinated and compiled by Joseph Mulema, Senior Scientist, Research.

Malaysia Centre: coordinated and compiled by Sathis Sri Thanarajoo, Scientist.

Pakistan Centre: coordinated Babar Bajwa, Senior Regional Director, Asia, and compiled by Yusuf Zafar.

[43] https://www.cabi.org/what-we-do/cabi-centres/

Trinidad and Tobago Centre contribution for the Bahamas and the Caribbean: coordinated by Naitram (Bob) Ramnanan.
Zambia Office: coordinated and compiled by Noah Phiri.

Acronyms and abbreviations

AHTEG – Ad Hoc Technical Expert Group
ABS – access and benefit-sharing
BAHFSA – Bahamas Agricultural Health and Food Safety Agency
BCA – biological control agent
BFA – biodiversity for food and agriculture
BOLD – Barcode of Life Data System
BRICS – Brazil, Russian Federation, India, China and South Africa
CARDI – Caribbean Agricultural Research and Development Institute
CBD – Convention on Biological Diversity
CABI – CAB International
CGIAR – Consultative Group on International Agricultural Research
CNV – copy number variants
COP – Conference of the Parties
CRI – Crops Research Institute
CRISPR – clustered regularly interspaced short palindromic repeats
CRISPR-Cas – CRISPR assisted protein
CSIR – Council for Scientific and Industrial Research
DDBJ – DNA Data Bank Japan
DMP – data management plan
DNA – deoxyribonucleic acid
DSI – digital sequence information
DSD – digital sequence data
EBI – European Bioinformatics Institute
EMBL – European Molecular Biology Laboratory
ENA – European Nucleotide Archive
FAIR – findability, accessibility, interoperability and reusability
FAO – Food and Agriculture Organization of the United Nations
G77 – Group of 77 (lower-income countries) at the United Nations
GBA– Global Biofoundry Alliance
GBC – Global Biodata Coalition
GBS – genotyping by sequencing
GI – genetic information
GRFA – genetic resources for food and agriculture
GRSD – genetic resource sequence data
GSD – genetic sequence data
GWAS – genome-wide association study
H3Africa – Human Heredity and Health in Africa
H3ABioNet – Pan African Bioinformatics Network for the Human Heredity and Health in Africa (H3Africa) consortium
HRMS – high resolution mass spectrometry
INSDC – International Nucleotide Sequence Database Collaboration
iBOL – International Barcode of Life
IAG – Informal Co-Chairs' Advisory Group
KUE – potassium use efficiency
LMICs – low- and middle-income countries
MAT – mutually agreed terms
mRNA – messenger RNA
MS/MS – tandem mass spectrometry
MTA – marker trait association

NAR – *Nucleic Acids Research* (journal)
NASD – nucleotide and amino acid sequence data
NASSI – nucleotide and amino acid sequence and structural information
NASSFI – nucleotide and amino acid sequence, structural and functional information
NBA – National Biodiversity Authority
NCBI – National Center for Biotechnology Information
NFP – national focal point
NGDC – National Genomics Data Centre
NGS – next-generation sequencing
NIH – National Institutes of Health
NSD – nucleotide sequence data
NSI – nucleotide sequence information
OECD – Organisation for Economic Co-operation and Development
OEWG – Open-ended Working Group
PCR – polymerase chain reaction
PGRI – Plant Genetics Research Institute
PIC – prior informed consent
PubMed – Public/Publisher MEDLINE
R&D – research and development
RNA – ribonucleic acid
SDG – Sustainable Development Goal
SNP – single nucleotide polymorphism
SPARC – Scholarly Publishing and Academic Resources Coalition
UWI – University of the West Indies
WDCM – World Data Centre for Microorganisms
WGS – whole genome sequence
WHO – World Health Organization
WiLDSI – "Wissenschaftsbasierte Lösungsansätze für Digitale Sequenzinformation" – Scientific approaches for digital sequence information
WIPO – World Intellectual Property Organization

References

AHTEG (Ad Hoc Technical Expert Group). 2020. *Report of the Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources, Montreal, Canada, 17–20 March 2020*. CBD/DSI/AHTEG/2020/1/7. Montreal, Canada. https://www.cbd.int/doc/c/ba60/7272/3260b5e396821d42bc21035a/dsi-ahteg-2020-01-07-en.pdf

Ambavaram, M.M.R., Ali, A., Ryan, K.P., Peoples, O., Snell, K.D. & Somleva, M.N. 2018. Novel transcription factors PvBMY1 and PvBMY3 increase biomass yield in greenhouse-grown switchgrass (*Panicum virgatum* L.). *Plant Science*, 273: 100–109. https://doi.org/10.1016/j.plantsci.2018.04.003

Antonelli, A., Fry, C., Smith, R.J., Simmonds, M.S.J., Kersey, P.J., Pritchard, H.W., Abbo, M.S. *et al.* 2020. *State of the world's plants and fungi*. Kew, UK, Royal Botanic Gardens. https://doi.org/10.34885/172

Arita, M., Karsch-Mizrachi, I. & Cochrane, G. 2021. The international nucleotide sequence database collaboration. *Nucleic Acids Research*, 49(D1): D121–D124. https://doi.org/10.1093/nar/gkaa967

Bagley, M., Karger, E., Ruiz Muller, M., Perron-Welch, F. & Thambisetty, S. 2020. *Fact-finding study on how domestic measures address benefit-sharing arising from commercial and non-commercial use of digital sequence information on genetic resources and address the use of digital sequence information on genetic resources for research and development. Convention on Biological Diversity, Montreal*. Annex. Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources, Montreal, Canada, 17-20 March 2020. CBD/DSI/AHTEG/2020/1/5. Montreal Canada, Secretariat of the Convention on Biological

Diversity. https://www.cbd.int/doc/c/428d/017b/1b0c60b47af50c81a1a34d52/dsi-ahteg-2020-01-05-en.pdf

Beagrie, N. & Houghton, J. 2021. *Data-driven discovery: The value and impact of EMBL-EBI managed data resources*. Salisbury, UK, Charles Beagrie, Ltd. https://www.embl.org/documents/wp-content/uploads/2021/10/EMBL-EBI-impact-report-2021.pdf

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. 2011. GenBank. *Nucleic Acids Research,* 39(Database issue): D32-7. https://doi.org/10.1093/nar/gkq1079.

Blackwell, M. 2011. The Fungi: 1, 2, 3 … 5.1 million species? *American Journal of Botany*, 98(3): 426–438. https://doi.org/10.3732/ajb.1000298

Bonanomi, G., De Filippis, F., Zotti, M., Idbella, M., Cesarano, G., Al-Rowaily, S. & Abd-ElGawad, A. 2020. Repeated applications of organic amendments promote beneficial microbiota, improve soil fertility and increase crop yield. *Applied Soil Ecology,* 156: 103714. https://doi.org/10.1016/j.apsoil.2020.103714

Calderini, D.F., Castillo, F.M., Arenas-M, A., Molero, G., Reynolds, M.P., Craze, M., Bowden, S. *et al*. 2021. Overcoming the trade-off between grain weight and number in wheat by the ectopic expression of expansin in developing seeds leads to increased yield potential. *New Phytologist,* 230(2): 629–640. https://doi.org/10.1111/nph.17048

Cantelli, G., Bateman, A., Brooksbank, C., Petrov, A.I., Malik-Sheriff, R.S., Ide-Smith, M., Hermjakob, H. *et al*. 2022. The European Bioinformatics Institute (EMBL-EBI) in 2021. *Nucleic Acids Research*, 50: D11–D19. https://doi.org/10.1093/nar/gkab1127

CBD (Secretariat of the Convention on Biological Diversity). 2021. *Policy options for access and benefit sharing and digital sequence information.* Summary of webinar held on line April 2021. Montreal, Canada. https://www.cbd.int/abs/*DSI*-webinar/DSIPolicyOptions2021.pdf

Chen, J.H., Chen, S.T., He, N.Y., Wang, Q.-L., Zhao, Y., Gao, W. & Guo, F.-Q. 2020. Nuclear-encoded synthesis of the D1 subunit of photosystem II increases photosynthetic efficiency and crop yield. *Nature Plants,* 6: 570–580. https://doi.org/10.1038/s41477-020-0629-z

Chibeba, A.M., Kyei-Boahen, S., de Fátima Guimarães, M., Nogueira, M.A. & Hungria, M. 2017. Isolation, characterization and selection of indigenous *Bradyrhizobium* strains with outstanding symbiotic performance to increase soybean yields in Mozambique. *Agriculture, Ecosystems & Environment,* 246: 291–305. https://doi.org/10.1016/j.agee.2017.06.017

CNCB-NGDC Members and Partners. 2022. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2022. *Nucleic Acids Research*, 50(D1): D27–D38. https://doi.org/10.1093/nar/gkab951

Cochrane, G. 2022. Genomic data for biodiversity – a global challenge**,** an exploration of where in the world genomics methods are applied and where the data are used. In: *EMBL-EBI*, Hinxton, UK. Cited 13 December 2022. https://www.ebi.ac.uk/about/news/perspectives/genomic-data-for-biodiversity-a-global-challenge/

Cowell, C., Paton, A., Borrell, J.S., Williams, C., Wilkin, P., Antonelli, A., Baker, W.J. *et al*. 2022. Uses and benefits of digital sequence information from plant genetic resources: Lessons learnt from botanical collections. *Plants People Planet,* 4(1): 33–43. https://doi.org/10.1002/ppp3.10216

de los Ríos-Pérez, L., Nguinkal, J.A., Verleih, M. Rebl, A., Brunner, R.M., Klosa, J., Schäfer, N., Stüeken, M., Goldammer, T., Wittenburg, D. 2020. An ultra-high density SNP-based linkage map for enhancing the pikeperch (*Sander lucioperca*) genome assembly to chromosome-scale. *Science Reports*, 10: 22335. https://doi.org/10.1038/s41598-020-79358-z

Ebenezer, T.E., Muigai, A.W.T., Nouala, S., Badaoui, B., Blaxter, M., Buddie, A.G., Jarvis, E.D. *et al*. 2022. Africa: sequence 100,000 species to safeguard biodiversity. *Nature,* 603(7901): 388–392. https://doi.org/10.1038/d41586-022-00712-4.

Edwards, D. & Batley, J. 2009. Plant genome sequencing: applications for crop improvement. *Plant Biotechnology Journal* 8(1): 2–9. https://doi.org/10.1111/j.1467-7652.2009.00459.x

El-Esawi, M.A. & Alayafi, A.A. 2019. Overexpression of rice Rab7 gene improves drought and heat tolerance and increases grain yield in rice (*Oryza sativa* L.). *Genes,* 10(1): 56. https://doi.org/10.3390/genes10010056

Ellison, C.A., Pollard, K.M. & Varia, S. 2020. Potential of a coevolved rust fungus for the management of Himalayan balsam in the British Isles: first field releases. *Weed Research,* 60(1): 37–49. https://doi.org/10.1111/wre.12403

Erenstein, O., Jaleta, M., Sonder, K., Mottaleb, K. & Prasanna, B.M. 2022. Global maize production, consumption and trade: trends and R&D implications. *Food Security,* https://doi.org/10.1007/s12571-022-01288-7

Faisal, A., Biswas, S., Zerin, T., Rahman, T. & Seraj, Z. 2017. Downregulation of the DST transcription factor using artificial microRNA to increase yield, salt and drought tolerance in rice. *American Journal of Plant Sciences,* 8(9): 2219–2237. https://doi.org/10.4236/ajps.2017.89149

FAO (Food and Agriculture Organization of the United Nations). 2021. *Digital sequence information on genetic resources for food and agriculture: innovation opportunities, challenges and implications*. Commission on Genetic Resources for Food and Agriculture Eighteenth Regular Session 27 September – 1 October 2021. CGRFA-18/21/5. Rome. https://www.fao.org/3/ng847en/ng847en.pdf

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J. *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* 269(5223): 496–512. DOI: 10.1126/science.754280

Fodor, N., Challinor, A., Droutsas, I., Ramirez-Villegas, J., Zabel, F., Koehler, A-K. & Foyer, C.H. 2017. Integrating plant science and crop modeling: Assessment of the impact of climate change on soybean and maize production. *Plant and Cell Physiology,* 58(11): 1833–1847. https://doi.org/10.1093/pcp/pcx141

Forin, N., Nigris, S., Voyron, S., Girlanda, M., Vizzini A., Casadoro, G. & Baldan, B. 2018 Next generation sequencing of ancient fungal specimens: the case of the Saccardo Mycological Herbarium. *Frontiers in Ecology and Evolution,* 6: 129. https://www.frontiersin.org/article/10.3389/fevo.2018.00129

Fukagawa, N.K. & Ziska, L.H. 2019. Rice: importance for global nutrition. *Journal of Nutritional Science and Vitaminology*, 65(Supplement): S2–S3. https://doi.org/10.3177/jnsv.65.S2

Gerard, G.S., Alqudah, A., Lohwasser, U., Börner, A. & Simón, M.R. 2019. Uncovering the genetic architecture of fruiting efficiency in bread wheat: a viable alternative to increase yield potential. *Crop Breeding & Genetics,* 59(5): 1853–1869. https://doi.org/10.2135/cropsci2018.10.0639

Guo, W., Chen, L., Chen, H., Yang, H., You, Q., Bao, A., Chen, S. *et al.* 2020. Overexpression of GmWRI1b in soybean stably improves plant architecture and associated yield parameters, and increases total seed oil production under field conditions. *Plant Biotechnology Journal,* 18(8): 1639–1641. https://doi.org/10.1111/pbi.13324

Gupta, S., Schillaci, M. & Roessner, U. 2022. Metabolomics as an emerging tool to study plant–microbe interactions. *Emerging Topics in Life Sciences,* 6(2): 175–183. https://doi.org/10.1042/ETLS20210262

Halewood, M., Lopez Noriega, I., Ellis, D., Roa, C., Rouard, M. & Hamilton, R.S. 2017. *Potential implications of the use of digital sequence information on genetic resources for the three objectives of the Convention on Biological Diversity. A submission from CGIAR to the Secretary of the Convention on Biological Diversity*. Rome, Bioversity International. https://cgspace.cgiar.org/handle/10568/92049

Hannah, L.C., Futch, B., Bing, J., Shaw, J.R., Boehlein, S., Stewart, J.D., Beiriger, R., Georgelis, N. & Greene, T. 2012 A shrunken-2 transgene increases maize yield by acting in maternal tissues to increase the frequency of seed development. *The Plant Cell,* 24(6): 2352–2363. https://doi.org/10.1105/tpc.112.100602

Hawksworth, D.L. & Lücking, R. 2017. Fungal diversity revisited: 2.2 to 3.8 million species. *Microbiology Spectrum,* 5(4): FUNK-0052-2016. https://doi.org/10.1128/microbiolspec.FUNK-0052-2016

Heinemann, J.A., Coray, D.S. & Thaler, D.S. 2018. *Exploratory fact-finding scoping study on "Digital Sequence Information" on genetic resources for food and agriculture*. Commission on Genetic Resources for Food and Agriculture: Background Study Paper No. 68. Rome. FAO. https://www.fao.org/3/CA2359EN/ca2359en.pdf

Hillson, N., Caddick, M., Cai, Y. Carrasco, J.A., Chang, M.W., Curach, N.C., Bell, D.J. *et al.* 2019. Building a global alliance of biofoundries. *Nature Communications*, 10: 2040. https://doi.org/10.1038/s41467-019-10079-2

Hotalinga, S., Kelleya, J.L. & Frandsen, P.B. 2021. Toward a genome sequence for every animal: Where are we now? *Proceedings of the National Academy of Sciences of the United States of America*, 118 (52): e2109019118. https://doi.org/10.1073/pnas.2109019118

Hurgobin, B. & Lewsey, M.G. 2022. Applications of cell- and tissue-specific 'omics to improve plant productivity. *Emerging Topics in Life Sciences*, 136(2): 163-173. https://doi.org/10.1042/ETLS20210286

Hurgobin, B. & Lewsey, M.G. 2022. How 'omics technologies can drive plant engineering, ecosystem surveillance, human and animal health. *Emerging Topics in Life Sciences*, 6(2): 137–139. https://doi.org/10.1042/ETLS20220020

Hutter, C. 2022. Genome-wide association studies. In: *National Human Genome Research Institute*. Bethesda USA. Cited 13 December 2022. https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies#:~:text=A%20genome%2Dwide%20association%20study,disease%20or%20a%20particular%20trait

Illumina. 2022. *Large-Scale Bull Genome Sequencing Enables Rapid Livestock Improvement*. Cited 22 December 2022. https://emea.illumina.com/science/customer-stories/icommunity-customer-interviews-case-studies/daetwyler-latrobe-interview-hiseq-1000bulls.html

Javal, M., Terblanche, J.S, Conlong, D.E., Delahaye, N., Grobbelaar, E., Benoit, L., Lopez-Vaamonde, C. & Haran J.M. 2021. DNA barcoding for bio-surveillance of emerging pests and species identification in Afrotropical Prioninae (Coleoptera, Cerambycidae). *Biodiversity Data Journal*, 9: e64499. https://doi.org/10.3897/BDJ.9.e64499

Kates, H.R., Doby, J.R., Siniscalchi, C.M., LaFrance, R., Soltis, D.E., Soltis, P.S., Guralnick, R.P. & Folk, R.A. 2021. The effects of herbarium specimen characteristics on short-read NGS sequencing success in nearly 8000 specimens: old, degraded samples have lower DNA yields but consistent sequencing success. *Frontiers in Plant Science*, 12: 669064. https://doi.org/10.3389/fpls.2021.669064

Kumar, M., Yusuf, M.A., Nigam, M. & Kumar, M. 2018. An update on genetic modification of chickpea for increased yield and stress tolerance. *Molecular Biotechnology*, 60: 651–663. https://doi.org/10.1007/s12033-018-0096-1

Land, M., Hauser, L., Jun, S.R. Nookaew, I., Leuze, M.R., Ahn, T-H, Karpinets, T. *et al.* 2015. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, 15(2): 141–161. https://doi.org/10.1007/s10142-015-0433-4

Lange, M., Alako, B.T.F., Cochrane, G., Ghaffar, M., Mascher, M., Habekost, P-K., Hillebrand, U. *et al*. 2021. Quantitative monitoring of nucleotide sequence data from genetic resources in context of their citation in the scientific literature. *GigaScience*, 10(12): giab084, https://doi.org/10.1093/gigascience/giab084

Lewin, H.A., Richards, S., Aiden, E.L., Archibald, J.M., Bálint, M., Barker, K.B., Baumgartner, B. *et al*. 2022. The Earth BioGenome Project 2020: Starting the clock. *Proceedings of the National Academy of Science U.S.A.*, 119(4): e2115635118. https://doi.org/10.1073/pnas.2115635118

Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J. Crandall, K.A., Durbin, R. *et al*. 2018. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, 115(17): 4325–4333. https://doi.org/10.1073/pnas.1720115115

Li, B. & Xue, D. 2019. Application of digital sequence information in biodiversity research and its potential impact on benefit sharing. *Journal of Biodiversity Science*, 27(12): 1379–1385. DOI: 10.17520/biods.2019242

Li, B., Liu, H., Zhang, Y., Kang, T., Zhang, L., Tong, J., Xiao, L. & Zhang, H. 2013. Constitutive expression of cell wall invertase genes increases grain yield and starch content in maize. *Plant Biotechnology Journal*, 11(9): 1080–1091. https://doi.org/10.1111/pbi.12102

Louca, S., Mazel, F., Doebeli, M. & Parfrey, L.W. 2019. A census-based estimate of earth's bacterial and archaeal diversity. *PLoS Biology,* 17(2): e3000106. https://doi.org/10.1371/journal.pbio.3000106.

Lyal, C.H.C. 2022. Digital sequence information on genetic resources and the convention on biological diversity. In: E. Chege Kamau, ed. *Global transformations in the use of biodiversity for research and development*. Ius Gentium: Comparative Perspectives on Law and Justice, 95, pp. 589–619. Cham, Switzerland, Springer.

Matthes, N., Pietsch, K., Rullmann, A. Näumann, G., Pöpping, B. & Szabo, K. 2020. The Barcoding Table of Animal Species (BaTAnS): a new tool to select appropriate methods for animal species identification using DNA barcoding. *Molecular Biology Reports,* 47: 6457–6461. https://doi.org/10.1007/s11033-020-05675-1

Mongabay. 2017a. *Scanning the barcode of wildlife.* Cited 23 December 2022. https://news.mongabay.com/2017/02/scanning-the-barcode-of-wildlife/

Mongabay. 2017a. *DNA barcoding helps identify endangered species from market specimens of sharks and rays*. Cited 23 December 2022. https://news.mongabay.com/2017/09/dna-barcoding-helps-identify-endangered-species-from-market-specimens/

Morand, S. 2018. Advances and challenges in barcoding of microbes, parasites, and their vectors and reservoirs. *Parasitology,* 145(5): 537–542. https://doi.org/10.1017/S0031182018000884

Muhich, A.J., Agosto-Ramos, A., & Kliebenstein, D.J. 2022. The ease and complexity of identifying and using specialized metabolites for crop engineering. *Emerging Topics in Life Sciences,* 6(2): 153–162. https://doi.org/10.1042/ETLS20210248

NIH (National Human Genome Research Institute). 2020. DNA Sequencing Fact Sheet. Cited 23 December 2022. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet

Nucleic Acids Research. 2017. Nucleic acid sequence, structure, and regulation. Nucleic Acids Research, 45:D1 (Database Issue). https://academic.oup.com/nar/issue/45/D1

OEWG (Open-Ended Working Group). 2021. *Digital sequence information on genetic resources. Addendum: Note by the Executive Secretary*. Open-Ended Working Group on the Post-2020 Global Biodiversity Framework, third meeting (resumed), Geneva, Switzerland, 12–28 January 2022. CBD/WG2020/3/4/Add.1. Montreal, Canada. https://www.cbd.int/doc/c/1081/7ad0/05a4577d6c756e8d2f6cb22f/wg2020-03-04-add1-en.pdf

OEWG. 2021. *Digital sequence information on genetic resources. Note by the Executive Secretary*. Open-Ended Working Group (OEWG) on the Post-2020 Global Biodiversity Framework, third meeting (online), 23 August – 3 September 2021. CBD/WG2020/3/4. Montreal, Canada. https://www.cbd.int/doc/c/afd4/4df3/d2d62f5f6a1bfe367c7448f4/wg2020-03-04-en.pdf.

OEWG. 2021. *Information from the Commission on Genetic Resources for Food and Agriculture related to digital sequence information on genetic resources*. Open-Ended Working Group on the Post-2020 Global Biodiversity Framework, third meeting (resumed), Geneva, Switzerland, 12–28 January 2022. CBD/WG2020/3/INF/9. Montreal, Canada. https://www.cbd.int/doc/c/986f/cb0e/07d17d0f56a7fac64bffc90f/wg2020-03-inf-09-en.pdf

Pickar-Oliver, A. & Gersbach, C.A. 2019. The next generation of CRISPR–Cas technologies and applications. *Nature Reviews Molecular Cell Biology,* 20: 490–507. https://www.nature.com/articles/s41580-019-0131-5

Pradhan, S., Babar, M.A., Bai, G., Khan, J., Shahi, D., Avci, M., Guo, J. *et al*. 2020. Genetic dissection of heat-responsive physiological traits to improve adaptation and increase yield potential in soft winter wheat. *BMC Genomics,* 21: 315. https://doi.org/10.1186/s12864-020-6717-7

Qin, H., Gu, Q., Zhang, J., Sun, L., Kuppu, S., Zhang, Y., Burow, M. *et al*. 2011. Regulated expression of an isopentenyltransferase gene (IPT) in peanut significantly improves drought tolerance and increases yield under field conditions. *Plant and Cell Physiology,* 52(11): 1904–1914. https://doi.org/10.1093/pcp/pcr125

Richardson, S.M., Mitchell, L.A., Stracquadanio, G., Yang, K., Dymond, J.S., Dicarlo, J.E., Lee, D. *et al*. 2017. Design of a synthetic yeast genome. *Science,* 355(53223): 1040–1044. https://www.science.org/doi/10.1126/science.aaf4557

Rigden, D.J. & Fernández, X.M. 2022. The 2022 Nucleic Acids Research database issue and the online molecular biology database collection, *Nucleic Acids Research*, 50: D1–D10. https://doi.org/10.1093/nar/gkab1195

Rodrigo, S., Hufton, A.L., Sett, S. & Scholz, A.H. 2022. *A technical assessment for the debate on benefit-sharing and digital sequence information*. Zenodo. DOI: 10.5281/zenodo.5849643.

Rohden, F., Huang, S., Dröge, G. & Scholz, A.H. 2020. *Combined study on digital sequence information in public and private databases and traceability.* Annex 1. Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources, Montreal, Canada, 17–20 March 2020. CBD/DSI/AHTEG/2020/1/4. Montreal, Canada. https://www.cbd.int/doc/c/1f8f/d793/57cb114ca40cb6468f479584/dsi-ahteg-2020-01-04-en.pdf

Rotter, A., Gaudêncio, S.P., Klun, K., Macher, J.-N., Thomas, O.P., Deniz, I., Edwards, C. *et al.* 2021. A new tool for faster construction of marine biotechnology collaborative networks. *Frontiers in Marine Science*, 8: 685164. https://doi.org/10.3389/fmars.2021.685164

Ruiz Muller, M. 2018. *Access to genetic resources and benefit sharing 25 years on: progress and challenges*. Issue Paper No. 44. Geneva, Switzerland, International Centre for Trade and Sustainable Development (ICTSD). https://www.voices4biojustice.org/wp-content/uploads/2018/12/Access-to-Genetic-Resources-and-Benefit-Sharing-25-Years-On-Progress-and-Challenges.pdf

Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R. *et al.* 2022b. Database resources of the National Center for Biotechnology Information, *Nucleic Acids Research,* 50: D20–D26. https://doi.org/10.1093/nar/gkab1112

Sayers, E.W., Cavanaugh, M., Clark, K., Pruitt, K.D., Schoch, C.L., Sherry, S.T. & Karsch-Mizrachi, I. 2022a. GenBank, *Nucleic Acids Research,* 50: D161–D164. https://doi.org/10.1093/nar/gkab1135

Schilling, R.K., Marschner, P., Shavrukov, Y., Berger, B., Tester, M., Roy, S.J. & Plett, D.C. 2013. Expression of the Arabidopsis vacuolar H+-pyrophosphatase gene (AVP1) improves the shoot biomass of transgenic barley and increases grain yield in a saline field. *Plant Biotechnology Journal*, 12(3): 378–386. https://doi.org/10.1111/pbi.12145

Schmidt, J., Garcia, M., Brien, C., Kalambettu, P., Garnett, T., Fleury, D. & Tricker, P.J. 2020. Transcripts of wheat at a target locus on chromosome 6B associated with increased yield, leaf mass and chlorophyll index under combined drought and heat stress. *PLoS ONE,* 15(11): e0241966. https://doi.org/10.1371/journal.pone.0241966

Scholz, A.H., Lange, M., Habekost, P., Oldham, P., Cancio, I., Cochrane, G. & Freitag, J. 2021. Myth-busting the provider-user relationship for digital sequence information, *GigaScience*, 10(12): giab085. https://doi.org/10.1093/gigascience/giab085

Shiferaw, B., Prasanna, B.M., Hellin, J. & Bänziger, M. 2011. Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. *Food Security,* 3: 307. https://doi.org/10.1007/s12571-011-0140-5

Si, T., Chao, R., Min, Y., Wu, Y., Ren, W. & Zhao. H. 2017. Automated multiplex genome-scale engineering in yeast. *Nature Communications,* 8: 15187. https://doi.org/10.1038/ncomms15187

Silvestri, L., Sosa, A., McKay, F., Diniz Vitorino, M., Hill, M., Zachariades, C. & Hight, S. 2019. The Nagoya Protocol and its implications for classical weed biological control. In: H.L. Hinz, M.-C. Bon, G. Bourdôt, M. Cristofaro, G. Desurmont, D. Kurose & H. Müller-Schärer, eds. *Proceedings of the XV International Symposium on Biological Control of Weeds, Engelberg, Switzerland,* pp. 304–309. https://bugwoodcloud.org/resource/files/15115.pdf

Simmons, C.R., Weers, B.P., Reimann, K.S., Abbitt, S.E., Frank, M.J., Wang, W., Wu, J., Shen, B. & Habben, J.E. 2020. Maize BIG GRAIN1 homolog overexpression increases maize grain yield. *Plant Biotechnology Journal,* 18(11): 2304–2315. https://doi.org/10.1111/pbi.13392

Singh, R. & Goodwin, S.B. 2022. Exploring the corn microbiome: a detailed review on current knowledge, techniques, and future directions. *PhytoFrontiers,* 3(2): 158–175. https://doi.org/10.1094/PHYTOFR-04-21-0026-RVW.

Skraly, F.A., Ambavaram, M.M.R., Peoples, O. & Snell, K.D. 2018. Metabolic engineering to increase crop yield: from concept to execution. *Plant Science*, 273: 23–32. https://doi.org/10.1016/j.plantsci.2018.03.011

Smith, D., Hinz, H., Mulema, J., Weyl, P. & Ryan, M.J. 2018. Biological control and the Nagoya Protocol on access and benefit sharing – a case of effective due diligence. *Biocontrol Science and Technology*, 28(10): 914–926. https://doi.org/10.1080/09583157.2018.1460317

Smith, D., Ryan, M.J., Luke, B., Djeddour, D., Seier, M.K., Varia, S., Pollard, K.M. *et al.* 2021. *CABI UK and Nagoya Protocol triggered benefit sharing*. CABI Working Paper 25. Egham, UK, CABI.

Sohn, H.B., Lee, H.Y., Seo, J.S., Jung, C., Jeon, J.H., Kim, J.H., Lee, Y.W., Lee, J.S., Cheong, J.J. & Choi, Y.D. 2011. Overexpression of jasmonic acid carboxyl methyltransferase increases tuber yield and size in transgenic potato. *Plant Biotechnology Reports*, 5: 27–34. https://doi.org/10.1007/s11816-010-0153-0

Sun, Y., Shang, L., Zhu, Q-H., Fan, L. & Guo, L. 2022. Twenty years of plant genome sequencing: achievements and challenges. *Trends in Plant Science*, 27(4): 391–401. https://doi.org/10.1016/j.tplants.2021.10.006

Supple, M.A. & Shapiro, B. 2018. Conservation of biodiversity in the genomics era. *Genome Biology*, 19: 131. https://doi.org/10.1186/s13059-018-1520-3

Vallarino, J.G., Kubiszewski-Jakubiak, S., Ruf, S., Rößner, M., Timm, S., Bauwe, H., Carrari, F. *et al.* 2020. Multi-gene metabolic engineering of tomato plants results in increased fruit yield up to 23%. *Science Reports*, 10: 17219. https://doi.org/10.1038/s41598-020-73709-6

Vogel, J.H., Muller, M.R., Angerer, K., Delgado-Gutiérrez, D. & Ballón, A.G. 2022. Bounded openness: a robust modality of access to genetic resources and the sharing of benefits. *Plants People Planet*, 4(1): 13–22. https://doi.org/10.1002/ppp3.10239

Whitfield, J. 2003. DNA barcodes catalogue animals. *Nature*. https://doi.org/10.1038/news030512-7

WHO (World Health Organization). 2017. *Comments by the World Health Organization on the draft factfinding and scoping study* "The emergence and growth of digital sequence information in research and development: implications for the conservation and sustainable use of biodiversity, and fair and equitable benefit sharing" dated 9 November 2017. Geneva, Switzerland. https://www.who.int/influenza/whocommentscbddsi.pdf

WiLDSI. 2020. *Finding compromise on ABS & DSI in the cbd: requirements & policy ideas from a scientific perspective*. WiLDSI White Paper. https://www.dsmz.de/fileadmin/user_upload/C

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, M., Axton, A., Baak, N., Blomberg, *et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Science Data*, 3: 160018. https://doi.org/10.1038/sdata.2016.18

Zhang, S., Peng, G. & Xia, Y. 2010. Microcycle conidiation and the conidial properties in the entomopathogenic fungus *Metarhizium acridum* on agar medium. *Biocontrol Science and Technology*, 20(8): 809–819. https://doi.org/10.1080/09583157.2010.482201

Zhang, Y., Andrews, H., Eglitis-Sexton, J., Godwin, I., Tanurdžić, M. & Crisp, P.A. 2022b. Epigenome guided crop improvement: current progress and future opportunities. *Emerging Topics in Life Sciences*, 6(2): 141–151. https://doi.org/10.1042/ETLS20210258

Zhang, Y., Li, W., Lu, P., Xu, T. & Pan, K. 2022a. Three preceding crops increased the yield of and inhibited clubroot disease in continuously monocropped Chinese cabbage by regulating the soil properties and rhizosphere microbial community. *Microorganisms*, 10(4): 799. https://doi.org/10.3390/microorganisms10040799

Zhang, Z., Wang, J., Wang, J., Wang, J. & Li, Y. 2020. Estimate of the sequenced proportion of the global prokaryotic genome. *Microbiome*, 8: 134. https://doi.org/10.1186/s40168-020-00903-z

Zhou, Y., Li, Y., Qi, X., Liu, R., Dong, J., Jing, W., Guo, M. *et al.* 2020. Overexpression of V-type H+ pyrophosphatase gene EdVP1 from *Elymus dahuricus* increases yield and potassium uptake of transgenic wheat under low potassium conditions. *Scientific Reports*, 10: 5020. https://doi.org/10.1038/s41598-020-62052-5

Zhou, Z, Tran, P.Q., Breister, A.M., Liu, Y., Kieft, K., Cowley, E.S., Karaoz, U. & Anantharaman, K. 2022. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* 10: 33. https://doi.org/10.1186/s40168-021-01213-8

Appendix 1. CAB Abstracts literature survey

CABI has been gathering data on agriculture for over 100 years, and much of this is presented via CAB Direct.[44] The CAB Abstracts bibliographic database is part of this resource and covers applied life sciences, including agriculture, plant sciences, animal sciences and related subjects. It includes over 10.9 million records dating from 1973 to the present (and an archive covering the period from 1912 to 1973). It is searchable on several platforms, including CAB Direct, which was used in this study. Our search strategy extracted the DSI-related studies from CAB Abstracts and grouped these into various categories, such as dominant crops in FAO regions, various types of terminology used to describe DSI, and examples of actual and potential applications of DSI in food and agriculture.

The search resulted in 1 180 915 hits, which represents the core dataset (DSI data pool) used for further analysis. The majority of the publications concerned plant genetic resources. The results demonstrate the extent and nature of the uses to which DSI is put to in the food and agriculture sector. The study presents trends in the growth and use of DSI and looks at the level of publications in each of the identified areas of use of DSI on GRFA. The searches enabled the selection of examples that demonstrate the impact and importance of DSI studies in the food and agriculture sector. The results of these searches showed that DSI on genetic material directly relevant to food and agriculture is of actual or potential value beyond the food and agriculture sector; and conversely, that much DSI originating from other sectors is of relevance to food and agriculture. Thus, DSI might also present scope-related issues.

Figure A1.1 shows the number of records in the database that cite DSI by publication year (2012 to 2022), reflecting the growth in DSI studies over the years, from the first citation in 1973 to 2022. The first 20 years, from 1973 to 1993, saw publications in the CAB Abstracts database rise to around 10 000 per year and the number rose eight-fold between 1993 and 2021, with almost 80 000 publications addressing DSI in the latter year. The total of 1 180 915 hits represents almost 11 percent of the papers in CAB Abstracts.

Figure A1.1 Number of publications in the CAB Abstracts database from 1973 to 2022 that address DSI

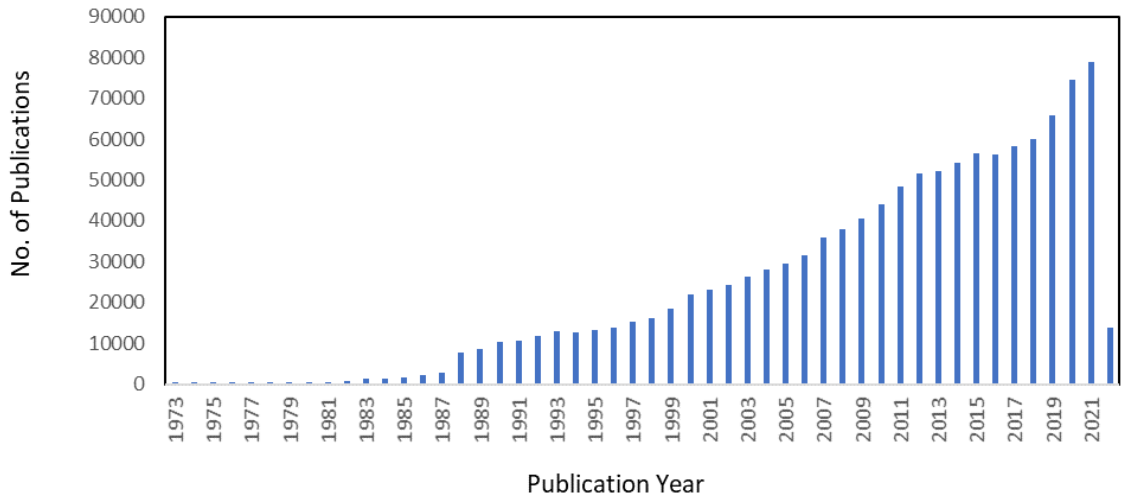


Table A.1.1 compares the FAO regions of the world for the number of publications on five crops (wheat, maize, rice, soybean and potato, where significant) that cite DSI as having impact. Asia and the Pacific have by far the greatest number of publications on DSI, followed by Europe and North America. The figures demonstrate the extent to which DSI is reaching into crop improvement and agriculture.

Table A1.1. number of hits for the top crops in the FAO regions

[44] https://www.cabdirect.org/

FAO Regions	Number of hits in DSI pool	Top crops [common name] (no. of hits)
Africa	26 981	*Triticum* [wheat] (967) *Zea mays* [maize] (743) *Oryza sativa* [rice] (313) *Glycine* [soybean] (266) *Solanum tuberosum* [potato] (245)
Asia and the Pacific	128 783	*Oryza sativa* [rice] (8 841) *Triticum* [wheat] (5 945) *Glycine* [soybean] (2 906) *Zea mays* [maize] (1 840) *Solanum tuberosum* [potato] (925)
Europe	53 811	*Triticum* [wheat] (2 179) *Zea mays* [maize] (1 193) *Solanum tuberosum* [potato] (730)
Latin America and the Caribbean	28 438	*Glycine* [soybean] (1 403) *Zea mays* [maize] (1 005) *Triticum* [wheat] (791) *Solanum tuberosum* [potato] (389)
Middle and Near East	22 166	*Triticum* [wheat] (1792) *Oryza sativa* [rice] (227) *Solanum tuberosum* [potato] (217)
North America	45 946	*Triticum* [wheat] (2 562) *Glycine* [soybean] (2 173) *Zea mays* [maize] (1 931) *Solanum tuberosum* [potato] (630)

Analysing the content of 1.18 million publications citing gene, protein and metabolite content in the CAB Abstracts database/Food Agriculture/ would be a huge task. Narrowing the analysis by including search parameters related to improved yield, gene technologies and cellular processes and to targeted crops such as wheat, rice and potato, and to livestock, reduced the number of hits considerably, as shown below. However, the task of reading thousands of papers to select examples of use meant that further targeting was needed. Including only records that had “yield increase” in the title, while also limiting the search to the last 11 years, resulted in 287 records. A review of these revealed how genomics and metabolomics have helped improve yields and resistance to disease, drought and increased temperature. Specifically, it was noted that genomics has provided a greater understanding of how crop plants function and how DSI can enable interventions in areas such as in photosynthesis

(40 papers), nitrogen metabolism (44 papers) and phosphate uptake (11 papers). Additionally, many of the papers were concerned with how DSI enables the exploration of different traits that could contribute to climate change adaptation, such as drought resistance (29 papers) and heat-stress resistance (50 papers). The types of BFA addressed in these papers were crops (129 papers) (e.g. wheat [40], rice [86], maize [30], potato [7], yam [5] and tomato [12]), livestock (13) (e.g. cattle [6]), bacteria (28) fungi (15) and viruses (8). Examples were selected to show how generation, analysis and use of DSI are contributing to the improvement of yields and to the future-proofing of food and agriculture with traits that increase drought and heat resistance and are hence contributing to climate change adaptation and helping to improve food security.

Comparing CABI data with other sources

The analysis focused on the validated and well-indexed CABI databases. The CABI team ruled out simple searches in Google to compare results but attempted to perform searches using the same search terms in PubMed and Google Scholar. Unfortunately, the string length for searches in Google Scholar did not allow the use of the terms used to compile the data pool on DSI for CAB Abstracts. Using a narrower set of terms for DSI in Google Scholar returned far fewer hits. The PubMed returns are presented in Figure A1.2.

Figure A1.2 Number of publications in PubMed from 1973 to 2022 that mention DSI

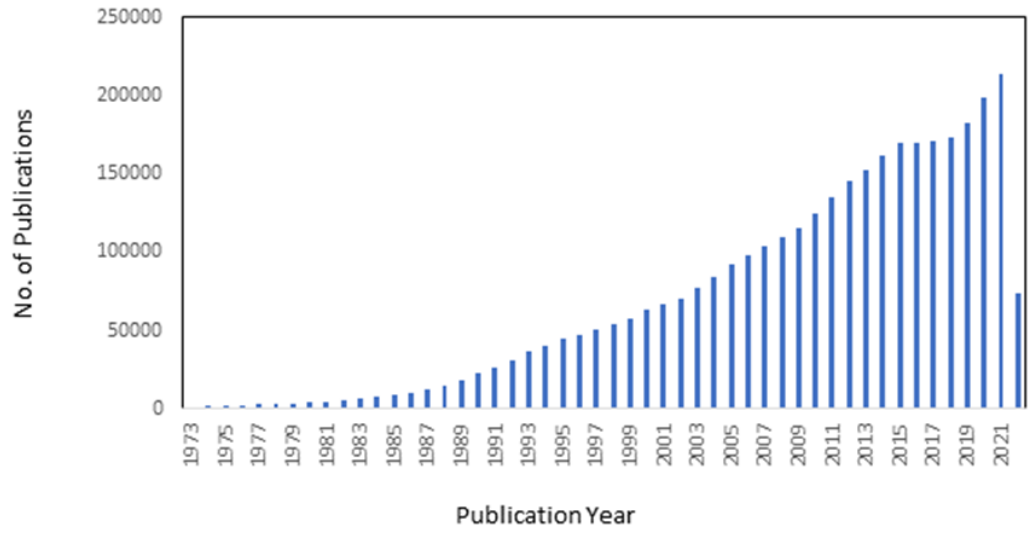


Figure A1.2 shows the percentage of total records over the years from 2012 to 2022 in CAB Abstracts and PubMed that cite DSI. The moving average trendlines illustrate the growth pattern of DSI-citing records. A steady growth is seen in the predominantly agriculture-focused CAB Abstracts from 14.2 percent of the records in 2012 to 19.7 percent of the records so far in 2022.

Figure A1.3 Percentage of publications citing DSI of total records added per year (2012 to 2022) in CAB Abstracts and PubMed

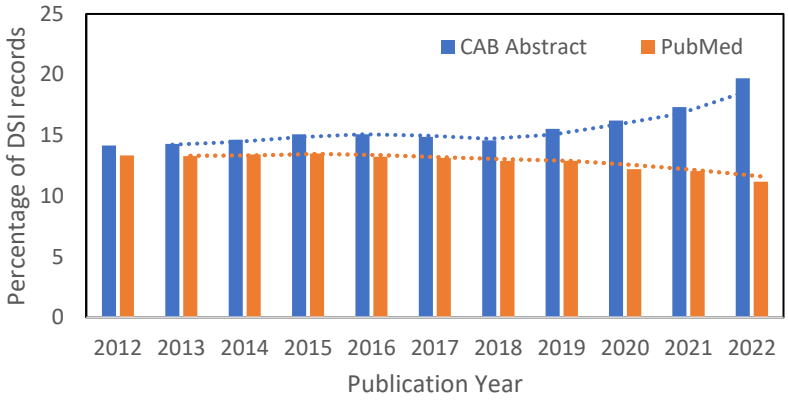

25
20
15
10
5
0
Percentage of DSI records
CAB Abstract
PubMed
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Publication Year

Figure A1.4 Distribution of publications related to DSI from different geographical locations[45]

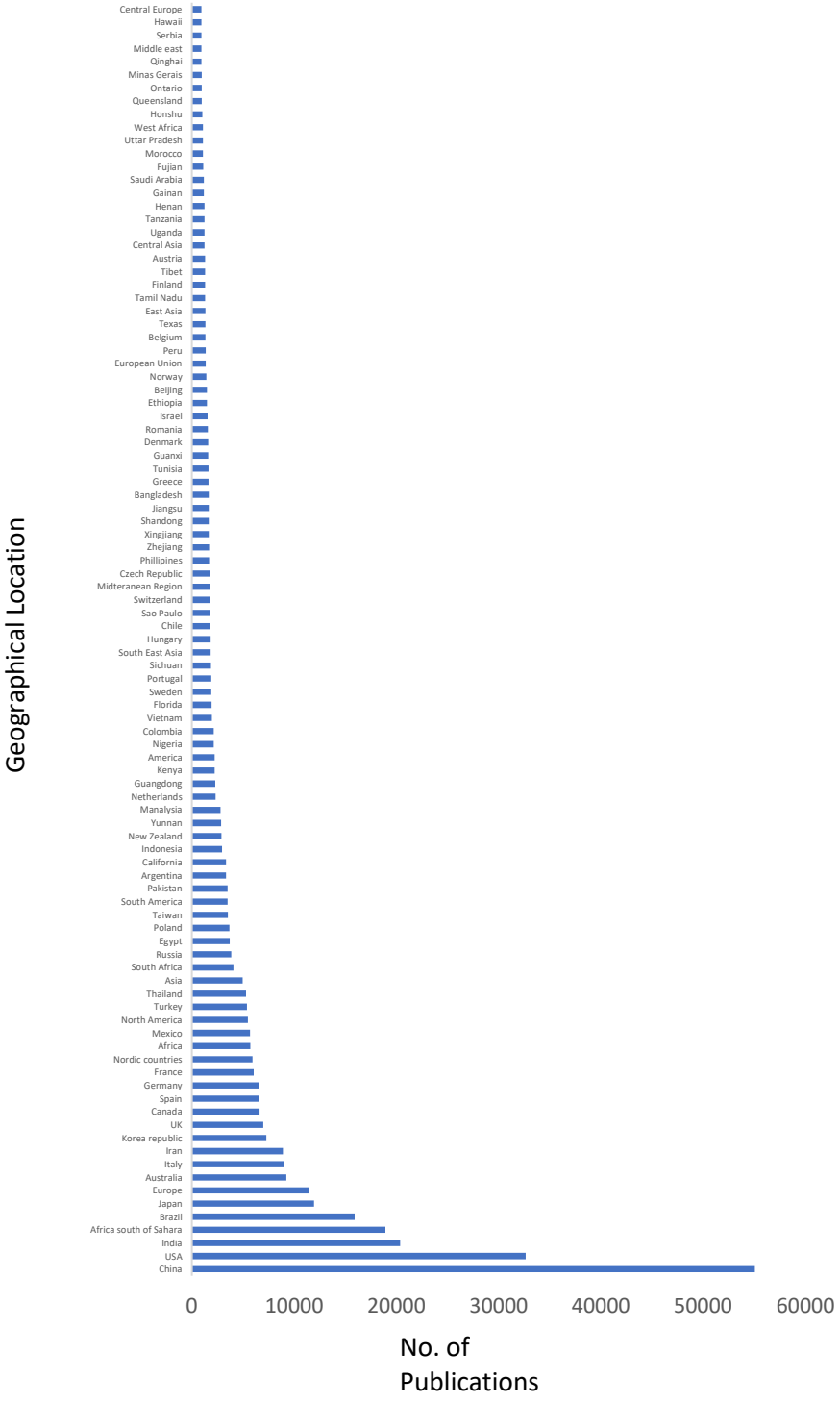


Figure A1.4 shows the distribution of publications related to DSI for 97 locations, with most publications coming from China (over 55 000), the United States of America (just under 35 000), India (approximately 20 000) and sub-Saharan-Africa (just under 20 000). As with the WiLDSI data, it is difficult to determine where the work was carried out, and again there were many North–South

[45] This table refers, for technical reasons, to some countries with inaccurate names. The accurate official names of these countries are: Czechia, Iran (Islamic Republic of), Republic of Korea, United Kingdom, Russian Federation, Taiwan Province of China, Republic of Tanzania, Türkiye, United States of America and Viet Nam.

collaborations. Although there are many publications from LMIC countries, there remains an imbalance in these countries' capacity to take full advantage of the DSI technology.

Appendix 2 CABI centre survey of obstacles to access and use of DSI

CABI centres[46] provided information on the generation and use of DSI in their host countries and regions, drawing both on their project work and on the experiences of the scientists at the centres and their partners. Where possible, the CABI centres contacted the national authorities of their host countries for input. The centres contributing are listed below and the information they provided is summarized in Section 5 of the main paper along with a table of commonalities. The full details are provided here (the contributors of this information are listed in the acknowledgements section of the main paper).

Bahamas

The Bahamas Agricultural Health and Food Safety Agency (BAHFSA) fully understands the importance of DSI to sustainable agriculture. One of BAHFSA's core objectives is to establish diagnostic laboratories with molecular capacity to access and utilize NSD to detect and identify pests and diseases. In the Bahamas, several obstacles, including a lack of human and financial resources to facilitate the establishment of the required technical infrastructure, prevent the widespread adoption and use of DSI. However, BAHFSA notes that scientific collaboration with local and regional research institutions could facilitate greater use of DSI.

Brazil

CABI's centre in São Paulo, Brazil operates across the whole of Latin America, providing scientific knowledge, information and expertise to the Latin American nations. CABI has projects that address agriculture in this region, investigating climate change, biodiversity and highly sensitive environments where a wide range of crops and livestock are farmed and traded, particularly coffee and cocoa. As in all regions, CABI works with a wide range of partners and has sourced information regarding the generation and use of DSI in their research in food and agriculture. At state level, a large number of public and private institutions generate and use DSI: EMBRAPA Cenargen is one of the key public institutions doing so. There are clear rules in Brazil that provide legal clarity to users of genetic resources and the DSI associated with them.[47,48] There is consensus in Brazil that legal measures that facilitate and foster research and development will generate more benefits, which can be channelled to biodiversity conservation and sustainable use, fulfilling the objectives of the international agreements on ABS.

DSI is generated locally, especially by research institutions, but it is difficult to estimate how much. It is generated mainly by researchers and postgraduate students in both public- and private-sector institutions. Brazilian genetic heritage can be freely accessed, but the results and products of its utilization are regulated by a registration or notification procedure. It is the national understanding that access, including through DSI, must be facilitated to generate the benefits that will fund biodiversity conservation and sustainable use.

DSI is easy to access for research institutions and researchers in Brazil, following the SisGen requirements.[49] A facilitated mechanism exists for access to genetic resources, with a focus on control of the economic exploitation of products or reproductive materials arising from access. An online registration system for tracing, tracking and overseeing access to genetic resources and associated traditional knowledge activities is in place, and the SisGen electronic system facilitates a procedure for the use of DSI within the framework of the CBD.

Despite all this being in place, there are problems that restrict the use of DSI. For instance, access to proper infrastructure and resources varies from one place to another. The north of Brazil has limited resources, thus limiting the full use of DSI. Training, budget and resource limitations also constrain the development and maintenance of information databases.

[46] https://www.cabi.org/what-we-do/cabi-centres/
[47] https://www.cbd.int/abs/DSI-views/2019/Brazil-DSI.pdf
[48] https://www.publicacoes.uniceub.br/rdi/article/viewFile/8079/pdf
[49] https://sisgen.gov.br/paginas/InstallSolution.aspx

Caribbean

CABI's office in Trinidad and Tobago[50] works with local partners in a region rich in natural resources, particularly commodity crops, which remain economically important for the area. The centre works across the whole of the Caribbean and Central America, carrying out work that is significant not just to the region, but globally. Projects have a focus on finding sustainable ways to manage crop pests and invasive species, conserve or enhance biodiversity, and support the commodity chains that flow from farmer to consumer. In executing and supporting projects in the region,[51] the centre works with several partners, some of whom provided feedback on the generation and use of DSI.

The Cocoa Research Centre under the Faculty of Science and Agriculture (FSA) of the University of the West Indies (UWI) was established in the 1930s, with a mandate to conserve, characterize, evaluate, utilize and distribute material from its internationally recognized germplasm collection (International Cocoa Germplasm, Trinidad). Research activities include germplasm conservation, morphological and molecular characterization of cacao accessions, screening of germplasm for resistance to diseases, germplasm enhancement (prebreeding for desirable traits), and quality and flavour assessment.[52] DSI generated is stored in the International Cocoa Germplasm Database, maintained by the University of Reading in the United Kingdom. The DSI in this database is freely accessible (Gillian Bidaisee, Genebank Characterization, personal communication, 2022).

In the late 1990s, the genetic basis for resistance to bacterial blight disease in anthurium was investigated. Most of the 60 anthurium cultivars grown in the Caribbean have been genotyped and their level of resistance to bacterial blight determined at the foliar and systemic levels using a two-stage screening method. The screening method developed and the understanding of the genetics of resistance allow targeted hybridizations between anthurium genotypes to obtain higher levels of resistance. The principles are being put into practice in the breeding programme at Kairi Cut-flowers Ltd. A number of novel varieties that combine resistance to bacterial blight with other horticultural characteristics have been developed. The Caribbean Agricultural Research and Development Institute (CARDI) has capacity but no equipment and needs to upgrade its facilities to house PCR and related equipment that can be used to genetically sequence Caribbean agrobiodiversity (personal communication from Fayaz Shaw, 2022).

The UWI is working on crosses and selection in minor crops, such as UWI F7 field maize and pigeon pea, but no materials have yet been sequenced. The crop pathology team uses part or whole genome sequencing and data processing of microorganisms, including soil microbes and plant pathogens. They note that interest in and use of DSI has increased in recent times at UWI, St. Augustine. The interest at St. Augustine is mostly related to ITS, 16SrNA gene sequencing, whole genome sequencing of microbes and genome-wide association studies (GWAS) in crops such as cocoa. However, they lack a database or centre for storing and processing their DSI data. DSI is used at UWI for reference purposes and to study similarity. However, limitations to technical capacity, computing infrastructure and logistics restrict its use. The situation is possibly the same across the Caribbean in all the few cases where DSI data are generated or used. Crop pathologists usually submit their sequences to databases such as NCBI and access DSI data from open databases.

The main conclusion for Trinidad and Tobago to date is that capacity and actual work on DSI is limited to the sequencing of the Cocoa Germplasm Collection, which is ongoing in collaboration with Reading University, United Kingdom. In addition, some work was done in the 1990s on about 60 varieties of anthuriums found in the Caribbean to investigate resistance to fungal and bacterial diseases.

China

The CABI Chinese centre consulted Chinese experts working in research fields related to DSI and reviewed relevant Chinese literature. This indicated that the generation and use of DSI is common in

[50] https://www.cabi.org/what-we-do/cabi-centre/trinidad-and-tobago/

[51] https://www.cabi.org/what-we-do/cabi-projects?section=1&region=central-america-and-the-caribbean&order=text-asc

[52] http://tt.biosafetyclearinghouse.net/0010.shtml

China and represents an important part of daily scientific research work. In the food and agriculture sector, this mainly involves sequence analysis of crop genome, transcriptome and protein groups and comparison with public databases. One expert reported that DSI is generated and used in 80 percent of their research. Sequencing data are generated locally, often through domestic contracted services, with the cost varying greatly each year depending on the sequencing type, the number of samples and the improvement of sequencing technology. Simple RNA samples cost CNY 700–1 100 (USD 100–1 700). These costs increase if bioinformatic analysis is also provided. Chinese researchers use data from other countries once they are released on a public database, as do researchers worldwide (e.g. GenBank, RefSeq,[53] SRA,[54] NSD) (Wu *et al.*, 2021).

Despite the extensive generation and use of DSI in China demonstrated by the figures available from the WiLDSI portal (see Section 3.1 of the main paper) researchers report limitations with regard to data accessibility, basic hardware conditions, such as network facilities and computers, data quality and format, data background information, and data analysis and utilization abilities. There is concern that researchers will not fully release the DSI data they generate after publishing their papers. Other concerns include the following: DSI data generated from a large number of enterprise/private sector grants are not released to the public; the public database storing DSI data is not being maintained; download speed is too slow or easy to interrupt in the case of access from China; and source information for released DSI data material is incomplete. Access to some databases is restricted, and some are no longer open to the public: often they can only be accessed if purchased by institutions. There are also a few databases with restricted access in China. DSI generation and use are extensively published in China by local researchers on local biodiversity in the fields of biotechnology, food and agricultural science, medical sciences and life sciences (Wu *et al.*, 2021; Li and Xue, 2019; Liu, 2021; Sun, Li and Zhao, 2021; Zhang, 2021).

Ghana

CABI's centre in Ghana serves West Africa, representing 15 countries with a combined population of about 300 million and where agriculture is of huge importance but not without significant challenges. It received two contributions on Ghana's ability to generate, access and use DSI. The first was from the Council for Scientific and Industrial Research (CSIR), Oil Palm Research Institute,[55] which is a division under Ghana's Crops Research Institute (CRI). This organization reported that although its researchers rely on DSI to a large extent, resource limitations mean that it is not able to generate and use DSI extensively. Researchers depend on external resources, which are expensive and not readily available or accessible. The problems affecting access and use include lack of technical infrastructure, financial and human resources, educational and training opportunities, scientific collaboration, computing infrastructure, reliable electricity, high-speed internet, capacity to store data and computing capacity, as well as the cost of data charges. DSI generated externally is used mainly as reference material. Examples of this are the genome sequencing of an African oil palm (*Elaeis guineensis*) (fruit type Dura) and whole genome resequencing of 72 oil palms from the African continent and Southeast Asian countries, which were performed on the Pacbio sequel II and Illumina sequencing platforms[56] by Temasek Life Sciences Laboratory, Singapore. Three of the Dura genotypes were from Ghana.[57] However, because of the lack of computing infrastructure, high-speed internet, technical infrastructure, and financial and human resources, the raw data were not analysed.

The second contribution came from the CSIR Crops Research Institute, which is an agriculture-based institution working on all food crops. This organization reported that the use of DSI is critical in its work but that researchers are not able to extensively generate and use it. They rely on platforms that produce results at a fee, which is not affordable for their resource-challenged systems. The CSIR-CRI currently has a 3730 Genome sequencer, which was purchased on a project. However, the installation was not completed, and staff were not trained in its use. Their ambition is to set it up as a hub to

[53] https://www.ncbi.nlm.nih.gov/Refseq
[54] https://www.ncbi.nlm.nih.gov/SRA
[55] https://opri.csir.org.gh/
[56] https://www.ncbi.nlm.nih.gov/bioproject/841085
[57] https://www.ncbi.nlm.nih.gov/sra/DRX281513; https://www.ncbi.nlm.nih.gov/sra/DRX281512; https://www.ncbi.nlm.nih.gov/sra/DRX281514

resource scientists in the subregions. Researchers in the institute generate as well as use DSI or data (both RNA and DNA). They use it for genetic analysis, including classification of plant viruses, bacteria, fungi and crop genotypes/varieties. This DSI is not generated locally: generally, it is generated by and sourced from international companies in the United States of America, Europe, South Africa or Australia. DSI is very important for the organization's research work. However, because of difficulty accessing it, its use is limited. DSI is available at websites such as NCBI, and with the right skills and internet availability it can be accessed when needed. Constraints to the availability of DSI include lack of technical infrastructure, financial and human resources, educational and training opportunities, scientific collaboration, computing infrastructure, reliable electricity, high-speed internet and capacity to store data, as well as the cost of charges.

The development of low-density single nucleotide polymorphism (SNP) panels and cost-effective genotyping platforms holds great promise for resource-limited breeding programmes in Africa. Prempeh *et al*. 2022 recommended the use of low-density SNP markers to characterize cassava accessions and for quality control in breeding activities. In crops such as rice and oil palm, different strategies have been employed to genotype large numbers of samples. SNP marker development through deep resequencing could provide all variation types on the resequenced region, but this is expensive (Xia *et al.*, 2019). However, low-density SNP microarrays are generated across known genes, and they therefore provide less-developed countries with a cost-effective option for analysing large sample data (Handyside and Wells, 2013). Using a high-density SNP panel, Bissah in 2016 (Bissah, 2016) mapped quantitative trait loci for salt tolerance in rice and compared the identified loci with mapped genes in SNPSEEK. Other examples of collaborative work with DSI include the following: determining the genetic makeup of some *Musa* (banana) hybrids (Quain, Agyeman and Dzomeku, 2018; Quain *et al.*, 2018); population studies of released and elite sweet potato;[12] comparison of the performance pepper varieties (Boateng *et al*., 2017); investigation of genetic relationships among genotypes of *Ceiba pentandra*, white silk-cotton tree (Abengmeneng *et al.*, 2016); characterization of *Solanum* species (Oppong *et al.*, 2015a); investigation of genetic relationships among cassava cultivars (Twumasi *et al.*, 2014); soybean diversity studies (Appiah-Kubi *et al.*, 2014); mapping of the distribution of maize streak virus genotypes across the forest and transition zones of Ghana (Oppong *et al.*, 2015b); and studying the distribution and spread of cassava mosaic virus disease in Ghana (Oppong *et al.*, 2021).

India

The CABI centre in India[58] has been implementing programmes and working with partners to help improve the lives of people and communities in the region since 1948. It manages and contributes to agricultural projects across South Asia, where more than 50 percent of the population is engaged in agriculture as its primary occupation. The region produces some key food security crops and is responsible for feeding a significant proportion of the world.

After consulting with partners in the region the centre reports that various private and public institutions generate and use DSI. The National Bureau of Plant Genetic Resources (NBPGR)-ICAR, New Delhi, deals with various aspects of the germplasm of agricultural and horticultural crops. The Forest Research Institute, Dehradun, deals with the germplasm of forest species, and the Botanical Survey of India, Calcutta, deals with the germplasm of the remaining plant species. The National Biodiversity Authority (NBA),[59] established by the Central Government in 2003, implements India's Biological Diversity Act of 2002. There are five national bureaus for genetic resources. Of these (under ICAR), the National Bureau of Plant Genetic Resources (NBPGR), National Bureau of Agricultural Insect Resources (NBAIR), National Bureau of Forest Genetic Resources (NBFGR) and National Bureau of Agriculturally Important Microbial Resources (NBAIMR) regularly generate DSI on plant, insect, animal and microbial genetic resources. Similarly, the Botanical Survey of India, Zoological Survey of India, Centre for Cellular and Molecular Biology (CCMB), state agricultural universities and many other universities generate DSI, with a focus on DNA and RNA sequences. Many crop-based institutes, such as the Indian Institute of Rice Research (IIRR), Indian Institute of

[58] https://www.cabi.org/what-we-do/cabi-centre/india/
[59] India-DSI.pdf (cbd.int)

Wheat and Barley Research (IIWR), Indian Institute of Millets Research (IIMR), Central Plantation Crops Research Institute (CPCRI), Indian Institute of Horticultural Research (IIHR) and many more, as well as private-sector players, such as the seed industry, generate DSI locally.

What started as a trickle in early 2000 is now a flood. ICAR itself has more than 60 institutes and universities, and most of the crop institutes and many of the universities under ICAR maintain small gene banks, and some institutes have their own germplasm information databases. Examples include the Germplasm Registration Information System,[60] the Musa Transcriptome Simple Sequence Repeat Database[61] and the CottonGen – Cotton Database Resources.[62] NBPGR[63] manages the second largest genebank in the world, with 400 000 accessions. It reports the use of genome technologies to improve breeding stock and even cites a patent (2011 Patent No. 245749) on a technology/process enabling simultaneous detection of two transgenes in transgenic maize.[64] Researchers have also accessed NSD from other countries for genetic improvement programmes. The response on DSI access and use indicated that some information is available to researchers, and indeed that some institutes provide training on how to use their databases (e.g. NBPGR).[48] Some institutions have online databases, while others do not but are working to create them. DSI information is also accessed by researchers in published literature. DSI is not available until it is published. Much of the DSI generated is for local use, and it is often not published or shared.

Partners in India explained that there are problems that restrict the use of DSI, including training constraints. There are apparent budget and resource constraints to the development and maintenance of information databases. Additionally, exploitation of resources, others patenting useful sequential information and reverse engineering, can prevent wide sharing of DSI.

Singh *et al*. in 2020 report on advances in sequencing technology and its outputs in their paper on the management and utilization of plant genetic resources in India. They explain that the "fundamental merit of an organized digital information system is that it provides fair and just opportunity for all to access. On-line portals, as a consequence of PGR Informatics, enable non-exclusive access to PGR information to a large number of users involved in overlapping research areas on PGR management."

Kenya

The CABI centre in Kenya serves the African region and is based in Nairobi. CABI runs projects that are essential for agriculture and economic growth in sub-Saharan Africa, where average crop yields are among the lowest in the world. DSI is starting to have an impact on science and agriculture in the region, but most of the DSI generated in Africa in recent years has been in the medical field, with most of this related to epidemics/pandemics such as Ebola and COVID-19. The DSI generated in food and agriculture has mainly been associated with DNA barcoding in cases where species need to be identified. Estimates of use are based on how much has been published, and data from the WiLDSI Data Portal[65] demonstrate that DSI from Kenya has been used by only 953 authors from Kenya but by over 9 000 authors from other countries.

Nucleotide sequences are generated and used locally, both within CABI projects and by others, as demonstrated again by the data from the WiLDSI Data Portal, but not to the same extent as in countries and regions such as China, Europe and the United States of America. These data are most often generated by universities, government-led research organizations, CGIAR centres, other international research organizations (such as CABI and ICIPE), national plant protection organizations and the private sector. African capacity is low, mainly because of a lack of infrastructure and financial resources, but data are available from the major nucleotide sequence databases such as NCBI, EMBL-EBI and DDBJ, which are all open access. Using DSI requires skills and knowledge in molecular sciences, and these are still limited in most of Africa. The cost of internet access is also an issue. The

[60] Germplasm Registration System (ernet.in)
[61] nrcb-bio (icar.gov.in)
[62] https://www.cottongen.org
[63] http://www.nbpgr.ernet.in/
[64] http://www.nbpgr.ernet.in/Technologies_and_IPRs.aspx
[65] https://apex.ipk-gatersleben.de/apex/wildsi/r/wildsi/home

data from the WiLDSI Data Portal indicate that access to and usage of DSI are still low and not significant.

Malaysia

The CABI centre in Malaysia works across the whole of Southeast Asia, a region that is still largely dependent on agriculture and is very rich in biodiversity and environmentally fragile. Projects run through the centre focus on crops such as rice and fruit crops, invasive species and strengthening agricultural ecosystems.[66] DSI is used at significant levels to identify genetic resources and their traits. CABI works with many partners and local agencies in Malaysia. The Department of Agriculture stated that it does not have any countrywide system for access to and use of DSI and that there are no centralized DNA/RNA databases. However, the Malaysian Agriculture Research and Development Institute (MARDI) has its own genebank system,[67] called AgrobiS.[68] AgrobIS is an information system developed by MARDI to provide the public with direct access to data on all the genetic resources conserved at MARDI. The system, once fully operational, will contain germplasm information for more than 40 000 accessions of plant genetic resources for food and agriculture, including fruits, rice, vegetables and medicinal plants. The system also includes information on 2 500 isolates of microbial genetic resources and about 30 000 specimens of arthropods.

MARDI explained that DNA and RNA sequences are generated using a next-generation sequencing platform for non-model species or organisms in food and agriculture. The sequences are used to develop molecular markers for use in genetic and molecular breeding studies, such as those on DNA fingerprinting, genetic diversity, population genetics and quantitative trait loci analysis. For bacterial genome sequencing, DNA and RNA sequences are used to harness beneficial microbes obtained from the soil microbiome to develop biofertilizer. For model organisms or species that are available in the public biological database, MARDI researchers download and use sequences to mine molecular markers or genes of interest. To publish in high-impact scientific journals, the researcher must deposit the sequence information in well-known databases (NCBI SRA, NCBI GenBank, ENA). This promotes reproducibility and transparency in the scientific community.

DSI is generated locally by MARDI for its projects: it could be as little as two to three organisms per year. The molecular biologist will send the samples (DNA/RNA/raw materials) to the sequencing provider. The bioinformatician will analyse the sequence data accordingly. Additionally, DSI generated by others is used extensively: for example, genome and transcriptome sequences in public databases are used to mine molecular markers or genes of interest and in comparative genomics analysis. DSI is easy to access and use from public databases, as instructions are provided. Moreover, the owner of each public database publishes a peer-reviewed article so that other researchers can learn about the database's function and availability.

MARDI currently describes the problems that reduce access and ability to use DSI in Malaysia as "low risk". For example, a user may be unable to access DSI because of a network issue. Some countries impose restrictions on access to data, often requiring users to log in using their organization's email. For commercial companies, a subscription is needed to access some databases.

Pakistan

Pakistan has made tremendous progress in developing biotechnology by establishing over 50 institutes/centres in the public sector during last five decades. Funding of over PKR 20 billion (more than EUR 88 million) was made available in the government sector under the Planning Commission of Pakistan, Higher Education Commission (HEC), Pakistan Science Foundation (PSF), and through various international development partners, such as the Committee on Scientific and Technical Commission (COMSTECH),[69] the United States Department of Agriculture, the United States Agency

[66] https://www.cabi.org/what-we-do/cabi-centre/malaysia/
[67] http://mygenebank.mardi.gov.my/
[68] http://agrobis.mardi.gov.my/agrobis_v2/admin/what-is-agrobis
[69] https://www.comstech.org/

for International Development and the International Centre for Genetic Engineering and Biotechnology.

As a result of these efforts, considerable human resources with sufficient scientific expertise have been developed. The country has therefore benefited from well-trained staff in this sector in the context of the mushrooming growth of PCR-based diagnostic tests for various diseases, especially hepatitis and recently COVID-19. Similarly, applications of modern biotechnology in forensic science have also made great strides, with a forensic laboratory in Punjab, Pakistan.[70] DNA marker studies are also carried out routinely in the food and agriculture sector in Pakistan, including studies of viruses, microbes, insects, plants and animals.

The response to the question "Is DSI generated locally, if so can you estimate how much and by whom?" indicated that, with support from Japan, over 90 000 plant accessions are being conserved and maintained at the Plant Genetics Research Institute (PGRI) at National Agriculture Research Centre (NARC) in Islamabad. In 2017, PGRI was restructured and renamed the Bio-resources Conservation Institute (BCI). Two new research programmes, including the Microbial Genetic Resources Programme and the Animal Genetic Resources Programme, were established to extend the research activities from plant genetic resources to microbial and animal genetic resources. The progress is slow but is continuing in the right direction.[71]

Pakistan Barcode of Life[42] (PakBOL) "offers a platform to the barcoding community in Pakistan and is a member of the International Bar Code of Life (iBOL)."[72] This enables countrywide collaboration to document and understand the country's biodiversity, building on DNA barcoding studies conducted in 2010. Collaboration is underway with research scientists at the Centre for Biodiversity Genomics at the University of Guelph, Canada, where iBOL's secretariat is located. To date, Pakistan's research community has generated around 50 000 barcode records from animals and 1 600 from plants. The country has also been actively participating in iBOL's Global Malaise Trap Programme,[73] with arthropod sampling completed at nine sites already and ongoing at 11.[74]

This forms a modest beginning, and further investment and work is needed. However, it illustrates that there is capacity and an impetus to enable more generation and use of DSI. The degree of effort needed is reflected in the state of microbial culture collections in Pakistan. Ahmed *et al*. in 2018 described the importance of microbes in biotechnological, agricultural and industrial applications in Pakistan. According to the World Data Centre for Microorganisms,[75] when consulted in 2018, there were five Culture Collection Centres registered from Pakistan, holding just over 3 500 strains. These include strains with plant growth promoting activity, strains that may have applications in bioremediation of heavy-metal polluted soils and water systems, strains useful in the food industry, pathogenic strains of bacterial blight from rice and citrus canker, and other extremophilic (e.g. salt-tolerant) strains for biotechnology. However, very few of these microbes have been truly identified at species level based on 16S ribosomal RNA gene sequence, and examples of new species of bacteria from the rich ecology of Pakistan are rare. Few of the holdings have been gene sequenced. Research collaborations with, for example, China and Saudi Arabia have made it possible to isolate and identify novel species of bacteria from Pakistan and sequence them (Amin *et al.*, 2016; Ali *et al.*, 2021). In addition to identification of microbial diversity, NSD is proving valuable in Pakistan for the genome sequence of multidrug-resistant novel candidate bacteria,[23] for DNA barcoding and biochemical profiling of medical plants, for the study of genotypic variation for drought tolerance in cotton (Rahman *et al.*, 2008), for characterizing insects, for example revealing cryptic species complexes by DNA barcode analysis of thrip (*Thysanoptera*) diversity in Pakistan (Iftikhar *et al.*, 2016), for a DNA barcode survey of insect biodiversity in Pakistan (Ashfaq *et al.*, 2022), for the identification of edible

[70] https://pfsa.punjab.gov.pk/
[71] http://www.parc.gov.pk
[72] https://ibol.org/about/ibol-consortium/
[73] http://globalmalaise.org/homecopy/
[74] https://ibol.org/press-release/the-launch-of-pakistan-barcode-of-life-pakbol/
[75] http://www.wfcc.info/ccinfo

fish species through DNA barcoding (Ghouri *et al.*, 2020), and for the use of DNA barcoding to control the illegal wildlife trade (Rehman *et al.*, 2015).

Global data, especially data available in public databases, are utilized frequently by specific research groups in Pakistan in the study of viruses and bacteria, and in plant and animal research. Bioinformatics departments are active in many public and private universities in Pakistan. To enable facilitated access and use of DSI, which is significant, the Higher Education Commission of Pakistan[76] has established a digital library for universities (more than 280) in the country. Additionally, G4/G5 connectivity to the internet is readily available in the country at a reasonable cost and speed. The use of sequence data in certain cases is mandatory, for example in ensuring the purity of basmati rice for export to European Union countries. Similarly, DNA fingerprinting is essential for approval/registration of a new plant variety by the Federal Seed Certification and Registration Authority. In order to comply with World Trade Organization and International Union for the Protection of New Varieties of Plants obligations, Pakistan established the Plant Breeders Registry and the requirement for DNA fingerprinting of novel plant varieties. The courts in Pakistan now accept DNA/RNA sequencing data for any dispute on ownership, paternity, theft and any other criminal acts. Furthermore, under Pakistan Biosafety Rules, 2005,[77] PCR-based testing using specific molecular markers to detect the presence of genetically modified organism material is mandatory; the Government of Pakistan Gazette notified four public research centres for this purpose.

The country is making progress in pockets on nucleotide sequence data (NSD), commonly called molecular markers and/or DNA fingerprinting. A few elite centres among the 50 institutes of biotechnology are capable of carrying out reasonable quality work on DSI. Trained human resources and sufficient infrastructure are available. However, there is a strong need for more collaboration and sustainability in DSI-related programmes.

Zambia

CABI's Zambia office serves the Southern African region, where agriculture is the main employer and source of income for the majority of the population. The office oversees projects and improves knowledge sharing to address agricultural and environmental challenges encountered by Southern African smallholder farmers. The office received feedback from Dr Paul W. Kachapulula, Head of Department, Department of Plant Sciences, School of Agricultural Sciences, and Dr Evans Kaimoyo, UNZA School of Natural Sciences, both at the University of Zambia (UNZA); and Dr Rabson Mulenga, Zambia Agricultural Research Institute (ZARI), Plant Pathology Department, Mount Makuru.

In his response, Dr Kachapulula (UNZA) estimated that about 20 percent of the members of staff at UNZA generate and use DSI. This is mostly sequencing of microorganisms (mainly fungi and bacteria) and genotypes as part of species identification, and research on gene expression and the qualities of new cultivars. Scientists normally extract DNA/RNA, ship it out of the country for sequencing and receive electronic sequences for further analyses. Such analyses normally require online access to databases such as NCBI and several others. Sequencing is normally done abroad at places such as INQABA, University of Cape Town (South Africa) and BECA (Kenya). The School of Veterinary Medicine at UNZA has sequencing capabilities, but the centre is not yet optimized for commercial use. Dr Kaimoyo (UNZA) indicated that relatively little DSI is generated in the country, and that it is mostly limited to universities, and medical and agricultural research institutes, which are few in number. Research scientists at the Schools of Veterinary Medicine, Natural Sciences, Agriculture Sciences and Health and Medical Sciences and at local and internationally affiliated research institutes are generating sizeable amounts of sequence data, equivalent to 40–50 percent of what is downloaded from sequence databases. Dr Mulenga, reporting on the generation and use of DSI at ZARI, explained that the institute generated and used DSI regularly to detect and characterize pathogens that infect crops, and used the data to generate technologies that reduce crop yield loss

[76] http://www.hec.gov.pk
[77] https://www.fao.org/faolex/results/details/en/c/LEX-FAOC053471

caused by pathogens. He reported that about 30 percent of the DSI used is generated locally. Data generated locally are deposited in public databases.

At UNZA, as with all university and research institutes in Zambia, staff are able to access DSI from publicly available online resources: internet facilities and quality are fairly good. However, access to databases often needs a subscription, and for high-throughput computations, the facilities at UNZA need upgrading. Dr Kaimoyo (UNZA) indicated that nucleotide and polypeptide sequence information use is to some extent dependent on the availability of sequencing services in the country. In Zambia, it is still relatively hard to find local gene-sequencing services, let alone genome-sequencing and polypeptide-sequencing facilities. Most of the work done by UNZA depends on sequencing services from abroad, where DNA samples are typically sent for sequencing at a relatively high cost. Dr Mulenga (ZARI) reported the main problem to be the actual generation of reliable DSI, mainly because of a lack of computing capacity and the complexity of bioinformatics. It is reported that DSI on Zambian biodiversity has been generated almost entirely in collaboration with the United States of America[29] and other countries, such as Australia (Mulenga *et al.*, 2015a,b; 2018; 2020; 2022).

Appendix 2 References

Abengmeneng, C.S., Ofori, D.A., Kumapley, P., Akromah, R., Jamnadass, R. & Quain, M. 2016. Genetic relationships among 36 genotypes of *Ceiba pentandra* (L.) as revealed by RAPD and ISSR markers. *American Journal of Agriculture and Forestry,* 4(4): 86–96.

Ahmed, I., Abbas, S. & Tariq, H. 2018. Importance of microbial culture collection in Pakistan: challenges and opportunities. *Bulletin of the BISMiS,* 7(2): 44–48. https://www.bismis.net/files/BulletinofBISMIS_7-2.pdf

Ali, A., Tariq, H., Abbas, S., Arshad, M., Li, S, Dong, L., Li., L., Li, W.J. & Ahmed, I. 2021. Draft genome sequence of a multidrug-resistant novel candidate *Pseudomonas* sp. NCCP-436 isolated from faeces of a bovine host in Pakistan. *Journal of Global Antimicrobial Resistance*, 27: 91–94. https://doi.org/10.1016/j.jgar.2021.08.011

Amin, A., Ahmed, I., Habib, N., Abbas, S., Hasan, F., Xiao, M., Hozzein, W.N. & Li, W.J. 2016. *Microvirga pakistanensis* sp. nov., a novel bacterium isolated from desert soil of Cholistan, Pakistan. *Archives of Microbiology*, 198(10): 933–939. https://doi.org/10.1007/s00203-016-1251-3

Appiah-Kubi, D., Asibuo, J.Y., Quain, M.D., Oppong, A., & Akromah, R. 2014. Diversity studies on soybean accessions from three countries. *Biocatalysis and Agricultural Biotechnology*, 3(2): 198–206. http://dx.doi.org/10.1016/j.bcab.2013.11.008

Ashfaq, M., Khan, A.M., Rasool, A., Akhtar, S., Nazir, N., Ahmed, N., Manzoor, F. *et al*. 2022. A DNA barcode survey of insect biodiversity in Pakistan. *PeerJ,* 10: e13267. https://doi.org/10.7717/peerj.13267

Bissah, M.N. 2016. *A study of genetic variability and quantitative trait loci (QTL) for salinity tolerance in rice (*Oryza sativa *L.).* Accra, University of Ghana. PhD Thesis.

Boateng, S.K., Aboagye L.M., Egbadzor, K.F., Gamedoagbao, D.K., Allotey, L.N. & Quain, M.D. 2017. The performance of five selected pepper accessions in comparison with two local varieties. *Agricultural and Food Science Journal of Ghana*, 10(1): 795–802.

Cowan, D., Lebre, P., Amon, C., Becker, R.W., Boga, H.I., Boulangé, A., Chiyaka, T.L., Coetzee, T. *et al.* 2022. Biogeographical survey of soil microbiomes across sub-Saharan Africa: structure, drivers, and predicted climate-driven changes. *Microbiome*, 10: 131. https://doi.org/10.1186/s40168-022-01297-w

Ghouri, M.Z., Ismail, M., Javed, M.A., Khan, S.H., Munawar, N., Umar, A.B., Nisa, M. *et al*. 2020. Identification of edible fish species of Pakistan through DNA barcoding. *Frontiers in Marine Science*, 7: 554183. https://doi.org/10.3389/fmars.2020.554183

Handyside, A.H. & Wells, D. 2013. Single nucleotide polymorphisms and next generation sequencing. In: D. Gardner, D. Sakkas, E. Seli & D. Wells, eds. *Human gametes and preimplantation embryos*, pp. 135–145. New York, USA, Springer.

Iftikhar, R., Ashfaq, M., Rasool, A. & Hebert, P.D.N. 2016. DNA barcode analysis of Thrips (Thysanoptera) diversity in Pakistan reveals cryptic species complexes. *PLoS ONE*, 11(1): e146014. https://doi.org/10.1371/journal.pone.0146014

Li, B. & Xue, D. 2019. Application of digital sequence information in biodiversity research and its potential impact on benefit sharing. *Journal of Biodiversity Science*, 27(12): 1379–1385. https://www.biodiversity-science.net/EN/10.17520/biods.2019242

Liu, Q. 2021. The development status and the China's choice on the issue of digital sequence information of genetic resources. *Journal of Ecology and Rural Environment*, 37(9): 1109–1114. [in Chinese with English abstract]. https://doi.org/10.19741/j.issn.1673-4831.2021.0326

Mulenga, M.R., Miano, D.W., Chikoti, P.C., Ndunguru, J., Legg, J.P. & Alabi, O.J. 2015b. First report of *East African cassava mosaic Malawi virus* in plants affected by cassava mosaic disease in Zambia. *Plant Disease*, 99(9): 1290. https://doi.org/10.1094/pdis-03-15-0264-pdn

Mulenga, R.M., Boykin, L.M., Chikoti, P.C., Suwilanji, S., Ng'uni, D. & Alabi, O.J. 2018. Cassava brown streak disease and Ugandan cassava brown streak virus reported for the first time in Zambia. *Plant Disease*, 102(7): 1410–1418. https://doi.org/10.1094/PDIS-11-17-1707-RE.

Mulenga, R.M., Legg, J. P., Ndunguru, J., Miano, D.W., Mutitu, E.W., Chikoti, P.C. & Alabi, O.J. 2015a. Survey, molecular detection and characterization of geminiviruses associated with cassava mosaic disease in Zambia. *Plant Disease*, 100(7): 1379–1387. https://doi.org/10.1094/PDIS-10-15-1170-RE

Mulenga, R.M., Miano, D.W., Al Rwahnih, M., Kaimoyo, E., Akello, J., Nzuve, F.M., Simulundu, E., *et al*. 2022. Survey for virus diversity in common bean (*Phaseolus vulgaris*) fields and the detection of a novel strain of *Cowpea polerovirus 1* in Zambia. *Plant Disease*, 106(9): 2380–2391. https://doi.org/10.1094/PDIS-11-21-2533-RE

Mulenga, R.M., Miano, D.W., Kaimoyo, E., Akello, J., Nzuve, F.M., Simulundu, E., & Alabi, O.J. 2020. First report of Ethiopian tobacco bushy top virus and its associated satellite RNA infecting common bean (*Phaseolus vulgaris* L.) in Zambia. *Plant Disease*, 105(2): 516. https://doi.org/10.1094/PDIS-03-20-0596-PDN

Oppong L.A., Quain M.D., Oppong, A., Doku, H.A., Agyemang, A. & Bonsu, O.K. 2015a. Molecular characterization of *Solanum* species using EST-SSRs and analysis of their zinc and iron contents. *American Journal of Experimental Agriculture*, 6(1): 30–44. https://doi.org/10.9734/AJEA/2015/6337

Oppong, A., Offei, S.K., Ofori, K., Adu-Dapaah, H., Lamptey, J.N.L., Kurenbach, B., Walters, M. *et al*. 2015b. Mapping the distribution of maize streak virus genotypes across the forest and transition zones of Ghana. *Archives of Virology*, 159: 1–10. https://doi.org/10.1007/s00705-014-2260-7

Oppong, A., Prempeh, R.NA, Abrokwah, L.A., Annang, E.A., Marfo, E.A., Appiah Kubi, Z., Danquah, N.A.O. *et al*. 2021. Cassava mosaic virus disease in Ghana: distribution and spread, *Journal of Plant Physiology and Pathology*, 9(8): 258.

Prempeh, R., Oppong, A., Amankwaah, V., Akomeah, B., Abrokwah, L., Allotey, L., Annang, E., Bosompem, A.N., Amoako, F. & Mbanjo, G.E. 2022. Characterization of cassava germplasm using a low-cost SNP panel. In: *Report of the Second Conference of the African Plant Breeding Association Conference (APBA) 25–29 October 2021, Kigali, Rwanda*, pp 43–44. APBA.

Quain, M.D., Adofo, K., Appiah-Kubi, D., Prempeh, R.N., Asafu-Agyei, J., Akomeah, B. & Dapaah, H. 2018. Use of expressed sequence tags-derived simple sequence repeat (SSR) markers for population studies of released and elite sweet potato. *International Journal of Genetics and Molecular Biology*, 10(2): 14–25. https://doi.org/10.5897/IJGMB2017.0159

Quain, M.D., Agyeman, A. & Dzomeku, B.M. 2018. Assessment of plantain (*Musa sapientum* L.) accessions genotypic groups relatedness using simple sequence repeats markers. *African Journal of Biotechnology*, 17(16): 541-551. https://doi.org/10.5897/AJB2017.16363

Quain, M.D., Agyeman, A., Okyere E., & Dzomeku, B.M. 2018. Unravelling genetic makeup of some *Musa* hybrids and selected *Musa* accessions using molecular and morphological characterization. *International Journal of Genetics and Molecular Biology*, 10(1): 1–13. https://doi.org/10.5897/IJGMB2018.0161

Rahman, M., Ullah, I., Ashraf, M. & Zafar, Y. 2008. A study of genotypic variation for drought tolerance in cotton. *Agronomy for Sustainable Development*, 28: 439–447. https://doi.org/10.1051/agro:2007041

Rehman, A., Jafar, S., Raja, N. A. & Mahar, J. 2015. Use of DNA barcoding to control the illegal wildlife trade: a CITES case report from Pakistan. *Journal of Bioresource Management*, 2(2): 19–22. DOI: 10.35691/JBM.5102.0017

Singh, K., Gupta, K., Tyagi, V., & Rajkumar, S. 2020. Plant genetic resources in India: management and utilization. *Вавиловский журнал генетики и селекции* [*Vavilov Journal of Genetics and Breeding*], 24(3): 306–314. https://doi.org/10.18699/VJ20.622

Sun, M., Li, Y. & Zhao, F. 2021. 生物遗传资源保护、获取与惠益分享现状和挑战 [Current status and challenges of protection, access to and benefit sharing of bio-genetic resources of China]. 环境保护[*Environmental Protection*], 49(21): 30–34. doi: 10.14026/j.cnki.0253-9705.2021.21.002

Twumasi, P., Acquah, E.W., Quain, M.D. & Parkes, E.Y. 2014. Use of simple sequence repeat (SSR) markers to establish genetic relationships among cassava cultivars released by different research groups in Ghanaian. *International Journal of Genetic and Molecular Biology*, 6(3): 29–36. https://doi.org/10.5897/IJGMB2014.0097

Wu, L., Shi, L., Gao, M. & Ma, J. 2021. Analysis on the status and suggestions for the development of digital sequence information of genetic resources. *China Science & Technology Resource Review*, 53(2): 36–43. [in Chinese with English abstract]. https://doi: 10.3772/j.issn.1674-1544.2021.02.005

Xia, W., Luo, T., Zhang W., Mason A.S., Huang D., Huang X., Tang W. *et al.* 2019. Development of high-density SNP markers and their application in evaluating genetic diversity and population structure in *Elaeis guineensis*. *Frontiers in Plant Science*, 10: 130. https://doi.org/10.3389/fpls.2019.00130

Zhang, X. 2021. Rules challenges and countermeasures of sharing of pathogens in the context of the implementation of the Nagoya Protocol. *Journal of Ecology and Rural Environment*, 37(9): 1098–1103. [in Chinese with English abstract]. doi: 10.19741/j.issn.1673-4831.2021.0423