



Convention on Biological Diversity

Distr.
GENERAL

CBD/SBSTTA/22/INF/2/Add.1
CBD/DSI/AHTEG/2018/1/2/Add.1*
18 January 2018

ENGLISH ONLY

AD HOC TECHNICAL EXPERT GROUP ON
DIGITAL SEQUENCE INFORMATION ON
GENETIC RESOURCES

Montreal, Canada, 13-16 February 2018
Item 3 of the provisional agenda**

SUBSIDIARY BODY ON SCIENTIFIC,
TECHNICAL AND TECHNOLOGICAL ADVICE

Twenty-second meeting
Montreal, Canada, 2-7 July 2018
Item 3 of the provisional agenda***

SYNTHESIS OF VIEWS AND INFORMATION ON THE POTENTIAL IMPLICATIONS OF THE USE OF DIGITAL SEQUENCE INFORMATION ON GENETIC RESOURCES FOR THE THREE OBJECTIVES OF THE CONVENTION AND THE OBJECTIVE OF THE NAGOYA PROTOCOL

Note by the Executive Secretary

Addendum

CASE STUDIES AND EXAMPLES OF THE USE OF DIGITAL SEQUENCE INFORMATION IN RELATION TO THE OBJECTIVES OF THE CONVENTION AND THE NAGOYA PROTOCOL

INTRODUCTION

1. Many of the submissions of views and information on the potential implications of the use of digital sequence information on genetic resources for the three objectives of the Convention and the objective of the Nagoya Protocol included cases studies and examples illustrating work being done in this area. These case studies and examples help to understand the breadth and complexity of this issue and a number of them are compiled here to elaborate on points made in the synthesis document.¹ The original submissions should be consulted for further details and for sources of information.²

* Reissued for technical reasons on 22 May 2018.

** CBD/DSI/AHTEG/2018/1/1.

*** CBD/SBSTTA/22/1.

¹ [CBD/DSI/AHTEG/2018/1/2.](https://www.cbd.int/abs/dsi-gr/ahteg.shtml#submissions)

² The submissions are available at: <https://www.cbd.int/abs/dsi-gr/ahteg.shtml#submissions>

I. USE OF DIGITAL SEQUENCE INFORMATION ON GENETIC RESOURCES FOR THE CONSERVATION AND SUSTAINABLE USE OF BIOLOGICAL DIVERSITY

General considerations on digital sequence information and the conservation and sustainable use of biodiversity

Example 1. Environmental DNA and endangered animals³

Analysing environmental DNA from water samples collected at watering sights in tropical forests enables detection of endangered species of mammals, thus providing insight into their survival rates. Unlike conventional methods to investigate wildlife with camera traps, the use of digital sequence information in the form of environmental DNA results in a more precise survey of wildlife and is also useful for species conservation.

Role of digital sequence information in projects that contribute to the conservation and sustainable use of biodiversity

Trees and forests

Example 2. Canada's National Forest Information System⁴

Canada developed the National Forest Information System⁵ in 2000 as a collaborative effort with Natural Resources Canada, the Canadian Forest Service and the Canadian Council of Forest Ministers to provide access to the most current, consistent and reliable forest resources information, integrating information across jurisdictional boundaries. This system contains data that allows for an accurate picture of Canadian forest practices and forest biodiversity, including assessing genetic conservation requirements of native tree species of Canada, information concerning threats to genetic diversity (e.g. invasive alien species), species biology and ecology.⁶

Example 3. Alien forest enemies – the BioSAFE project⁷

A project funded by Genome Canada is involved in the protection of Canadian biodiversity through biosurveillance of alien forest enemies (BioSAFE). It is led by researchers from Canadian universities in collaboration with the Canadian Food Inspection Agency. This national-level project will enable forest health professionals to track and identify forest invasive pests and diseases using a genomics based approach. This project will enable the development of DNA-based diagnostic tests to identify and monitor these pathogens quickly (within hours) and accurately which will then be used as a decision support tool to mitigate threats. The success of this project requires the availability of open access to genetic sequence data of known pathogens for identification.

Example 4. Responding to the Ash Dieback outbreak⁸

In mid-September 2012, conservation volunteers recognized that ash trees (the third most common tree in the United Kingdom) in an ancient woodland in Norfolk, United Kingdom were showing unusual disease symptoms. Following analysis of the diseased material, DNA sequencing confirmed that the infection was a fungal pathogen spreading across Europe. Recognizing that Norfolk was likely at the edge of the

³ Japan.

⁴ CGRFA – Canada. For an explanation of the short forms used in the present document, see the annex to [CBD/DSI/AHTEG/2018/1/2](https://www.cbd.int/doc/DSI/AHTEG/2018/1/2/Add.1.pdf).

⁵ https://ca.nfis.org/index_eng.html

⁶ See also <https://pfc.cfsnet.nfis.org/CAFRIS/home.jsp>

⁷ CGRFA – Canada.

⁸ Group of genomics experts.

epidemic spreading from Europe, scientists at several research institutions formed an ad hoc emergency response group to start sequencing the genomes of the pathogen and infected trees. In the hope of encouraging a rapid response, they agreed to make all data immediately available with open access. They built websites and software tools to make data available and accessible to genomic scientists, biologists and the public, for who they built a Facebook game that received over 63,000 plays in the first year enabling the public to be involved in the rapid assembly of the genomes. The result of the “Open Dieback Project” was an unprecedented speed of discovery. In just a few years, the consortium was able to identify genetic markers for trees with low susceptibility to the disease and the fundamentals required to select and breed trees with enhanced tolerance to the disease.

Mammals

Example 5. Conservation of endangered gorilla populations⁹

Mount Tschierimu in the Democratic Republic of the Congo is home to a highly endangered population of gorillas. With a population of one female, four males and one baby of unknown sex, the colony is no longer viable and needs new gorillas to provide enough genetic variability for the colony to continue. Living on Mount Tschierimu, they were thought to be mountain gorillas, but some unusual characteristics put this in question. To confirm their species, researchers at the Sanger Institute working with researchers from the United States and conservationists from the Democratic Republic of the Congo, took a sample from a gorilla, which was sequenced at the Sanger Institute and compared to openly available digital sequence information from three gorilla species. The analysis showed that it was not a mountain gorilla but an eastern lowland subspecies, which was necessary information to develop a conservation strategy. The digital sequence information of the gorilla was also deposited in a public database. Identifying the gorilla’s species was only possible because the researchers were able to access the digital sequence information of the other gorillas.

Microorganisms

Example 6. Study of Soil and Water Microbiome: the EcoBiomics project¹⁰

Agriculture and Agri Food Canada (AAFC) is leading an interdepartmental project that aims to develop new knowledge to improve water quality and soil health by comprehensively characterizing aquatic microbiomes, soil microbiomes, and invertebrate zoobiomes, and testing hypotheses in order to enhance environmental monitoring, assessment and remediation activities. It also aims at establishing comprehensive biodiversity baselines for assessing future changes to water and soil biodiversity at important long-term environmental monitoring sites in Canada.

The EcoBiomics project directly supports the objective of conserving biological diversity. This project uses metagenomics approaches to profile microbial and invertebrate communities in varying habitats tied to water and soil, including pristine natural areas, agricultural, forestry and fishery systems, oil sands, and lakes and rivers. The result of the sampling in these locations based on standardized molecular approaches will define the biodiversity within a taxonomic, ecological and functional context. The ability to do this analysis is absolutely dependent on known sequence information from resources such as GenBank to identify the taxa in the sample and potentially their role in the ecosystem being studied. This project may also contribute to the identification and naming of taxa that are not known to science and openly share this knowledge. This biomonitoring is essential to understand taxa and communities that may need conserving and thus potential regulatory actions. There is also a strong economic tie to the production systems involved, as maintaining diversity is typically correlated to healthy systems and thus production as well as monitoring for invasives and functional changes that could damage productivity.

⁹ Wellcome Trust and Wellcome Sanger Institute.

¹⁰ Canada.

The objectives of EcoBiomics also directly support sustainable use of the components of biodiversity as its production systems rely on sustainable biodiversity-based ecosystem services to be productive. The biomonitoring will produce metrics to assess sustainability. It would also indicate what remediation in terms of soil or water diversity is necessary to allow ecosystems to return to “normal” function.

Plants and fungi

Example 7. Catalogue of Plants and Fungi of Brazil and List of Species of Brazilian Flora¹¹

The relationship between conservation and the List of Species of Brazilian Flora is exemplified when, in 2010, the country published the Catalogue of Plants and Fungi of Brazil and launched the first online version of the List of Species of Brazilian Flora, meeting Target 1 of the Global Strategy for Plant Conservation. This botanical milestone was only achieved due to the commitment of more than 400 Brazilian and foreign taxonomists who worked on a platform where information about the flora was included and disseminated in real time. The “Brazilian List”, as it was popularly known, closed in November 2015 with the publication of five papers and their respective databases dealing with the different groups of fungi and plants. It is important to emphasize that the list of species of “Flora do Brasil” was evaluated on the basis of studies of molecular phylogeny and taxonomic reviews.

Biodiversity for food and agriculture

Example 8. Studying pollinator diversity and understanding threats to their populations¹²

At least 35 per cent of global crop production, and the majority of fruits, depend on bees and other insects for pollination services. While most interest in insect pollinators has focused on commercial honey bees, there is growing understanding that native bees are more efficient and effective crop pollinators. The documented alarming declines in both commercial honey bees and wild bee populations has led to increased fears that current agricultural productivity may be unsustainable without concerted efforts to maintain and enhance both wild and cultivated bee populations. AAFC conducts research focusing on documenting bee diversity in Canadian agricultural landscapes and studying the impacts of land use, pesticides and pathogens on bee populations. This work requires AAFC to sample large numbers of bees which all need to be identified to species level.

In order to efficiently and accurately identify species, AAFC often sequences various genomic regions of these bees and compares sequences to those that have been made available in online databases through collaborative efforts of the scientific community for well over a decade. Having access to sequences of bees from other countries is vital to be able to detect any newly introduced species that AAFC’s samples may be the first to detect in Canada. AAFC also makes use of worldwide sequences of pathogens known to negatively impact bees in other parts of the world, in order to screen bees sampled in Canada for these pathogens. Equally as important are the sequences from the thousands of bees and pathogens that Canada has produced that represent invaluable tools for other researchers worldwide.

Example 9. Understanding gene flow related to herbicide resistance¹³

Currently, there are over 450 weed species globally that have evolved certain levels of resistance to at least one, but often many, herbicides. Within Canada, just over 60 species are herbicide-resistant and within the United States there are over 150, some of which are spreading or are at risk for spreading into Canada. Furthermore, with the introduction of crops modified to have herbicide resistance, the genes that confer these traits could move into wild relatives, providing additional challenges for weed management for the farming community. Herbicide-resistant weeds, whether they have evolved or have received this resistance through gene flow, are a clear threat to agricultural production within Canada and globally. Within the Brassicaceae, it has been documented that transgenes for herbicide resistance can move from

¹¹ CGRFA – Brazil.

¹² Canada.

¹³ Canada.

crops (e.g. canola; *Brassica napus*) into weeds (e.g. Birdrape, *Brassica rapa*). This creates herbicide-resistant weeds that are difficult to control. In addition, if canola crops with different herbicide resistance genes are grown, it is possible for these traits to combine in the weed, resulting in weeds with multiple resistances. As a result, it is important to assess the potential of transgenes escaping from Brassicaceae crops.

The research conducted at AAFC to address this threat relies at every stage on free and open access to genomic resources. As a first step, AAFC uses the freely available information on genes from species around the world to determine which weed species are most closely related to novel crops and therefore most likely to receive transgenes. This enables efficient allocation of resources to where the risk of gene flow is the highest. Second, AAFC uses similar information to develop molecular markers to detect hybrids between crops and their wild relatives, allowing determination of the rate at which this hybridization is likely to occur. This information is required by the Plant Biosafety Assessment Office within the Canadian Food Inspection Agency to facilitate informed, science based decisions on the unconfined release of enhanced cultivars. Finally, the availability of accessible published genomes allows AAFC to create the tools needed to dissect the consequences of hybridization so that an informed risk benefit analysis can be made, including: (a) pinpointing the potential frequency of this hybridization in nature; (b) determining the portions of the genome that are most likely to be exchanged; and (c) quantifying how quickly a transgene may spread once introduced. For example, a glyphosate resistant (e.g. Roundup resistant) ecotype of the weed *Kochia scoparia*, has recently invaded Canada from the United States and the gene that confers this resistance is spreading through Canadian populations. This species has the potential to cause devastating losses in wheat and soybean. Its genome has just been made available allowing for AAFC to have an unprecedented ability to understand how seed and pollen movement contribute to the spread of these genes through populations. This information will contribute to the development of more effective and sustainable weed management strategies for this species.

*Example 10. Understanding the genetic diversity of the Bactrian camel for conservation purposes*¹⁴

With their tolerance to cold, drought and high altitudes, Bactrian camels (*Camelus bactrianus*) have been particularly appreciated in the steppes and mountains of Central Asia and could be a source of potentially useful traits in breeding programmes, but their number has been decreasing rapidly in recent years. Strategic conservation and breeding strategies are hindered by, among other factors, the very limited knowledge about the genetic diversity of Bactrian camels and the relationship among existing populations. With research partners from China and Mongolia, scientists at the International Livestock Research Institute applied genetic analysis using microsatellites markers to characterize populations of Bactrian camels in these two countries. The study revealed significant differences among Chinese and Mongolian populations, showed gene flow within the target populations (possibly associated with trading along the Silk Road and transhumance) and confirmed that Bactrian camels from China and Mongolia are genetically distant and should be considered as distinct populations in conservation and breeding programmes.

*Example 11. International Potato Center genotypes accessions to understand the genetic structure of ex situ sweet potato collection*¹⁵

In Peru, farmers traditionally cultivated 20-40 different landraces of potatoes as a form of insurance; by planting such diversity, some landraces will produce a crop even in bad years, sustaining farmers until the following harvest. Over the past several decades, the planting of many different potato landrace varieties by some communities has gradually decreased and many families are now planting fewer than ten varieties. A programme by the genebank of the International Potato Center (known as CIP from its Spanish name) has been operating to give back, or repatriate, landrace varieties collected from these areas

¹⁴ CGIAR.

¹⁵ CGIAR.

and thus contribute to building back-up systems whereby indigenous communities can continue traditional farming practices where diversity plays a key role in long-term sustainability. One challenge associated with this restoration work is that it is not known what diversity existed 50 or 100 years ago or even what diversity exists today in the *ex situ* genebank. Thus, the CIP genebank has recently genotyped its entire collection of cultivated landrace potato and sweet potato, laying the foundation for assessing the diversity present in the *ex situ* collection.

Role of digital sequence information in plant and animal breeding

Example 12. Gene editing at the International Maize and Wheat Improvement Center: Disease Resistance and Grain Quality¹⁶

The International Maize and Wheat Improvement Center (known as CIMMYT from its Spanish name) is in the process of editing genes using the CRISPR/Cas9¹⁷ system for stress tolerance and quality traits in maize and wheat. Maize Lethal Necrosis (MLN), a disease prevalent in East Africa that is caused by a combination of two viruses, poses a significant threat to food security in that region. CIMMYT, in collaboration with DuPont Pioneer (now DowDuPont), has identified a strong source of MLN resistance and is close to isolating the responsible gene. Most of the hybrids in Africa are generated from combining three parents. The MLN resistance CIMMYT has identified is recessive, which means that each of the parent lines in hybrids would need to be modified to confer a resistant phenotype on the hybrids. Conventional backcrossing is a time-consuming and resource-intensive process. In addition, it is nearly impossible to recreate the original makeup of the hybrid parents as the residual genome of the donor-resistant parent causes variable drag on grain yield. CIMMYT and DowDuPont will edit the susceptible gene to its resistant form directly in the parents of the susceptible commercial African hybrids. This will not only save years of time but also eliminate any chance of yield drag by precisely modifying a single locus. In wheat, CIMMYT plans to focus its gene editing efforts on creating additional variation for durable rust resistance, the preferred mode of resistance by the breeders, and on improving available metal ions (zinc and iron) through downregulation of phytate in the grain.

Example 13. Arabidopsis – a model system for plant breeding¹⁸

Arabidopsis thaliana is a small weedy plant that belongs to the Brassicaceae family. The genome of *Arabidopsis thaliana* was fully sequenced in 2001 and now serves as a model species, not only for other Brassicaceae family members but for all seed plants. Digital sequence information is made available through the Arabidopsis Information Resource (TAIR).¹⁹ Arabidopsis is used to study almost all aspects of plant life, such as development, metabolic pathways, disease resistance and adaptation. The knowledge gained using this model system and the digital sequence information made available through TAIR have influenced breeding and helped to breed for a wide variety of traits, such as salt tolerance in rice, flowering time in sugar beet and resistance to clubroot disease in *Brassica napus*. Important for the success of such breeding efforts, which, in essence, build on a comparative analysis between a crop and a model species, is that a database with digital sequence information is also available for the crop species.

Example 14. Unlocking the genetic diversity of creole wheats²⁰

Mexican wheat landraces, also known as “creole wheats”, were brought to the Americas from the 16th through 18th centuries and gradually became adapted to the local environments including many heat and

¹⁶ CGIAR.

¹⁷ CRISPR is an abbreviation of “clustered regularly interspaced short palindromic repeats” while Cas9 refers to a specific CRISPR-associated (Cas) protein, which is an enzyme that cuts DNA. CRISPR/Cas9 is used for targeted gene editing by directing the Cas9 protein to a specific location within the DNA where it cuts. This allows for the DNA to be edited at the location where the cut was made.

¹⁸ CGRFA – Prof. Sundstrom.

¹⁹ <https://www.arabidopsis.org/>

²⁰ CGIAR.

drought-stressed regions. As such, they should have useful genetic variation for stress tolerance. The introduction of the genetic diversity of these creole wheats into breeding pipelines has potential for developing the next generation of wheat varieties. With this objective in mind, a team of scientists from CIMMYT, the Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (INIFAP) in Mexico and the Punjab Biodiversity Board (India) carried out a study to: (a) characterize the collection of Mexican wheat landraces conserved in the CIMMYT germplasm bank; and (b) develop a core reference set using multiple variables. A core reference set is a subset of a genetic resources collection representing the diversity present in the whole collection but small enough for breeders to evaluate for interesting traits. Core reference sets have been established in the past on the basis of one variable, for example genotypic data or phenotype measures or geographical distribution. Simultaneous use of multiple types of variables (genotype, phenotype, geography, etc.) provides a robust diversity estimate for its application in plant breeding. As a result, 8,416 wheat landraces, representing a range of Mexican agro-ecologies, were characterized by genetic markers (DarTseq) and also phenotypically for yield potential, drought and heat tolerance, and yellow rust resistance to identify a core reference set that can represent this important variation. This core reference set captures 89 per cent of the rare alleles present in the complete set.

*Example 15. Creating “mini-core” collections to enhance utilization of germplasm for crop improvement*²¹

Germplasm diversity is basic to crop improvement programmes. However much of the germplasm in genebanks has yet to be used in crop improvement programmes. Greater use of germplasm in crop improvement programmes is needed for sustained and enhanced agricultural production for food security. The genebank of the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) conserves over 125,000 accessions of six mandate crops and five small millets from 144 countries. The main reason for low use of germplasm is lack of information on traits of economic importance and the large size of the collections. To enhance utilization of germplasm in crop improvement programmes, representative core collections (10 per cent of the entire ICRISAT collection) were developed using data on quantitative and qualitative traits in chickpea, pigeonpea, groundnut, sorghum, pearl millet, finger millet, foxtail millet, Proso millet, barnyard millet, kodo millet and little millet. Unfortunately, the number of accessions in these core reference sets remains too large for its meaningful evaluation in breeding programmes. To overcome this, ICRISAT scientists developed the mini-core (10 per cent of the core collection or 1 per cent of the entire collection) concept and proposed a two-stage strategy using qualitative and quantitative trait data of the mini-core collection. Extensive multidisciplinary evaluation of mini-core collections has identified new sources of variation for multiple traits, including tolerance to biotic and abiotic stresses, and for nutritional and agronomic traits. These mini-core reference sets have been distributed to breeders in 36 countries for use in improvement programmes. Sequencing these mini-core collections would be of great value to determine sequence variation associated with traits and help identify the most useful germplasm lines for use as parents in breeding programmes.

*Example 16. Combining participatory farmer variety evaluation and selection with genotyping: wheat landraces in Ethiopia*²²

In a recent study carried out by Bioversity International under the Seeds for Needs programme, a combination of participatory approaches, genomics and quantitative genetics was used to trace the genetic basis of smallholder farmer preferences of durum wheat traits in Ethiopia. Two smallholder communities evaluated 400 Ethiopian wheat varieties, mostly landraces, for traits of local interest in two locations in the Ethiopian highlands. For each wheat variety, farmers provided quantitative evaluations of their preference for flowering time, spike morphology, tillering capacity and overall quality. A total of 10 agronomic and phenology traits were simultaneously measured on the same varieties, providing the means to compare them with farmer traits. The durum wheat varieties were genotyped for more than

²¹ CGIAR.

²² CGIAR.

80,000 single nucleotide polymorphisms markers, and the resulting data were used in a genome wide association study that resulted in a molecular dissection of smallholder farmers' choice criteria. A total of 124 putative quantitative trait loci (QTL) affecting farmer traits and 30 putative QTL affecting metric traits were found. The study showed that smallholder farmers' traditional knowledge can be associated with QTL for desired phenotypes. These results demonstrate that it is feasible and appropriate to involve farming communities to directly evaluate broad collections of genotypes using a selected set of previously agreed summary traits. The combination of participatory variety selection and modern plant breeding can not only speed up the genetic gain in breeding targeting smallholder farming systems but also lead to improved varieties more closely addressing smallholder farmers' needs.

*Example 17. Responding to a wheat blast outbreak in Bangladesh*²³

In 2016, the wheat crop in Bangladesh suffered from an outbreak of an aggressive fungus known as wheat blast. In some regions, losses reached 70 per cent. To rapidly determine the precise identity and likely origin of the outbreak pathogen, scientists applied field pathogenomics (transcriptome sequencing of symptomatic and asymptomatic leaf samples collected from infected wheat fields in Bangladesh). This method, based on sequencing technology, allows scientists to acquire data directly from field samples of pathogens, warning growers about new races of disease emerging on previously resistant varieties. To encourage more experts to use their resources and expertise to find a solution, all raw sequence data was immediately and openly released on the project's website.²⁴ Phylogenomic and population genomic analyses revealed that the Bangladesh wheat blast outbreak was likely caused by isolates belonging to the South American wheat-infecting lineage of *M. oryzae*. Data sharing drew together an international group of experts, who compared global data on blast pathogens and identified that the new Bangladesh pathogen was most likely introduced into Asia from South America (fungal pathogens are known to be transported across continents by wind as well as by trade). The result was that, in record time, the knowledge acquired to manage wheat blast in Brazil using disease-resistant cultivars and fungicides could be directly applied to treat a Bangladeshi epidemic. This highlights the need for intensive monitoring and surveillance of crop pathogens and also the power of making digital sequence information open and accessible in order to rapidly address the sustainability of food production.

Role of digital sequence information in substituting for “natural” products

*Example 18. Sustainable bioproduction of Taxol for chemotherapies*²⁵

The effort to develop a low-cost, sustainable source of paclitaxel provides a good example how digital sequence information can support sustainable use. Paclitaxel, currently sold under the brand name Taxol, is a chemotherapy medication used to treat a number of types of cancer. Discovered in 1971 in the bark of a tree, Pacific Yew (*Taxus brevifolia*), and approved for medical use in 1993, it is now recognized by the World Health Organization as one of the most effective and safe medicines needed in a health-care system. Clinical trials in the mid-1980s required the felling of thousands of trees and led to the source trees becoming an endangered species. Since then paclitaxel has been produced via semi-synthetic chemistry and plant cell culture. These advances have taken pressure off the Pacific Yew, reducing the threat of species loss. However, supply and price vary, and the drug frequently features on the Drug Shortages List compiled by the American Society of Health System Pharmacists. The biosynthetic pathway has now been mined from the *T. brevifolia* genome and, because these data are shared in public databases, a large number of international research groups are able to use the digital sequence information to work towards the monumentally difficult task of reprogramming species that are suitable for low-cost, large-scale production (e.g., yeast, bacteria) with the large and complex paclitaxel biosynthetic pathway. As well as providing the potential for meeting global demands through sustainable, low-cost production,

²³ Group of genomics experts.

²⁴ Open Wheat Blast: www.wheatblast.net.

²⁵ Group of genomics experts.

this method has the potential to enable bioproduction of novel chemical derivatives of the natural product – e.g., with reduced side-effects.

Digital sequence information for public health purposes

Example 19. Global spread of dysentery²⁶

There are four species of *Shigella*, and most cases of dysentery are caused by *S. flexneri*. Little was known about the genetic makeup of *S. flexneri* strains, which made tracking the spread of infections challenging and has significantly impeded public health efforts to control outbreaks. A total of 351 strains of *S. flexneri*, from accredited public health laboratories in South Africa, Bangladesh, France, Viet Nam and the Republic of Korea were sequenced at the Sanger Institute, with the samples themselves originating from outbreaks in Bangladesh, Pakistan, Algeria, Egypt, Haiti, Senegal, Guinea, Burkina Faso, Chad, Cameroon, Ivory Coast, the Republic of Korea, Madagascar, Haiti and the Dominican Republic. By sequencing these many different strains, the researchers were able to determine that unique *S. flexneri* strains occupy distinct geographic areas, do not spread but are able to persist over a long time. This indicates the bacteria does not travel over wide areas and is not the underlying cause of dysenteric pandemics. These findings strongly reinforce the importance of sanitation and the provision of clean drinking water as well as informing public health strategies for tackling persistent outbreaks.

Example 20. Placing genomic surveillance at the heart of viral epidemic response²⁷

In recent outbreaks of disease, such as Ebola and Middle East respiratory syndrome coronavirus, sequencing samples of the viral pathogen revealed critical insights into the origins of the infection and the evolution and transmission of the disease. For example, in the 2013-2016 West African Ebola outbreak, researchers collected and sequenced over 1,600 Ebola virus samples to understand how the virus was evolving and the factors responsible for its transmission. Analysis of the digital sequence information of the virus showed how different strains crossed borders and spread within countries. Access to a subset of this data during the outbreak led to border closures aimed at limiting the spread of the virus and suggests the potential value of sequencing to control future outbreaks, if it can be shared and analysed quickly enough to inform response efforts. However, the impact of these new sequencing technologies has yet to be realized, in part due to the complexities and time taken to ship samples for sequencing and the resulting delays in the production of digital sequence information. Wellcome Trust is funding a project to develop a field-deployable virus sequencing system and accompanying information-sharing platforms so that real-time viral genome sequencing can have a greater impact on the public health response for the next outbreak.

Example 21. Tracking resistance to artemisinin collaboration (TRAC) studies²⁸

Tracking and understanding the genetic lineage of resistance is vital for improving malaria containment and elimination, as well as patient treatment. TRAC I and II are international collaborations, coordinated by the Mahidol-Oxford Tropical Medicine Research Unit, to gather, share and analyse “real-time” genetic information on malaria drug resistance. The TRAC I study mapped the extent and severity of artemisinin resistance in South-East Asia and TRAC II will continue to map resistance and also assess the safety and efficacy of new artemisinin combination treatments. It is an international effort, as TRAC II has over 60 investigators from more than a dozen countries. The researchers post malaria parasite digital sequence information on TRAC and other databases, such as the WorldWide Antimalarial Resistance Network, so that other researchers can build on their findings. Rapid sharing is critical to help track and anticipate the geographic routes of drug resistance and inform national and regional patient treatment strategies to stop resistance to artemisinin spreading to other malaria endemic regions.

²⁶ Wellcome Trust and Wellcome Sanger Institute.

²⁷ Wellcome Trust and Wellcome Sanger Institute.

²⁸ Wellcome Trust and Wellcome Sanger Institute.

*Example 22. The Global Initiative on Sharing All Influenza Data*²⁹

The establishment of the Global Initiative on Sharing All Influenza Data (GISAID) as a new mechanism for encouraging the international sharing of data on influenza viruses was initially spurred by the global threat posed by the highly pathogenic avian influenza (H5N1) in the mid-2000s. The Initiative has been designed to overcome some of the challenges associated with the international sharing of virus data by being sensitive to the needs of stakeholders. It includes a data access agreement by which users must abide in order to have access to the data shared through the Initiative.

The core provisions of the agreement include that users: (a) will share their own data and allow other users to access it; (b) that they will not share or distribute data submitted directly to the GISAID sharing mechanism to non-GISAID servers or to individuals/institutions who are not registered GISAID users; (c) that they will credit the use of others' data in publications; (d) that they will make best efforts to collaborate with the originating laboratory and involve them in analyses and further research involving the data; (e) that they will analyse findings jointly; and (f) that they will maintain common access to technology derived from the data so that it can be used not only for research but also for the development of medical interventions, such as diagnostics, vaccines, or antivirals. According to the agreement, GISAID users thus have the right to develop a commercial product on the basis of data obtained through GISAID, but they may not impose any terms on the data itself (which remains the sole property of the contributor), and they must also seek to collaborate with the data contributors.

Limiting access – and even outright exclusion – of those who violate the terms of the access agreement remains a credible sanction, and one that has been used in the past. According to GISAID, the percentage of all active users whose access credentials to the GISAID platform had been revoked at the time of writing is around 0.16 per cent. At present, however, GISAID is only able to trace who is accessing information, not whether people are passing this information on to others. GISAID maintains that, if such data subsequently surfaces, they do have means to prove that someone has illegally obtained the data – meaning that data contributors who suspect violations can seek to pursue this through legal channels. Nor, of course, can there be ultimate guarantees that people will adhere to these rules when confronting all the pressures of a pandemic situation in the future, although the initiative has now developed a successful track record of navigating such situations.

II. DIGITAL SEQUENCE INFORMATION AND FAIR AND EQUITABLE BENEFIT-SHARING

Regulation of digital sequence information through prior informed consent and mutually agreed terms for access to the information itself

*Example 23. Development of a new consensus phytase to improve the nutritional value of animal feed*³⁰

It has been demonstrated that a “consensus protein” has the potential to be more thermostable than any of the individual protein sequences used in its design. For the design of a consensus protein (e.g. a consensus phytase), a (large) set of homologous sequences for the protein of interest (e.g., all known microbial phytase amino acid sequences) are aligned, and for each amino acid position, the most frequent amino acid is calculated and selected. Next, this consensus amino acid sequence is converted into a corresponding DNA sequence, and this DNA sequence is then synthesized chemically. The consensus protein is then expressed in a suitable host, and tested for its thermostability. If it is found to have the desired thermostability, the product will be developed for commercialization. This consensus approach has been applied to the design of many proteins and is nowadays an established protein design tool.

²⁹ Global Initiative on Sharing All Influenza Data.

³⁰ ICC.

Benefit-sharing through technology transfer and capacity-building

*Example 24. Partnerships and capacity development through CIMMYT-led “MasAgro Biodiversidad”*³¹

The CIMMYT-led “MasAgro Biodiversidad” (Seeds of Discovery – SeeD) Initiative has the goal of increasing effective and equitable use and benefit-sharing from maize and wheat genetic resources conserved in germplasm banks. A platform of publicly accessible germplasm, data, tools and services is being developed through public-private partnerships that contribute expertise and resources to the project.³² Capacity development, including graduate student thesis projects, technical workshops, visiting scientist projects, and publicly available software tools, forms the cornerstone of a strategy to enhance and extend project impacts, and provide an equitable framework for scientific innovation and benefit-sharing.

*Example 25. Free and accessible to all online learning platform to accelerate the development of improved maize and wheat varieties*³³

An online learning platform created in partnership among the Seeds of Discovery (SeeD) Initiative at CIMMYT (in Mexico), the National Institute of Agricultural Botany (NIAB) (Cambridge, United Kingdom) and Diversity Arrays Technology Pty (Canberra, Australia) comprises distance learning practical and theory modules about how to enhance the use of genetic diversity in wheat and maize. The online modules are available free of charge and to anyone who wishes to access them. The modules are complemented with videos that show how the modules may help prospective users to solve problems found in their research and that explain the aims and outputs of the Seeds of Discovery Initiative. The creation of the platform and modules was driven by the felt need of reaching out to a much larger number of researchers than through the limited spaces of face-to-face courses and workshops offered by SeeD on genetic diversity analysis, breeding and use of datasets and software tools. The modules are aimed at postgraduate students, researchers, crop breeders and university members. Initially developed in Spanish to respond to the capacity-building needs of users in Mexico and Latin America, it will be available in English to reach a wider spectrum of people interested in the characterization and use of genetic diversity.³⁴

*Example 26. Levelling the playing field on use of genomic digital information through CIMMYT’s SAGA*³⁵

The Genetic Analysis Service for Agriculture (known as SAGA from its Spanish name) of the Seeds of Discovery (SeeD) Initiative, funded by the Government of Mexico and run by CIMMYT since 2011, has used DArTseq technology to characterize genetically 100 per cent of the CIMMYT maize collection (approximately 29,000 accessions) and approximately 40 per cent of the 150,000 accessions in CIMMYT’s wheat collection. Public and private institutions such as UNAM (Mexico), INIFAP-Sinaloa (Mexico), CNRG (the Mexican Genetic Resources Center), CATIE (Costa Rica), ICARDA, IITA, and a Mexican private seed company have availed themselves of the genotyping services of the SeeD project to enhance their own research. Many Mexican researchers work with SeeD to address their specific programme needs (e.g. heat or drought tolerance, disease resistance) and opportunities (e.g. enhanced nutritional qualities, enhanced forage productivity) within their projects.

*Example 27. Building bioinformatics capacity in Africa*³⁶

Over the last decade, several international organizations and better-resourced African establishments (e.g., The African Society of Human Genetics and the African Society for Bioinformatics and

³¹ CGIAR.

³² See <http://seedsofdiscovery.org/es/catalogo/>

³³ CGIAR.

³⁴ See <http://seedsofdiscovery.org/new-online-learning-platform-offers-capacity-development-for-all/>

³⁵ CGIAR.

³⁶ Group of genomics experts.

Computational Biology) have focused efforts on building capacity in bioinformatics. The popularity of bioinformatics is due to its versatility and infrastructure requirements, and these efforts have been incredibly successful. For example, there is now an extensive Pan-African Bioinformatics network, H3ABioNet, comprising 32 bioinformatics research groups distributed among 15 African countries. In 2014, researchers in Kenya and South Africa led the sequencing and genome assembly of the tsetse fly, the vector of human African trypanosomiasis. Nigerian bioinformatics research groups have applied bioinformatics techniques to a number of domestic issues, including malaria, while Ghanaian bioinformaticians have contributed to and led projects to analyse the sequence diversity of a wide range of human and plant pathogens and crop species. These represent a considerable and rapidly expanding knowledge base in genomics and bioinformatics, poised to maximize the use of digital sequence information and other bioinformatic resources. Without access to open sequence databases and software, these knowledge exchanges and capacity-building exercises would not have been possible.

*Example 28. Fast and frugal lab tools for developing countries*³⁷

The synthetic biology community aims to make software, hardware and wetware for biology and biotechnology cheaper, easier to operate and internationally compatible. This is partly to facilitate scaling for industry, but this so-called “frugal science” movement also aims to make open, cheap and easy tools specifically to facilitate the adoption of biotechnologies that utilize digital sequence information in developing countries. Efforts in the development of low-cost, open-source laboratory hardware have been particularly successful and include molecular biology essentials, e.g., OpenPCR and Biopette, the latter deployed in a range of workshops across several countries by TReND (Teaching and Research in (Neuro)science for Development) in Africa. Open-source versions of microscopes (e.g., The FlyPi and The Waterscope) and even robotics e.g., OpenTrons are enabling researchers in less resourced institutions to thrive. For example, community platforms, such as Hackteria, enable researchers in developing countries to equip biology labs from OpenSource Hardware at less than 10 per cent of the commercial price. Similar efforts are under way to equip researchers with open-source wetware (e.g., The BioBricks Foundation and OpenPlant), including molecular (DNA-based) tools to stimulate innovation and entrepreneurship.

Digital sequence information, databases and benefit-sharing

*Example 29. Global Biodiversity Information Facility*³⁸

An example of a large database that has a multinational “ownership” is the Global Biodiversity Information Facility (GBIF). GBIF has signed partnership agreements with many organizations within the governmental, academic or private sector. Voting participants of the Governing Board of GBIF include developing and developed countries. GBIF’s primary task is to create and manage a centralized portal for providing free and open online access to biodiversity data by envisioning a global biodiversity information commons. GBIF is “a multilateral initiative [...] to promote, coordinate, design and implement the compilation, linking, standardization, digitization and global dissemination of the world’s biodiversity data, within an appropriate framework of rights [...] that will establish and support a distributed information system that will enable users to access and utilize quantities of existing and new biodiversity data”.³⁹ It is considered vital for “delivering biodiversity knowledge in the information age”.⁴⁰ However, sustainability of the benefits of access to biodiversity data on the GBIF portal depends on the willingness of the voting participants of GBIF (all of which are countries) to keep up its funding.

³⁷ Group of genomics experts.

³⁸ CETAF.

³⁹ Reichman, J.H., Uhler, P. and Dedeurwaerdere, T. 2016. Governing Digitally Integrated Genetic Resources, Data and Literature: Global Intellectual Property Strategies for a Redesigned Microbial Research Commons. Cambridge University Press, Cambridge.

⁴⁰ Ibid.

*Example 30. Global Genome Biodiversity Network and the GGBN Data Standard*⁴¹

The Global Genome Biodiversity Network (GGBN) is an international network of institutions that share an interest in long-term preservation of genomic samples representing the diversity of non-human life on Earth. GGBN makes the DNA and tissue collections of its members discoverable for research through a networked community of biodiversity biobanks. In doing so, GGBN provides trusted and transparent access to genomic samples for all through an access and benefit-sharing framework. GGBN serves its members by sharing a code of conduct and best practice tools that are compliant with the Nagoya Protocol as part of a responsible approach to making samples discoverable and accessible subject to any restrictions on their use, and when known, links to associated genetic sequence data.

GGBN does not itself produce sequence data, but the GGBN Data Portal (<http://www.ggbn.org>) cross-links to sequencing portals – the International Nucleotide Sequence Database Consortium (INSDC)⁴² and the Barcode of Life Data System (BOLD)⁴³ – to establish links between the stored samples and the sequence accession numbers related to the samples. Additionally, the GGBN Data Portal stores information about the samples permit status. Thus, the GGBN Data Portal increases the visibility of samples and availability as well as any restrictions on their use. GGBN does not conduct research on or produce genetic sequence data.

GGBN has developed the GGBN Data Standard to complement existing biodiversity standards, such as Darwin Core or ABCD (the Access to Biological Collections Data schema). The GGBN Data Standard is intended to provide a platform based on a documented agreement to promote the efficient sharing and usage of genomic sample material and associated specimen information in a consistent and open manner. It is a set of terms and controlled vocabularies designed to represent any and all sample facts. This also includes vocabulary for permits and loans according to the requirements of the Nagoya Protocol. GGBN is working on a tool that enables tracking of parent and offspring use of samples.

GGBN proposes the GGBN Data Standard as the global biodiversity data exchange standard for fulfilling the Nagoya Protocol and is already in contact with INSDC, BOLD and GBIF to enable support of this standard in other global portals. GGBN seeks to make sure that all samples created since the ratification of the Nagoya Protocol will provide permit information by the end of 2020. Furthermore, work is ongoing on automated submission pipelines to INSDC, which includes permit information. This is an example of transparency and accountability regarding permits.

Benefits shared from use of digital sequence information

*Example 31. The promise of maize provitamin A biofortification through discovery of natural genetic variation*⁴⁴

In a project funded by, among others, the United States Department of Agriculture, Harvest Plus, and the Borlaug Fellowship, researchers from CIMMYT and several other educational and research institutions have been involved in efforts to develop maize crops with enhanced levels of the precursors to vitamin A. Maize is an important subsistence crop in sub-Saharan Africa, where vitamin A deficiency is common and can lead to blindness and increased susceptibility to infections. Researchers have been working to identify natural variation in the amount of carotenoids produced in kernels of maize through association analysis, linkage mapping, expression analysis and mutagenesis. Natural variation at a lycopene cyclase locus was shown to affect the flux down certain carotenoid pathways, further affecting the development of provitamin A compounds. The selection of the alleles with molecular markers identified through sequence

⁴¹ GGBN.

⁴² <http://www.insdc.org>

⁴³ <http://www.barcodinglife.org>

⁴⁴ CGIAR.

data will enable breeders to produce maize grain with enhanced vitamin A levels, potentially having a transformational effect to improve the nutritional status of millions in sub-Saharan Africa.
