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

A Fact-Finding and Scoping Study Undertaken for the Secretariat of the Convention on Biological Diversity

Sarah A. Laird and Rachel P. Wynberg,

with contributions from Arash Iranzadeh and Anna Sliva Kooser

9 November 2017

DRAFT FOR PEER REVIEW

Contents

Acronym Table

ABS	Access and Benefit Sharing
AHTEG	Ad Hoc Technical Expert Group
ARK	Archival Resource Key
BiOS	Biological Innovation for Open Society
BLAST	Basic Local Alignment Search Tool
CETAF	Consortium of European Taxonomic Facilities
CITES	Convention on International Trade in Endangered Species of Fauna and Flora
COI	Cytochrome c oxidase I
BOLD	Barcode of Life Database
CBD	Convention on Biological Diversity
CBOL	Consortium for the Barcode of Life
cDNA	Complementary DNA
CGRFA	Commission for Genetic Resources for Food and Agriculture
DAA	Data Access Agreement
DICOM	Digital Imaging and Communications in Medicine
DIYbio	Do-it-Yourself Bio
DNA	Deoxyribonucleic Acid
DOI	Digital Object Identifier System
DSI	Digital Sequence Information
eDNA	Environmental Genomics
EMBL-EBI	EMBL European Bioinformatics Institute
EPIC	Epigenomics of Plants International Consortium
FAO	Food and Agriculture Organization of the United Nations
GBIF	Global Biodiversity Information Facility
GCM	Global Catalogue of Microorganisms
GGBN	Global Genome Biodiversity Network
GISAID	Global Initiative on Sharing All Influenza Data
GMI	Global Microbial Identifier
GO	Gene Ontology
GOC	GO Consortium
GR	Genetic Resource
GSC	Genomic Standards Consortium
GSD	Genetic Sequence Data
GWAS	Genome Wide Association Studies
Handle	Handle System

HIV	Human Immunodeficiency Virus
iBOL	International Barcode of Life Project
iGEM	International Genetically Engineered Machine
IMCAS	Institute of Microbiology, Chinese Academy of Sciences
INSD	International Nucleotide Sequence Database
INSDC	International Nucleotide Sequence Database Collaboration
IP	Intellectual Property
IPD	Immuno Polymorphism
ITPGRFA	International Treaty for Plant Genetic Resources for Food and Agriculture
IVTM	Influenza Virus Traceability Mechanism
JCVI	J Craig Venter Institute
LSID	Life Science Identifiers
MAT	Mutually Agreed Terms
MGD	Mouse Genome Database
MIxS	Minimum Information about any (x) Sequence
MHC	Major Histocompatibility Complex
MIGS	Minimum Information about a Genome Sequence
MOU	Memorandum of Understanding
MTA	Material Transfer Agreement
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NGS	Next Generation Sequencing
NIST	National Institute of Standards and Technology (US)
NP	Nagoya Protocol
OBO	Open Biomedical Ontologies
OECD	Organisation for Economic Cooperation and Development
OSDD	Open Source Drug Discovery
PatSeq	Patent Sequence
PC	Partnership Contribution
PCR	Polymerase Chain Reaction
PIC	Prior Informed Consent
PIP	Pandemic Influenza Preparedness
PURL	Persistent Uniform Resource Locator
QTL	Quantitative Trait Locus
R&D	Research and Development
RNA	Ribonucleic Acid

SBOL	Synthetic Biology Open Language
SGD	*Saccharomyces* Genome Database
sMTA	Standard Material Transfer Agreement
SNP	Single Nucleotide Polymorphisms
SO	Sequence Ontology
SRA	Sequence Read Archive
TWEG	Technical Expert Working Group
UK	United Kingdom
UPOV	International Convention for the Protection of New Varieties of Plants
URN	Uniform Resource Name
US	United States
WFCC	World Federation for Culture Collections
WHO	World Health Organization
WHO GISRS	WHO Global Influenza Surveillance and Response System

Executive Summary

Background to the Study

In December 2016, the 13th meeting of the Conference of the Parties (COP) to the Convention on Biological Diversity (CBD) and the second meeting of the Conference of the Parties serving as the meeting of the Parties to the Nagoya Protocol on Access and Benefit-sharing adopted decisions to address the cross-cutting issue of "digital sequence information on genetic resources" (decisions XIII/16 and NP-2/14, respectively). The decisions included formation of an Ad Hoc Technical Expert Group (AHTEG) on Digital Sequence Information (DSI) on Genetic Resources, and an invitation to governments, indigenous peoples and local communities, and relevant organizations and stakeholders to submit views and information on the potential implications of the use of digital sequence information for the three objectives of the CBD and the Nagoya Protocol.

In addition, the COP requested the Executive Secretary of the CBD to commission a fact-finding and scoping study to clarify terminology and concepts and to assess the extent and the terms and conditions of the use of digital sequence information in the context of the CBD and the Nagoya Protocol. This study is the result of that decision. The present study references and in some cases also complements work on this issue undertaken as part of other international policy processes. These include the UN General Assembly process on biodiversity in areas beyond national jurisdiction; the World Health Organization's (WHO) Pandemic Influenza Preparedness (PIP) Framework; the International Treaty for Plant Genetic Resources for Food and Agriculture (ITPGRFA); and the Commission for Genetic Resources for Food and Agriculture (CGRFA).

The research for this study took place over four months, and included a literature review, as well as semi-structured interviews with academic researchers, industry representatives, database managers, civil society groups, policy makers, and others. In total, 55 individuals from 17 countries were interviewed.

Overview of the Report

Seven sections comprise this report:

- Section 1 introduces the study and its terms of reference;
- Section 2 explores the term "digital sequence information";
- Section 3 reviews the diverse and rapidly evolving ways digital sequence information is used in academic research and industry today;
- Section 4 examines how digital sequence information is accessed, stored, and managed, including via public and specialized databases, and registries of standard parts;
- Section 5 explores the generation of "new" digital sequence information from physical samples derived from field and *ex situ* collections;
- Section 6 reviews tools used to manage digital sequence information accessed through databases or registries, including conditions of use notices, click through agreements, and open source and user agreements;
- Section 7 reviews ways that digital sequence information contributes to the conservation and sustainable use of biodiversity, and some of the conservation impacts of technologies that make

use of sequence information;

- Section 8 explores the implications of digital sequence information for fair and equitable benefit sharing, including opportunities and challenges that arise.

Terminology

The term "*digital sequence information*" is used in decisions CBD XIII/16 and Nagoya Protocol (NP) 2/14, but has grown from the CBD policy process. Terms more commonly employed by the scientific community and databases include *genetic sequence data, nucleotide sequence data, nucleotide sequence information*, and *genetic sequences*. Differences in terminology in scientific circles reflect differences in the material referred to, as well as the speed and transformative nature of technological change today, which make it difficult to harmonize terminology. In ABS policy discussions, differences in terminology often reflect divergent views of what falls within the scope of the Nagoya Protocol and national laws.

Terminology also varies between international policy processes. The ITPGRFA elected to use the term "*sequence data*" in its recently commissioned scoping study on synthetic biology. The UN General Assembly's policy process on marine biodiversity in areas beyond national jurisdiction began with the term *resources in silico* but has moved to *digital sequence data*. The WHO PIP Framework uses the term *genetic sequence data,* which they define as: "The order of nucleotides found in a molecule of DNA or RNA... contain[ing] the genetic information that determines the biological characteristics of an organism or a virus". Steps have been taken to harmonize terminology across international policy processes, but this has yet to take hold. For the purpose of this study, we use the terms fluidly, but for the most part, use the term digital sequence information, in keeping with decision XIII/16.

The Use of Digital Sequence Information

Digital sequence information is the product of sequencing technologies that have become faster, cheaper and more accurate in recent years. It may be natural or synthetic, identical to sequences found in nature, or designed, mutated, or degenerated. Digital sequence information permeates nearly every branch of the life sciences and modern biology today, allowing for computational analyses and simulations that are significantly cheaper and quicker than biological experiments run in a conventional laboratory. It contributes to understanding the molecular basis of life, evolution, and how genes might be manipulated to provide new therapies and cures for disease, industrial products, energy sources, chemicals, and other products. It also plays an important role in deepening knowledge about biodiversity, identifying and mitigating risks to threatened species, enhancing our ability to track illegal trade, identifying species and the geographic origins of products, and assisting with biodiversity planning and conservation management.

Genomic technologies used to study genes and their functions generate an unprecedented amount of information, making this an intensely data-rich field. As a result, bioinformatics – the collection, classification, storage and analysis of complex biological data – has grown alongside genomic technologies in order to store, retrieve, and analyze these vast and growing amounts of information. Advances in sequencing and bioinformatics have in turn spawned metagenomics, also known as environmental genomics, in which researchers sequence and analyze the genomes of species found in an environmental sample, usually from soil or water.

These technological and scientific advances have changed the way researchers work, making possible dynamic knowledge hubs, and diffuse scientific collaborations. They take place in an increasingly globalized research context in which collaborative and inter-disciplinary approaches are now the norm. Diverse networks of researchers from industry, government, academia, and community laboratories commonly span the globe in a system of "open innovation" in which users add incremental value through data and knowledge sharing along a chain that involves multiple databases and gene sequences. Distinctions between academic, governmental, or industry research using genetic sequences have become increasingly blurred, as have those between different industrial sectors.

Synthetic biology is one part of this rapidly transforming scientific landscape, and has wide application across sectors. As synthetic biology technologies have become cheaper and more widely accessible, an explosion of small-scale, publicly accessible community laboratories, DIY (do-it-yourself) bio, and open science collaborations have grown up. Inspired by the open source software movement, groups exchange and use digital sequence information within an open source framework that seeks to develop products and technologies while 'democratizing' science, and even the means of production.

Synthetic biology and other research approaches that make use of sequence information are also used in commercial research and development, including within the industrial biotechnology, pharmaceutical, and agriculture industries. For example, within industrial biotechnology, genes might be combined from a number of different organisms into an artificial DNA construct, and incorporated into a host organism which produces bio-based products such as chemicals, food and feed, detergents, pulp and paper, electronics, automotive, packaging, cosmetics, bioprocessing catalysts, textiles and bioenergy.

In drug discovery and development, pharmaceutical companies are also making use of the cheaper and more rapid sequencing technologies, and advances in bioinformatics. For example, predictive biomarkers allow trials to be smaller and potentially reach significance faster, and 'synthetic' candidate vaccine viruses are being generated using particular genetic sequences. Genomic information also plays a role in conservation, pre-breeding and breeding within agriculture, most commonly when plant genomic information is mined to identify sequences and genes of interest.

How digital sequence information is accessed, stored, and managed

Digital sequence information is accessed from a range of private, governmental, and research institution collections, companies that synthesize sequence information, journal articles, supplementary files linked to published papers, and from public databases and genetic parts registries. In this section, we review public and specialty databases and registries of parts, which are the largest repositories of digital sequence information.

Public databases

In the late 1970s, when DNA sequence data began to accumulate in the scientific literature, public databases were set up to store and organize sequences, and it soon became best scientific practice to publish new genetic sequences in sequence databases. There are now more than 1,500 publicly accessible biological databases, organized based on heterogeneity, data type, scope and curation. The largest databases are part of the International Nucleotide Sequence Database Collaboration (INSDC). This is comprised of three global partners:

- The European Nucleotide Archive), based at the EMBL European Bioinformatics Institute (EMBL-EBI) in Cambridge, UK;
- GenBank, based at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, USA; and
- DNA Data Bank of Japan (DDBJ), based at the National Institute for Genetics in Mishima, Japan.

These partners, funded by their respective governments, "capture, preserve, share and exchange a comprehensive collection of nucleotide sequence and associated information". A common means of using these databases is to run a Basic Local Alignment Search Tool (BLAST) search, which finds regions of local similarity between query sequences and those in the databases by searching every record. This means that all sequences in a database are regularly used as part of these searches. The INSDC's policy emphasizes free, unrestricted access to all of the data records in their database.

The amount of data flowing into databases is exponentially increasing, as is the use of that data. The INSDC databases have almost doubled in size in the last few years, and the EMBL-EBI search engine, for example, runs on average 12.6 million jobs every month. The number of bases and sequences, individuals and species sequenced, and the depth of genomic coverage obtained per sample are all increasing. Journals increasingly require that genetic sequence data be deposited in these public databanks as a condition of publication, and an INSDC accession number is often necessary to publish research. Databases work with publishers to ensure a flow of data into repositories for release before, or at the time of, publication, often creating embargo periods prior to publication during which data remains confidential.

There have been multiple efforts to standardize and unify the terminology and data standards of databases. In recent years, this has increasingly included information on the environmental context and locations from which organisms originate. Earlier records, however, rarely contain metadata on the origins of sequences, and contemporary records are not always complete.

Registries of Standard Parts

Another common source of genetic sequence data is in shared repositories like the Registry of Standard Biological Parts, and the Inventory of Composable Elements. "Parts" are DNA sequences that encode for a specific biological function, and that can be combined to create new, longer and more complex parts. The Registry creates a library of standard parts that have been tested, characterized and organized (each with an identification code), making it easier for researchers to share parts and collaborate. The Registry is available for use by researchers around the world, many of whom contribute new parts back to the Registry following validation. The Registry currently holds over 20,000 documented parts.

Generation of 'new' digital sequence information from physical samples

Most research is based on sequences accessed through databases or parts registries, but some groups sequence and analyze physical samples from field collections, citizen science sourcing programs, or *ex situ* collections.

Field collections of physical samples are a much smaller part of research strategies in high tech industries than they were twenty years ago. Today, few companies undertake regular and systematic

collections, although there are exceptions. Academic groups continue to have an interest in physical samples, in particular the wide diversity of microbial species that can now be studied using metagenomic sequencing technologies. Interest in organisms from areas of high species diversity, extreme environments, and unique ecological niches also persists.

A few companies and research institutes continue to collect field samples, most of which are then sequenced. Citizen science programs that solicit samples from around the world as part of efforts to understand biological and genetic diversity, particularly of microorganisms, are increasingly common. In these programs, samples are shared in exchange for data analysis for contributors, and research programs receiving samples avoid the cost and time of field collecting expeditions. As a result, the scope of these efforts can be enormous, generating massive quantities of data and covering vast geographic distances.

A wide and varied range of *ex situ* collections are held by public entities, non-profits, scientific research institutions like botanical gardens and natural history museums, culture collections, universities, companies and others. Many of these groups are digitizing their collections, which might include producing digital images and sharing data about specimen collection, as well as producing digital sequence information from physical samples.

Although the science is moving away from physical material, its use is still necessary and important for most research projects. Physical samples provide information a sequence alone cannot, including the relationship of genotype to phenotype, and interactions between organisms and their environment. Discovering things that are completely unknown from a genome alone is still largely in the future.

A significant technological advance with relevance for access and benefit sharing is the MinION, the world's "first and only nanopore DNA sequencer", which is portable and low-cost and designed to make biological analyses widely available. The day has arrived when individuals can easily and affordably sequence genes from physical material anywhere in the world, and send it via the internet to researchers, databases, foundries, and other institutions in regions far from the site of collection.

At the other end of the process, advances in automation are making it simpler and cheaper to synthesize DNA parts. A digital-to-biological-converter has been developed to produce functional biologics in an automated fashion from digitally transmitted DNA sequences, in particular DNA templates, RNA molecules, proteins and viral particles. Synthesizers can now churn out strings of several thousand base pairs rather than a few hundred, at a fraction of the cost of even a few years ago. The technology is moving so quickly that it will soon be possible for most researchers to inexpensively synthesize DNA in their laboratory.

Tools to Manage Digital Sequence Information: Conditions of use notices and agreements

A range of approaches attach conditions to the use of digital sequence information. These include notifications on databases and websites, conditions of use notices, click through agreements, open source Material Transfer Agreements (MTAs), and user agreements. In most approaches, negotiation of an agreement between a commercial user and a contributor of sequence information is envisioned at some point in the future, once a commercial use has been established.

Conditions of use notices and click-through agreements

A number of websites and databases include conditions of use notices that might include asserting that downloaded digital sequence information is the patrimony of the country of collection, that users of the information agree to acknowledge the country of origin in any publication, or that national focal points should be contacted if sequence information is used for commercial purposes.

One step beyond a conditions of use notice is a click-wrap, or click-through, agreement that requires users to click their assent to certain terms in order to gain access to the website or database. These are commonly used by software companies. Concerns about both conditions of use notices and click-through agreements include that users do not understand what they are agreeing to, do not read the fine print, and that these are not legally enforceable.

Open source and user agreements

Open source agreements are designed to promote innovation and avoid the high transaction and legal costs associated with traditional MTAs or other forms of licensing agreements. They are intended to facilitate the free exchange of information, technology and materials, and support increasingly networked and collaborative research. Contributors may request attribution and reporting for materials, but materials are immune from the assertion of intellectual property, and may be transferred between researchers within the open source community, whether academic or commercial. Some agreements require that anything developed from materials be shared with the community of contributors and users, but others do not, and none include royalties for the use of materials or methods.

User agreements, often with similar features to open source agreements, are employed by some targeted databases and research institutions. For example, the Global Initiative on Sharing All Influenza Data (GISAID) has developed a Database Access Agreement (DAA) that issues licenses for the use of data and includes benefit sharing. The J Craig Venter Institute (JCVI) has negotiated more involved Memoranda of Understanding (MOUs) that address digital sequence information as part of marine microbe collections, some inside territorial waters. JCVI, along with many other academic and research groups undertaking field collections, include language in their agreements clarifying that sequence information will be uploaded to public databases.

Digital Sequence Information and the Conservation and Sustainable Use of Biodiversity

Digital sequence information is a critical tool and resource for the conservation and sustainable use of biodiversity. Understanding the Earth's biodiversity and its dynamic changes relies heavily on access to appropriate information, yet our knowledge of some of the most basic aspects of biodiversity remains inadequate. Increasingly, cost-effective genetic sequence-based diagnostic techniques are part of the toolkit of biodiversity researchers. Examples include:

- DNA barcodes, used extensively to identify species;
- the characterization of national biodiversity;
- the use of genetic sequence data in taxonomy, especially in cases where morphological identification is difficult;
- understanding genetic variability in populations;

- analyzing relationships between populations, and thus minimizing further genetic loss in endangered populations;
- identifying invasive alien species or pests;
- understanding pollinators; and
- monitoring environmental change, including developing models about the impacts of climate change on species and their distribution.

Genetic sequence analysis is also a powerful tool for implementation of CITES (the Convention on International Trade in Endangered Species of Fauna and Flora) and related agreements and supports the fight against illegal logging and seafood fraud, including the mislabeling of products. Databases containing sequence data comprise reference libraries for comparing specimens and samples that are confiscated by law enforcement officials. For example, using DNA sequence markers, it is possible to distinguish between wild and cultivated species, identify the source of samples thought to be from threatened or endangered species, or monitor processed products which otherwise might be difficult to identify.

Identifying and characterizing genetic resources also contributes to the development of new crops that are resilient to climate change, pathogens, soil degradation, salinity and drought. The application of digital sequence information is also invaluable in molecular epidemiology, and helps to trace the origin and evolution of pathogens in emergency situations.

In addition to its valuable role in conservation science, planning and management, digital sequence information is integral to technologies and applications, like synthetic biology, that have potentially positive and adverse effects on biodiversity. Possible positive impacts include reduced consumption of fossil fuels by relying on biological processes that use renewable raw materials to produce biofuels and cleaner, more efficient manufacturing processes that pollute less and reduce waste. They might also include microorganisms designed for bioremediation or new manufacturing processes to produce chemicals, plastics, and drug-precursors currently extracted unsustainably from natural resources or synthesized from petrochemicals. In the future, synthetic biology could also potentially be used to control invasive species, tackle threats to endangered species, restore habitats through modification of genomes, or even recreate extinct species.

Although not explored in this study, some of these technologies raise environmental, social justice and ethical concerns which are currently under discussion in the synthetic biology AHTEG. For example, there are concerns about the unsustainable production of the biomass that feeds biological factories producing biofuels, chemicals, plastics, pharmaceuticals and other products. The pressure placed on land, forests, and so-called marginal lands for biomass production, linked to land grabs that impact indigenous peoples and local communities and displace food crops and traditional agriculture, has raised significant social and environmental concerns. The replacement of cash crops with new biotechnology products also has potential impacts on small farmer livelihoods. Concerns have also been expressed about the unpredictable ecological impacts of gene drives or invasive species toxic to other non-target organisms, or which damage native genetic diversity.

Digital Sequence Information, Fair and Equitable Benefit Sharing, and the Nagoya Protocol

It is difficult to generalize about benefits that might result from the use of digital sequence information given the rapid and transformative nature of the science and technology associated with sequences.

However, a number of potential benefits, as well as challenges to benefit sharing, have emerged over the course of this research. In addition to more speculative monetary benefits that might accrue from the system that manages, disseminates, and uses digital sequence information, new forms of non-monetary benefit sharing have emerged, in keeping with those identified in the Annex to the Nagoya Protocol. These include wider access to databases, knowledge and technology; technology transfer, capacity-building, and collaboration; and research directed at priority public needs.

Wider accessibility of databases, knowledge, and technology

An important form of benefit sharing is access to publicly available databases. Tax payers in the countries and regions that undertake the bulk of research using digital sequence information (the US, Europe and Japan), provide funds, expertise, and technological capacity to store, analyze and manage data within the public databases. Most countries do not have the funds or capacity to manage comparable systems, and so the INSDC databases serve as a resource for the global community. Every contributor of data or research results from around the world adds value to a shared global system, and in return gains access to the greater value of the collection. In addition, these databases house information, and provide analyses, on global biodiversity, and serve as an important resource for biodiversity conservation and sustainable use. However, some consider access to databases and technology an insufficient benefit, involving a loss of control over national patrimony. Furthermore, countries rich in biodiversity may lack sufficient molecular research capacity or biotechnology infrastructure to make use of global database systems.

Benefit sharing is also impacted by the different approaches taken to access bulk sequence information held in databases. The two main approaches include *open access* or *public domain* (free and unrestricted access), and *open source* (in which some conditions attach to access). The open access approach allows the free and unencumbered use of digital sequence information to fuel innovation and scientific research. The open source approach ensures smaller groups and individuals are not locked out of these innovations and technologies, can attach conditions to the use of data to ensure wider forms of benefit sharing, and might involve user agreements or MTAs. Although proponents of these approaches to access differ in their view of how to ensure the 'greatest good', both support making as much data publicly available as possible, for easy use by a wide range of researchers across the globe.

Technology transfer, capacity-building, and collaboration

Capacity development and research collaborations present a significant opportunity for benefit sharing. In a similar way to conventional biodiscovery, such benefits growing from the use of digital sequence information may outweigh any potential financial benefits over time. The nature of research collaborations associated with sequence information can be quite different from those undertaken for biodiscovery, however. They might occur through cloud laboratoriews, involve the sharing of software, materials and technology, the provision of samples in exchange for sequencing and analysis, and other exchanges that do not include bi-lateral agreements, or perhaps even direct interaction between groups and individuals.

Research directed at priority public needs

Open science non-profit networks that share knowledge, technology and materials see the provision of these benefits as significant, but also view the broader research collaborations they spawn as contributing benefits to humankind. These collaborations address critical healthcare, environmental,

food security and other challenges we face today. Much of this research is also intended to address the needs of marginalized or under-served communities around the world.

Monetary benefits

Monetary benefits growing from the use of digital sequence information are largely speculative to date, and are potentially complex due to challenges in identifying provenance and the value of any given sequence or part. The negotiation of monetary benefits through database and registry conditions of use notices, MTAs, licenses and user agreements, is generally deferred to a point in the future when a commercial product has been developed, although as noted most open source agreements eschew monetary benefits. The practicalities of implementation remain undeveloped, however.

Some have suggested a standard access fee, or subscription, in which users pay a small charge for accessing a sequence, or an annual subscription. Given the blurring boundaries between commercial and non-commercial user, all might gain access on the same terms. Most database managers and researchers are opposed to any fee-based approach, however, given the significant cost and potential bureaucracy associated with creating a payment system and monitoring use. There is also concern that a fee-based system might isolate data or reduce the effectiveness of databases. As a result of these difficulties, many have suggested the establishment of a global fund to address benefit sharing from public databases. Experience from funds established under the ITPGRFA and the WHO PIP Framework may provide relevant lessons in this regard.

Determining value

The challenges of determining the value of digital sequence information are especially intractable. For example, products, processes and technologies growing from digital sequence information might involve genes from multiple countries and organisms combined together to create new biosynthetic pathways. Additionally, homologous, or identical, sequences vital to life, and in which natural selection has eliminated mutations, might be found in different organisms around the world. This means that if companies cannot acquire legal certainty for a sequence of interest from one country, they can search for, and often find, the sequence in another country. Further complicating matters is that sequence information is regularly modified and can be re-used indefinitely, raising questions about whether benefits attach to each transaction, or if there is a cut-off point after which benefit sharing does not apply. Additionally, the value of digital sequence information is often found in the aggregate, rather than an individual sequence, when it is part of a larger collection of sequences within databases against which searches and analyses are run. Finally, the commercial applications of sequence information are so enormously varied, and so rapidly changing, it is extremely challenging to characterize the utilization, and commercial value, of sequences.

Identification challenges

A range of challenges for benefit sharing are linked to the identification of contributors, users and the provenance of sequences.

Identification of contributors and users of digital sequence information. The bulk of digital sequence information is accessed through public databases, which do not require contributors or users to register or log in, agree to terms and conditions, or sign user agreements. Internal policies, and the governments that fund the databases, require that such databases do not erect barriers to free access, or apply

conditions to their use; this might be understood to include ABS conditions, and user and contributor identifications. However, many of the hundreds of specialized sequence databases directed to particular organisms, gene groupings, or diseases have developed policies and regulations, including the protection of personal privacy and confidentiality. One example is the GISAID Database Access Agreement that is free and open to anyone who positively identifies themselves and agrees to respect the rights of contributors. Open source agreements similarly require that contributors and users identify themselves as part of joining a community of researchers. Unique identifiers for researchers have also been proposed as a way to support ABS; these follow researchers through their careers, and link to publications. Unique identifiers could also potentially link to sequence data that is deposited in or accessed from databases.

Identification of the provenance of digital sequence information. There are increasing efforts to better link original physical material with digital sequence information, including metadata on the location of specimen collections. Many in the database and research community support inclusion of the provenance of digital sequence information, which is important for science, and might also support benefit sharing. A number of groups holding specimens are working to link sources, physical samples, and international databases. However, there are concerns about how effectively identification can work for sequence information, since sequences from the same species from the same habitat might differ due to natural mutations over very short periods of time. An additional challenge for identifying digital sequence information is that it is not immediately recognizable as belonging to a particular source, particularly as it undergoes modification.

Monitoring the Utilization of Digital Sequence Information

Monitoring is critical for effective benefit sharing, yet genetic sequences are far more difficult to monitor than physical genetic resources. These challenges increase over time as sequences pass through multiple hands, are modified, and the unique identity of a sequence erodes. As noted, a number of groups are working to identify provenance, and strengthen links between samples and sequences. These include the INSDC and other databases, ontology and standards organizations, and a number of governments. Some groups have tried 'watermarking' a DNA sequence in a non-coding region of DNA, while the Global Genome Biodiversity Network (GGBN) Data Standard is working on ways to share and use genomic sample material and associated specimen information as part of a monitoring system. Others are adapting national permitting systems to facilitate monitoring by giving each permit a unique identifier that would accompany material through the research process, including after it is sequenced and uploaded to databases.

Some are skeptical of the potential to monitor digital sequence information in any meaningful way, and express concern about the management, bureaucracy and expense involved in adding layers of legal documents and information to databases. It has been suggested that the separation of legal and scientific databases could help to address this concern. For example, scientific databases that hold sequence information could be separate from, but linked to, legal databases that are managed by governments and which contain permits and agreements associated with data.

Distinguishing between non-commercial and commercial research.

The lines between academic and commercial research have grown increasingly blurred in recent decades, as academic and government researchers partner with industry. Additionally, sequences move

fluidly between commercial and non-commercial institutions, and once uploaded to public databases are available for all to use. When genetic resources or digital sequence information are accessed, it is also not always clear how the material and information will be used in the future. For example, samples or sequences might be accessed under academic research terms, uploaded onto databases, and eventually used commercially, potentially by multiple different users, without the original providers aware of or involved in this process.

Conclusion

Digital sequence information is clearly a critical resource and tool for the conservation and sustainable use of biodiversity. The use of this information through transformative science and technologies also creates significant opportunities for non-monetary, and possibly monetary, forms of benefit sharing. There are, however, a range of challenges to realizing many of these benefits, linked in part to the difficulties of monitoring and identifying contributors, users and the provenance of sequences; the problems of determining value; and the increasingly grey area between non-commercial and commercial research.

It behooves ABS policy makers to stay abreast of the profound developments shaping research today. Sequencing platforms have become faster, cheaper and more accurate in recent years, producing massive quantities of sequence information. Researchers can now edit and synthesize genes. In the last year, new affordable and portable devices allow researchers to sequence physical samples, and then upload them to the internet or databases. Physical samples are still of interest to researchers, but their role in the research and commercialization process is changing, and the future is unclear.

Paralleling dramatic changes in science and technology are developments in the institutional, legal and social context of research. These include new, open and multi-party collaborations and diffuse research networks. Such collaborations are typically underpinned by a philosophy supporting unencumbered and free exchange of materials and technology, often as a way of serving the greatest public good, and to avoid intellectual property and transaction costs. New and significant benefits result from these innovative approaches, but use novel forms of benefit sharing that have not traditionally featured in ABS agreements. It might be that the strengths of ABS, open science, and other approaches could be combined in pioneering and inventive ways to develop flexible and adaptive policies that ensure benefits for the global community from the use of digital sequence information, including the important role it plays in the conservation and sustainable use of biodiversity.

1. Introduction

In December 2016 at the 13th meeting of the Conference of the Parties to the Convention on Biological Diversity (CBD), and the second meeting of the Conference of the Parties serving as the meeting of the Nagoya Protocol on Access and Benefit-sharing, adopted decisions to address the cross-cutting issue of "digital sequence information on genetic resources" (decisions XIII/16 and NP-2/4, respectively). The decisions included formation of an Ad Hoc Technical Expert Group (AHTEG) on Digital Sequence Information on Genetic Resources, and an invitation to governments, indigenous peoples and local communities, and relevant organizations and stakeholders to submit 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 Nagoya Protocol. The Executive Secretary will prepare a synthesis of the submitted views and information that will be considered by the AHTEG.

In addition, the Conference of the Parties requested the Executive Secretary of the CBD to commission a fact-finding and scoping study, the subject of this report, to clarify terminology and concepts and to assess the extent and the terms and conditions of the use of digital sequence information on genetic resources in the context of the CBD and the Nagoya Protocol (paragraph 3(b)). This study references and in some cases complements work undertaken as part of other international policy processes. These include the implications of digital sequence information for benefit sharing under consideration by the UN General Assembly process on biodiversity in areas beyond national jurisdiction, where the issue of access and benefit sharing (ABS) for digital information from marine genetic resources has been raised; the World Health Organization (WHO) as part of its Pandemic Influenza Preparedness (PIP) Framework; and the International Treaty for Plant Genetic Resources for Food and Agriculture (ITPGRFA) and the Commission for Genetic Resources for Food and Agriculture (CGRFA), which are both considering the implications of digital sequence information for genetic resources for food and agriculture.

The research for this study was undertaken over the course of three months, and included a review of primary and secondary literature by the project team, as well as interviews and meetings with a wide range of stakeholders and experts, including academic researchers, industry representatives, database managers, civil society groups, policy makers, and others. Discussions were held with the project team for the ITPGRFA scoping study on "how current synthetic biology technologies and practices related to the exchange and use of sequence data are relevant for the Treaty" and with the CGRFA, which launched a study on digital sequence information in October 2017. As a result of these parallel research processes, and production of scoping studies, the emphasis in this study is the use of digital sequence information for the conservation and sustainable use of biodiversity, and academic and commercial research oriented towards pharmaceutical, industrial biotechnology, and other applications, outside of food and agriculture. However, there are clearly significant overlaps in the issues addressed, including the widespread use of public databases, and the use of agricultural plant genetic sequence data in sectors other than agriculture.

In total, the research team conducted semi-structured interviews with 55 individuals from 17 countries. Despite the short time-frame for the study we aimed to capture as broad and diverse a range of views as possible. This report focuses more narrowly on the terms of reference for the scoping study, as outlined in decision XIII/16, producing a resource for the AHTEG and others, and does not explore the broader policy implications of digital sequence information, or make recommendations.

2. Terminology

This section provides an overview of the range of terminologies employed in discussions associated with digital sequence information, current practices within the research and database community, and terms employed in policy processes. It responds directly to a request by the Parties in decision XIII/16 for further clarifications on terminology. We do not explore issues of scope associated with terminology, nor the evolution of the term "digital sequence information" within the CBD and Nagoya Protocol policy processes, since these are not part of the terms of reference for this study, and will be examined by the AHTEG.

Although the term "digital sequence information" is used in decisions CBD XIII/16 and Nagoya Protocol (NP) 2/14, a number of related terms are used within the scientific community, by governments, and as part of other international policy processes. These include *resources in silico*, *genetic sequence data*, *genetic sequence information, digital sequence data, genetic information, dematerialized genetic resources, in silico utilization*, *information on nucleic acid sequences*, *nucleic acid information*, and *natural information.* A related term and concept with implications for this discussion, also raised in many recent submissions in response to decision XIII/16, is that of *intangible* genetic resources, which include digital sequence information, in contrast to *tangible* physical genetic resources as defined within the Convention.

2.1 Exploring Terminology within Scientific and Policy Circles

Genetic sequence data appears to be the term most widely used within scientific research circles, but the large databases joined into the International Nucleotide Sequence Database Collection consortium (discussed below) employ slightly different variations of terms. The DNA Data Bank of Japan uses the term "*nucleotide sequence data*"; the European Molecular Biology Laboratory (EMBL) European Bioinformatics Institute (EMBL-EBI) uses "*nucleotide sequence information*" and GenBank in the US uses "*genetic sequences*". In part, differences in terminology reflect differences in what is referred to, for example, if a database includes DNA, RNA, or amino acid sequences. Within ABS policy discussions, differences in preferred terminology usually grow from divergent views of what falls within the scope of the Nagoya Protocol and national laws. The term "digital sequence information" is not employed within scientific or database circles, however, and has grown from the CBD policy process.

Processes within the CBD, the ITPGRFA, the UN General Assembly, and the WHO have explored terminology associated with genetic sequence use, the transmission of this data or information digitally, and the implications of employing different terms, including the words "digital", "sequence" and "information".

Within the UN General Assembly's policy process on marine biodiversity in areas beyond national jurisdiction, the first term used in discussions was *resources in silico*, but in order to follow more closely the language from the CBD decision, *digital sequence data* became the term of choice. The ITPGRFA has elected to use the term "*sequence data*" in its recently commissioned scoping study (Welch et al, 2017). In a background study paper for the FAO and ITPGRFA, Manzella (2016) uses the term *genetic information* (processed sequenced data) under which is subsumed *genomic data* (raw sequence data); he notes that in biological research, "data" is a building block that, once organized and processed (eg through context and structure) is turned into "information" (Manzella, 2016; see Table 1. below on Categories of Information developed by Jaspars, 2017).

The WHO PIP Framework[1] uses the term *genetic sequence data (GSD),* which they define as: "The order of nucleotides found in a molecule of DNA or RNA…contain[ing] the genetic information that determines the biological characteristics of an organism or a virus". This term is also used by the Global Initiative on Sharing All Influenza Data (GISAID), that acts as the main collection of genetic sequence data of influenza viruses and related clinical and epidemiological data for the global community. Its EpiFlu Database Access Agreement (discussed further below), defines "Data" as "…any and all (i) sequence data and other associated data and information contained in the GISAID EpiFlu Database pertaining to influenza viruses, (ii) any annotations, corrections, updates, modifications, improvements, derivatives or other enhancements to any such data contained in the GISAID EpiFlu Database, and (iii) any safety information relevant to use of the data or to regulatory approval of vaccines or other therapies that embody or utilize the data contained in the GISAID EpiFlu Database." (www.gsaid.org)

All policy processes that have addressed digital sequence information have included significant discussions around terminology, including ambiguities on the terms used. The respective international policy processes have taken steps to harmonize terminology but, as one researcher noted, "harmonizing terminology is something that is difficult if not impossible to achieve for dynamic terminologies that are used in multiple disciplines, and in fields that are actively evolving and changing over time, but in unpredictable ways".

Table 1. Categories of information and types of data incorporating different levels of processing and analysis

Categories of information	Explanation	Types of data
Data only	**Raw data (e.g. genetic sequence data)**	• **Metadata associated with the samples** • Initial taxonomic analysis of the samples • **Genetic sequence data (DNA)** • Transcriptome data (RNA of the genes that are functional at that time) • Automatic gene/transcriptome function annotations • Protein sequence data (DNA/RNA data automatically translated to give amino acid sequence)

Data and analysis	Genetic sequence data which has been annotated with putative gene functions using an algorithm	• Initial taxonomic analysis of the samples (DNA methods?) • Automatic gene/transcriptome automatic function annotations • Protein sequence data (DNA/RNA data automatically translated to give amino acid sequence) • *Protein structure data (Embargo)* • *Metabolite data (mainly commercial databases)*
Data, analysis and interpretation	Critical evaluation of the data and its analysis conducted by an expert	• Full taxonomic analysis of the samples • Manual gene/transcriptome function annotations • *Protein structure data (Embargo)* • *Metabolite data (mainly commercial databases)*

Source: Marcel Jaspars, 2017

Following is a brief review of the elements of the term digital sequence information – "digital", "sequence" and "information" – and views expressed during interviews undertaken as part of this project, in the literature, and submissions to the CBD Secretariat. We do not synthesize these various views, but instead present them as background for the AHTEG to consider in their deliberations.

"DIGITAL"

Researchers were not largely supportive of inclusion of the term "digital", with one claiming it was "confusing and unnecessary since all gene sequences are digitized anyway." Another researcher, who also manages a database, said that he has not heard the term "digital" in relationship to sequences used outside of CBD circles: "It sounds like it is describing the way information is stored – as in digitally - but how does that clarify what we are discussing?"

Others noted that the term "digital" describes the form of transmission, rather than the sequence information itself. In theory, sequence data accessed through print books and articles, and other non-digital means, would not be covered by "digital sequence information". As the Peruvian Society of Environmental Law (2017) notes: "In addition to the digital and print media employed to transmit natural information are film recordings, sound-analog recordings and, more fundamentally, gas liquid and light for the sensory perceptions of smell, sound, taste, touch and sight." They cite as examples photos of burrs from the *Arctium lappa* plant and the rudimentary sketches submitted in the 1958 patent application of Velcro, and sound recordings of "bats" and/or "dolphins" that have been cited in

347 patent applications on "echo-location". Another researcher noted that data and information are now stored on synthetic DNA, using technology developed by the company Twist Bioscience.

Another researcher echoed this point: "The crux is what sequences are we talking about, and does it need to be digital or not? It could be on a piece of paper carried into another country and would have the same implication. So 'digital' is not crucial as part of the terminology – maybe 99% of the transfers are currently in digital format, so for practical reasons it works to use digital, but the focus of what we are talking about is the sequence". Hammond (2017) also notes that since future information, or computer, systems, may not be "digital", and since sequence information that is not stored digitally should also be included in the CBD discussions, it might be worthwhile to remove "digital" from the definition.

"SEQUENCE"

The CBD definitions are often re-evaluated by various groups in light of scientific and technological changes. *Genetic resources* ("genetic material of actual or potential value") and *genetic material* ("any material of plant, animal, microbial or other origin containing functional units of heredity") have received particular attention within the context of digital sequence information. At the time of CBD negotiations, researchers focused on full sequences that coded for proteins, accessed via journal articles, conference proceedings, books, fax and the internet to some extent. Today genetic *parts* are of most interest to researchers and it is unclear whether a partial coding sequence or a DNA sequence that regulates gene expression constitutes a functional unit of heredity, and so qualifies as a "genetic resource", or how proteomes or metabolomes would be addressed. Earlier discussions focused on DNA sequences, but today sequence information is generally considered to extend beyond DNA. Sequences result from the process of determining the order of nucleotides or amino acids in a genome, transcriptome, or proteome of an organism and might include whole genome sequences, RNA sequences, short RNA sequences, exome sequences, degradome sequences or amino acid sequences. Digital sequence information might include metagenomics/metabarcoding, various epigenomic markers, and other molecular information.

Digital sequence information may have different qualities, including: DNA barcodes (short stretches of DNA that are used as a fingerprint to identify an organism); gene sequences (that include the start and stop instructions and all the necessary DNA codons to create a protein); regulatory DNA (stretches of DNA that do not code for proteins but have effects on, for example, the processing of genes); and whole genomes (the complete sequences of an organism) (BIA, 2017).

One researcher suggested a more accurate term might be *biomolecular data*, "which would include not only DNA and RNA but also the results of proteomics and metabolomics." Others claim that "sequence" narrows the scope too much, and would not, for example, cover expressions of natural information other than nucleic acids and amino acids (Vogel et al, forthcoming).

"INFORMATION"

The word "information" has generated perhaps the greatest discussion, with significant differences in opinion on whether the subject of discussions is *information* or *data*, and whether genetic resources, defined within the CBD as *genetic material containing functional units of heredity*, would include digital sequence information.

Dutfield (2012) distinguishes between ways that "information" is used in discussing DNA: information *about* DNA is used in relation to "growth, development, regeneration, reproduction, disease, resistance to disease, and general cell functioning, of which vast amounts are being generated…" but which cannot be acquired by looking only at the sequence of bases. Alternatively, some researchers describe *DNA information* and mean the arrangement of the letters ACGT (an acronym for the four types of bases found in a DNA molecule: adenine (A), cytosine (C), guanine (G), and thymine (T)) in a sequence, or 'raw data'. Dutfield argues that the former is more accurate - information science and digital technology are applied to DNA sequence *data*, to generate *information* that is intelligible, usable and sharable.

Several research groups and companies have recommended that the CBD policy process maintain a clear distinction between genetic material itself, and the data describing the order of DNA or RNA nucleotides in genetic material, or information analytically inferred from that material. They also propose distinguishing between *tangible* (physical) and *intangible* (including digital sequence information) materials and information. As one researcher noted: "If you talk about digital sequence *information*, rather than sequence *data*, you put a lot of emphasis on the fact that it is information, and is not tangible, is not physical material".

For others, the emphasis on physical material rather than the informational dimensions of genetic resources creates risks for benefit sharing (Ruiz Muller, 2015). They recommend modifying 'information' with either 'natural' or 'artificial'. Because the provenance of a sequence is not clear from the term "digital sequence information on genetic resources", and since sequences can also be synthesized and artificial, it is argued that the term runs the danger of extending the scope of ABS to artificial sequences, while not addressing the full range of natural information that should be included (Vogel et al, forthcoming).

There is clearly a great deal more discussion required on the terminology associated with this issue. An over-arching goal is to find the balance between terminology that is on the one hand adaptive, dynamic and fluid enough to reflect the pace of scientific, technological, market and other change, and on the other hand is clear and solid enough to provide legal certainty, and resolution around the scope of ABS (eg Schei and Tvedt, 2010; Tvedt et al, 2016; Vogel et al, forthcoming; Laird and Wynberg, 2016; Ruiz Muller, 2015). Although the term "digital sequence information" is a place-holder and will receive further consideration from the AHTEG, and although it raises numerous questions and concerns as noted, we will use the term throughout this document in line with decision XIII/16.

3. The Use of Digital Sequence Information

Digital sequence information permeates nearly every branch of the life sciences and modern biology today, allowing for computational analyses and simulations that are significantly cheaper and quicker than biological experiments run in a wet laboratory[2]. It contributes to understanding the molecular basis of phenotype, evolution, and how we can manipulate genes to provide new therapies and cures for disease, industrial products, renewable energy sources, chemicals, and other products and solutions (Field *et al*. 2008; GGBN, 2017). Digital sequence information may be natural or synthetic, identical to sequences found in nature, or designed, mutated, or degenerated (Patron, Earlham Institute, in Scott and Berry, 2017).

In this section, we briefly review how digital sequence information is produced, and how it is used by researchers in an increasingly networked, global, inter-disciplinary, and collaborative research environment, including its use in synthetic biology research and applications within some industries. The important role of digital sequence information in deepening our knowledge about biodiversity, identifying and mitigating risks to threatened species, enhancing our ability to track illegal trade, identifying species and the geographic origins of products, biodiversity planning, and other conservation research and management uses, will be discussed in Section 7.

Digital sequence information is the product of sequencing technologies that have become faster, cheaper, and more accurate in recent years. The aim of DNA sequencing is to determine the order in which each of the four DNA bases are arranged in the molecule. There are two major sequencing techniques, the first being early or *first generation sequencing*. These methods were based on the use of labor intensive chain termination DNA amplification and electrophoresis methods to visualize the resulting sequence. This method was limiting in that only high quality, single source DNA could be sequenced and some prior knowledge of the target DNA sequence was needed. With advances in molecular biology the methods used to sequence DNA evolved to *next generation sequencing* (*NGS*), which is also called *deep sequencing* or *high throughput sequencing,* and makes it possible to re-sequence entire genomes or sample entire transcriptomes more efficiently, cheaply, and in greater depth (Martyniuk et al, 2017). New *third generation sequencing* platforms are currently under development, consisting of single molecule sequencers and do not require DNA amplification (Heather and Chain, 2016). [3]

All *NGS* platforms produce massive amounts of sequencing data because millions of DNA fragments can be sequenced in parallel and simultaneously. As a result, bioinformatics goes hand-in-hand with NGS. Computational algorithms are used to develop tools and software to analyze tremendous amounts of biological data (EBI, 2017; National Academy of Sciences, 2017). Advances in these information technologies, including massive storage capacity, powerful data manipulation techniques, and graphical capabilities have transformed molecular biology and lead to new fields such as metagenomics (Reichman and Okedji, 2012).

Metagenomics, also known as environmental genomics, or environmental DNA (eDNA) sequencing, allows researchers to sequence and analyze the genomes of all the microorganisms present in a sample of soil or water, which may contain thousands of different species.[4] Metagenomic analysis produces data from millions of small fragments of the genome of each organism in the sample, in contrast to whole genome sequencing (WGS) data, which describes the entire genome of one specific organism (SFAM, 2017). Metagenomics has vastly increased our knowledge of genetic and biological diversity (Escalente et al, 2014).

Another type of digital sequence information with particular relevance to biodiversity is DNA barcodes. DNA barcoding focuses on genes that are present in most organisms, however the sequence of the gene is unique to each species, like a genetic fingerprint, and so allows for species identification, although this may not possible for all species (Herbert et al, 2003; Woese et al, 1985; Clarridge, 2004; Schindel et al, 2015).

3.1 How is digital sequence information used and by whom?

The digital transmission of sequence information is taking place in an increasingly globalized research context, where collaborative, global, and inter-disciplinary approaches are now the norm. Advances in science and information technologies have changed the way researchers work, making possible dynamic knowledge hubs, and diffuse scientific networks and collaborations (Reichman and Okedji, 2012). Networks of researchers from diverse institutional homes (e.g. industry, government, academia, community laboratories) commonly span the globe in a system of "open innovation" in which users add incremental value through data and knowledge along a chain that involves "swift compilation, comparison and reanalysis of genetic information from a variety of sources, across multiple databases and gene sequences" (IFPMA, 2017; ICC, 2017). These differentiated research structures and collaborations across disciplines – including biologists, molecular life scientists, mathematicians, and computer scientists – are "highly decentralized and based increasingly on a service model in which sequencing, synthesis, storage, assembly, screening and other activities are conducted by numerous different actors" (Welch et al, 2017; National Academy of Sciences, 2017).

Distinctions between academic, governmental, or industry research using genetic sequences have become blurred, as have distinctions between different industrial sectors. For the purposes of illustration, however, below we will provide a snapshot of how digital sequence information is used in synthetic biology research, and the three primary areas of biotechnology – industrial, healthcare, and agriculture. Following this, we review the rise of community laboratories and the DIY (Do It Yourself) bio use of digital sequence information.

3.1.1 Synthetic biology research

Synthetic biology was defined by the AHTEG on Synthetic Biology, as "a further development and new dimension of modern biotechnology that combines science, technology, and engineering to facilitate and accelerate the understanding, design, redesign, manufacture and/or modification of genetic materials, living organisms and biological systems" (UNEP/CBD/SBSTTA/20/8, March 2016)[5]. Synthetic biology was founded in multiple sectors, and makes use of various techniques that include DNA-based circuits; synthetic metabolic pathway engineering based on naturally occurring DNA sequences that are computer optimized; synthetic genomics; protocell construction; and xenobiology or chemical synthetic biology (Scott et al, 2015).

The field is guided by digital sequence information in order to apply gene editing techniques like CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)/Cas9, and increasingly gene synthesis, to create new organisms and systems. Nature is often viewed as an 'inspiration' or jumping-off point from which metabolic pathways are modified, genomes edited, and sequences combined from many sources (Scott and Berry, 2017).

Synthetic biology techniques include taking genes from a number of different organisms and combining them into a "vector" – an artificial DNA construct – which allows the genes to work together. Genes are selected from the genomes of micro- or other organisms collected from soil, water, or other natural environments, *ex-situ* collections, or the millions of genetic sequences in public databases. The vector containing the genes is incorporated into the "host" organism, a modified, easy to grow microorganism that can express the genes. Both the vector and the hosts are also often owned by companies that have associated intellectual property. The vectors within the hosts produce the proteins or small molecules of

interest (Jaspars, PHARMASEA, 2017). This process turns microorganism hosts into biological or microbial 'factories' fed by biomass feedstocks that produce sugars.

An example of the highly networked and global nature of how access to sequence information has changed the way research is conducted with the advent of synthetic biology are the establishment of "Biofoundries". Examples of biofoundries include Imperial College's SynbiCITE[6], and the National University of Singapore's Synthetic Biology Foundry (Eisenstein, 2016). These research facilities use robotic assembly lines to create, test and optimize microbes at a much larger scale than could be done by hand. This work is based on standardized parts – small sequences of DNA – which might be identical to sequences found in nature, either cloned from an organism or synthesized from information held in a public database or private collections, or parts that are designed, mutated, or degenerated (Patron, Earlham Institute in Scott and Berry, 2017). [7]

The use of such advance technologies accelerates the commercialization of organisms and products of synthetic biology by moving promising foundational research into industrial and clinical applications. It facilitates the development of organisms that contain synthetic pathways and networks using several genes from many different organisms, as well as mutating and editing genomes to ultimately end up with complex engineered organisms. This process is underpinned by the automated assembly of "complex and bespoke DNA molecules" (Patron in Scott and Berry, 2017).

Products and processes that utilize synthetic biology include new ways of producing pharmaceuticals like opioids and the anti-malarial artemisinin, biofuels, detection devices, cleaning up toxic spills, as well as a means to grow organs for transplant, manipulate the microbiome, and produce cosmetics (National Academy of Sciences, 2017). Estimates of the value of synthetic biology products and processes vary, with a recent estimate of $5,245.7 million in global revenues in 2015, with annual growth of 15.5% projected through 2022 (Allied Research, 2016). US annual revenues from genetically engineered plants and microbes are estimated at more than $300 billion (National Academy of Science, 2017).

A well-known example of an unusually valuable application of synthetic biology is *Aequorea victoria,* a bioluminescent jellyfish found off the cost of the US. It is cited in numerous patent searches and is the source of one of the top ten most used parts in the iGEM Registry of Biological Parts (discussed below). These include its use as a component of a microbial insecticide and the engineering of silkworms to produce yellow fluorescent cocoons used to make silk clothes and wallpapers (Slobodian et al, 2017).

3.1.2 Industrial biotechnology

Commercial industrial biotechnology uses enzymes and micro-organisms to make bio-based products in sectors such as chemicals, food and feed, detergents, pulp and paper, electronics, automotive, packaging, cosmetics, bioprocessing catalysts, textiles and bioenergy (https://www.europabio.org/industrial-biotech). Products range from high volume, low value products like biofuels, through to chemical intermediates, bio-plastics, cosmetics and fragrances, and high value pharmaceutical production and fine chemicals. This industry migrates away from traditional petroleum-based processes to engineered fermentation-based manufacturing. It is difficult to place a value on industrial biotechnology since information on its use and value rarely makes its way into the public domain because industrial biotechnology processes and products are often neither sold nor patented (so do not require disclosure). They are frequently used internally or sold between companies rather

than publicly. Many companies in this sector are privately owned and so do not report to shareholders; and governments have been slow to collect data on these activities (Laird, 2015). [8]

3.1.3 Healthcare biotechnology

Healthcare biotechnology creates an advanced class of drugs and therapies called biologics, including gene and stem cell therapies, but also vaccines and diagnostic tools such as HIV test kits (www.europabio.org).[9] The US and Europe continue to dominate in healthcare biotechnology, followed by the Asia Pacific region. China, Japan, South Korea, and Singapore are growing in importance and size. Within Europe, the UK, Switzerland, Germany, France, Sweden, Ireland, Denmark, The Netherlands, and Norway are leaders (Ernst and Young, Debra Yu, 2017; Grandview Research, 2017).[10]

Life science companies increasingly focus their strategies on digital technologies, which can impact research and development (R&D) and healthcare strategies. For example, cloud-based secure data-sharing platforms that facilitate research collaborations across geographic distances allow pharmaceutical companies to store and analyze their own data alongside publicly available genomic datasets. Drugs developed with predictive biomarkers allow trials to be smaller and potentially reach significance faster, and personalized medicines, supported by advances in genome sequencing, diagnostics, and biomarker identification reduce failure rates and time to clinical trial approval (Ernst and Young, 2017; Deloitte, 2016; Grandview Research, 2017).

The Technical Expert Working Group (TEWG) of the WHO PIP Framework has provided a summary of the ways digital sequence information has contributed to influenza-related technologies, products, inventions and patents (http://www.who.int/influenza/pip/advisory_group/gsd/en/). This includes:

- *direct use* of a particular sequence to develop a product, including production of 'synthetic' candidate vaccine viruses for vaccine development generated when a particular genetic sequence is used to design synthetic DNA (eg Novartis' synthetic H7N9 vaccine using a sequence shared by the Chinese Center for Disease Control through GISAID in 2013);

- *bulk sequences*, which, for example, might consist of multiple genes or genome sequences that share a common denominator, such as a subtype, a mutation or a conserved region, and that are analysed or used in bulk in basic research, applied research, public health and epidemiology; and

- *indirect uses* that include proteins generated by genetic sequence data to derive antibodies for therapy and diagnostics, prediction of vaccine efficacy which is dependent upon data related to viral evolution obtained through genomic analysis of sequences from circulating viruses, and understanding global migration and persistence to aid in vaccine strain selection.

The TEWG has noted that genetic sequence data has led to the development of new and better vaccines, as well as significantly decreasing the time required to manufacture pandemic vaccines (TEWG, 2014; see, too, the primary collection of GSD for influenza viruses: www.gisaid.org, 2017; Section 6.2.3).

3.1.4 Agriculture[11]

Genomic information plays a role in conservation, pre-breeding and breeding within agriculture (Manzella, 2016; see Section 6.2.2)[12]. Most commonly, plant genomic information is mined to identify

genes of interest, which can then be used to edit agricultural crop genomes. Plant genomic information might also be mined for use outside of agriculture (Welch et al, 2017). Emerging technologies are also focused on harnessing the potential for plants that have been modified to produce vaccines, high value chemicals, and pharmaceuticals, as is done with microorganisms (Welch et al, 2017; James et al, 2015). The Open Plant Synthetic Biology Research Center, a joint initiative of the University of Cambridge, John Innes Centre and the Earlham Institute, and part of the UK Synthetic Biology for Growth programme, is engineering plant systems for bioproduction, which could be far more productive than microorganisms. They are also working to share the next generation of DNA tools for 'smart' breeding of crop systems, including reprogramming crop metabolism and plant architecture to address urgent threats and challenges like climate change, soil degradation, new pathogens, restricted land use, salinity and drought (www.openplant.org).

Analysis of digital sequence information obtained from livestock also plays an important role in animal breeding and conservation of animal genetic resources. This includes: insight into the origin and domestication of farm animal species, analysis of genetic diversity amongst different breeds, marker assisted selection, QTL mapping, genome wide association studies (GWAS), genomic selection, proteomics, metabolomics, phenomics, landscape genomics, identification of genetic defects, maximizing genetic progress while maintaining genetic variability, and authenticity of products (see full discussion in Martyniuk et al, 2017).

3.1.5 Community laboratories, DIYbio, and open science

As costs of the technologies associated with obtaining digital sequence information have dropped, and become more widely accessible, an explosion of small-scale, publicly accessible community laboratories, DIY (do-it-yourself) bio, and open science collaborations that use digital sequence information have flourished within an 'open source' framework to develop products and processes to address a broad variety of issues. A range of non-profit organizations facilitate this new paradigm of innovation involving diverse participants from universities and governments to companies and high school students.

The open science approach is based on the free exchange of knowledge, materials, technologies and tools and is an effort to "democratize problem solving to enable diverse solutions through decentralized innovation", as well as the means of production (www.bios.net; Swetlitz, 2017). In part this movement is an effort to maintain the flow of research materials and methods necessary for today's "digitally integrated scientific research" at a time when the public domain is receding under pressure from expanding copyright and related laws (Reichman and Okedeji, 2012).

Examples of groups active in this arena include the following:

- Cambia, a non-profit based in Australia that creates new technologies, tools, and paradigms to promote change and enable innovation. Cambia founded the Biological Innovation for Open Society (BiOS) Initiative. Both groups are based in Australia, but are global in reach, and share a vision to "democratize, decentralize and diversify" research and "design, develop and disseminate" tools and technology, in order to share "information, knowledge, and wisdom within and between communities that have been marginalized or inadequately served" (www.bios.org; www.cambia.org).

- Open Source Drug Discovery (OSDD), a platform based in India. It similarly seeks to achieve broader social goals through a new research paradigm. Their work focuses on the provision of affordable healthcare and includes a 'virtual laboratory' to facilitate global collaboration on diseases of the developing world like malaria and tuberculosis, which are not addressed by "traditional closed-door and market driven approaches for drug discovery", and are hampered by limitations in collaboration, data sharing, and confidentiality requirements (Bhadwarj et al, 20111; 2011). Like other non-profits working in this area, their model is inspired by the open source software movement (Singh, 2008).
- The BioBricks Foundation, a non-profit founded in 2006, and leader in this movement, was established with a mission to ensure that the engineering of biology is conducted in an open and ethical manner to benefit all people of the planet. The goal is a new paradigm for research that is based on the idea that fundamental scientific knowledge should be freely available for ethical, open innovation (www.biobricks.org). They have managed numerous spin-off groups, including OpenWetWare, which promotes the sharing of information, know-how and wisdom among researchers working in biology and biological engineering (www.openwetware.org).

Contests and awards promoting Open Science approaches and the development of synthetic biology are increasing in number, with the largest being the iGEM Competition. The International Genetically Engineered Machine (iGEM) Foundation is an independent non-profit based in the US, dedicated to education and competition, the advancement of synthetic biology and the development of an open community and collaboration. iGEM runs an annual competition for college, high school and community laboratories, that encourages students to work together "to solve real-world challenges by building genetically engineered biological systems with standard, interchangeable parts." The competition draws teams from around the world, with 339 competing in 2017 (www.igem.org). [13]

4. How Digital Sequence Information is Accessed, Stored and Managed

Genomic information is accessed through journal articles; supplementary files linked to published papers; online; public, industry, or research institution collections; synthesis companies; foundries; or the millions of genetic sequences in public databases or genetic parts registries. It is also found in emails, and online.

This section reviews the most common ways that digital sequence information is accessed: from databases and registries. In the next section, we will review how "new" digital sequence information makes its way to research and databases from field collections involving physical samples of soil, water, or other natural environments, and from *ex-situ* collections.

4.1 Public Databases

Genomic technologies used to study genes and their functions generate an unprecedented amount of information, making this "an intensely data-rich field". As a result, bioinformatics – the collection, classification, storage and analysis of complex biological data - has grown alongside genomic technologies in order to store, retrieve, and analyze these vast and growing amounts of information and the large-scale datasets generated (Pevsner, 2015; www.ebi.ac.uk). In the late 1970s, when DNA sequence data began to accumulate in the scientific literature, the early databases were set up to store

and organize these sequences, and it soon became best scientific practice to publish new genetic sequences in sequence databases[14] (www.ebi.ac.uk).

There are now more than 1,500 publicly accessible biological databases (*Nucleic Acid Research*, 2014), organized based on heterogeneity, data type, scope and curation. They might include sequence data for nucleic acids (such as databases with RNA expression information), genome databases for model organisms, RNA databases for various RNA types (for microRNAs, snoRNAs, tRNAs, piRNAs, etc.), and amino acid databases with information about known proteins. Based on the level of curation, they are classified as "Primary"– containing raw data (eg Sequence Read Archive (SRA) https://www.ncbi.nlm.nih.gov/sra) – or "Secondary" – containing curated and analyzed data (e.g. Refseq https://www.ncbi.nlm.nih.gov/refseq/).

Databases are further classified as comprehensive or specialized. Examples of specialized databases include: WormBase (http://www.wormbase.org/#012-34-5); Banana Genome Hub (http://banana-genome-hub.southgreen.fr/); and SPGDB (http://pranag.physics.iisc.ernet.in/SPGDB/). Comprehensive databases contain different data types from numerous species, and include those within the International Nucleotide Sequence Database Collaboration, the largest public databases.

4.1.1 The International Nucleotide Sequence Database Collaboration

The International Nucleotide Sequence Database Collaboration (INSDC) (www.insdc.org) is comprised of three global partners:

- **The European Nucleotide Archive**, based at the EMBL European Bioinformatics Institute (EMBL-EBI) in Cambridge, UK. The EMBL-EBI is funded by 23 member states and two associate member states and contains the world's most comprehensive range of freely available molecular data resources. This includes the PRIDE Archive, a centralized public repository for proteomics data; The IPD (Immuno Polymorphism)-MHC Database for sequences of the Major Histocompatibility Complex (MHC) from a number of different species; the European Variation Archive including genetic variation data from all species; and ENSEMBL, a genome browser for vertebrate genomes (Martyniuk et al, 2017).

- **GenBank**, based at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, USA. The NCBI includes more than 30 databases related to genes, genomes, proteins and chemicals, as well as bibliographic records from MEDLINE and other sources, and is funded by the US government. The Entrez retrieval system provides integrated access to medical literature and nucleotide and protein sequence databases, including complete genomes and schematics of entire chromosomes, as well as associated mapping. In addition to GenBank, NCBI hosts the following databases, also part of the INSDC: the High Throughput Genomic Sequences Database; the GSS Database of unannotated short single-read primarily genomic sequences from GenBank; the SNP Database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations; and the Gene Database that provides detailed information for known and predicted genes defined by nucleotide sequence or map position, containing more than 17 million entries and including data from all major taxonomic groups (a record may include nomenclature, Reference Sequences, maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide) (Martyniuk et al, 2017).

- **DNA Data Bank of Japan**, based at the National Institute for Genetics in Mishima, Japan, which primarily collects sequence data from Japanese researchers, but also from other countries, and shares data with the EMBL-EBI and GenBank.

These partners "capture, preserve, share and exchange a comprehensive collection of nucleotide sequence and associated information" for use by the scientific community, and develop new services to handle "the changing landscape of data types". They exchange data, standard formats and share technology, and incorporate everything from raw data (e.g. next generation sequencing reads) through to assembly data, experimental design details, taxonomic information, functional annotation and information about the projects and biological samples associated with sequencing efforts (Cochrane et al, 2016; Toribio et al, 2016). All INSDC partners are publicly funded by their host governments, and the INSDC's policy (http://www.insdc.org/policy.html) emphasizes the mandate to free, unrestricted access to all of the data records in their database (Cochrane et al, 2016).

Since there are so many databases containing digital sequence information, some of which overlap, it can be a challenge to navigate between them, and so several meta databases have been established to collate information from other databases. Some meta databases simply merge information into a different viewing format, while others focus on curating data with a focus on a particular disease or organism (Bolser et al, 2012). For example, the field of epigenetics is growing rapidly alongside advancements in next generation sequencing, and a meta database – the Human Epigenome Atlas (http://www.genboree.org/epigenomeatlas/index.rhtml) – was established to keep up with the huge influx of epigenetic data for humans. Databases focused on epigenomic studies in plants can be found at the EPIC (Epigenomics of Plants International Consortium) website (https://www.plant-epigenome.org/).

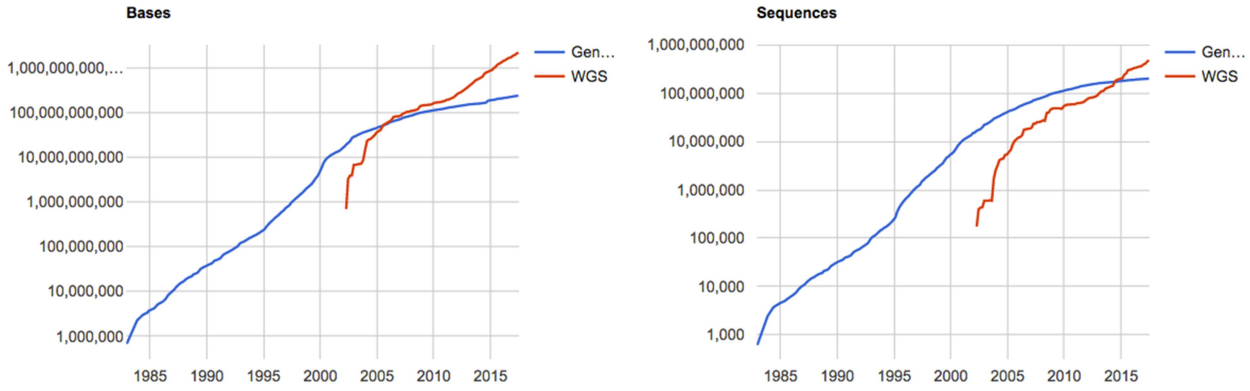
Other examples of metadata sources include Fairsharing (https://fairsharing.org/databases/), a searchable portal of three linked registries covering standards, databases, and data policies in the life sciences, broadly encompassing the biological, environmental and biomedical sciences, launched to build the social and technical infrastructure necessary to openly share data. Ark DB at the Roslin Institute (http://www/ed/ac/uk/roslin/facilities-resources/bioinformatives) is a generic, species-independent database built to capture the state of published information on genome mapping for a given species (Martyniuk et al, 20017).

A common means of using databases is to run a Basic Local Alignment Search Tool (BLAST) search which finds regions of local similarity between query sequences and sequences on the databases. To do this every record is searched, which means that all of the data contained in a database is accessed on a regular basis. BLAST compares nucleotide or protein sequences to sequence databases to calculate the statistical significance of matches. It can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families (https://blast.ncbi.nlm.nih.gov/blast.cgi). As one researcher described: "The power of these databases is in the ability to compare hundreds of sequences quickly. Any restriction in the open availability of genetic sequence information uploaded to these databases will reduce their value and utility." But BLAST might also help researchers find an identical sequence in a different organism as a way of avoiding the use of a sequence that raised legal uncertainties under ABS, or to avoid monitoring (Welch et al, 2017; Bagley, 2017).

4.1.2 Increase in data flow and use

The amount of data flowing into databases and registries is exponentially increasing. For databases, this includes the number of bases and sequences, the numbers of individuals and species sequenced, and the depth of genomic coverage obtained per sample. Databases such as EMBL, GenBank, Sequence Read Archive (SRA) and the DNA Data Bank of Japan have almost doubled in size in the last few years. They now serve as repositories of quadrillions (>10 to the 15th) of nucleotides of DNA sequences. This figure will soon be in the quintillions (>10 to the 18th). These base and sequence records have been collected from over 300,000 organisms (IFPMA, 2017; Pevsner, 2017; Cochrane et al, 2016; NHM, 2017).

From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months (www.ncbi.nlm.nih.gov/genbank/statistics) (Figure 1). The number of sequence entries have increased from 606 in 1982 to 201,663,568 in June 2017 (www.ncbi.nlm.nih.gov/genebank/statistics/). The INSDC assembled/annotated sequence dataset trebled between 2012 and 2015 (Cochrane et al, 2016).



Source: www.ncbi.nlm.nih.gov/genbank/statistics

Figure 1. GenBank Sequences: since 1982, the number of bases has doubled approximately every 18 months (blue = GenBank; red = Whole Genome Shotgun)

On the user side, the EMBL-EBI search engine runs on average 12.6 million jobs every month (www.ebi.ac.uk). Other statistics illustrating the scale of EMBL-EBI's engagement with the global research community include:

- scientists at over 5 million unique sites use EMBL-EBI websites every month;
- in 2016, EMBL-EBI had 186 grants jointly funded with researchers and institutes in 62 countries throughout the world;
- every weekday, more than 27 million requests are made to EMBL-EBI websites; and
- EMBL-EBI data centres can store over 120 Petabytes (80,000 Terabytes) of data (www.ebi.ac.uk).

Database managers and others note that many datasets are not entered into international public databases due to concerns about confidentiality, control, and benefit sharing. The extent of what is not uploaded is difficult to estimate, but the amount of data flooding into these databases remains

enormous. Projects and smaller databases established for specific research areas also show massive increases in data flow into the databases, and use, in the last five years. For example, Zhi-Liang et al (2015 in Martyniuk et al, 2017) describe how the Animal QTL Database (http://www.animalgenome.org/QTLdb) has undergone dramatic growth in new data curated, data downloads and new functions and tools. Qiita, the technical knowledge sharing and collaboration platform for the Earth Microbiome Project (discussed below), has also seen a substantial increase in data usage in the last few years, with the project submitting studies and samples to EBI, as well.

4.1.3 Standards for digital sequence information sharing and compatibility between databases

Essential to the use of digital sequence information have been efforts to standardize and unify the terminology of genetic databases. By standardizing electronic data, it "can be exported, translated, queried, and unified across independently developed systems and services" (Gruber, 2009). Adherence to agreed data standards allows INSDC partners to develop complementary data-submission tools, to exchange data on a daily basis, and to present the same content in different ways according to user needs[15] (Cochrane et al, 2016).

In 1998, the Gene Ontology[16] (GO) Consortium (http://www.geneontology.org/) was founded to unify the genetic terminology of databases for three model organisms widely used in biomedical research: FlyBase (*Drosophila*), SGD (*Saccharomyces*), and the Mouse Genome Database. Over the years, the GO project has expanded to include additional organisms, and is now an integral part of the larger, overarching ontology classification effort called Open Biological and Biomedical Ontologies (OBO) (http://obofoundry.org/). The Genomic Standards Consortium (GSC) was founded in 2015 to promote the capture of genomic data electronically, in a standard format, including information on the environmental context and locations from where organisms originate (www.gensc.org). Early on, GSC collaborators found that the inclusion of environmental context data was not common, and that even for bacterial and archaeal species with validly published names, strain names were not routinely captured in genomic annotation documents before the sequencing of large numbers of genomes from the same species. Now, they emphasize, such information is considered essential. "As the number of habitats and communities sampled using metagenomics approaches increases, we are also being forced to rethink our understanding of the minimum information required to adequately describe a genome sequence. Without adequate description of the environmental context and the experimental methods used, such data sets will be of less value for researchers wishing to conduct comparative genomic studies or link genetic potential with the diversity and abundance of organisms" (Field et al, 2008; www.gensc.org). Annex 1 further explores the important role of ontology initiatives in creating unified and standardized terminology associated with digital sequence information.

The increasing inclusion of environmental context data over the last decade makes it easier to trace sequences back to source countries, a critical step for ABS implementation. This data often includes geographical coordinates, and information about collections from which sequences might have come. As EMBL-EBI describes the value of metadata: "For example, if you're involved in sequencing samples from the environment, perhaps to understand biodiversity in different conditions, or to investigate associations between crop yield and differences in soil flora, it would be useful to know when and where your samples were collected. Standardised descriptors of collection time and geographical location can then be associated with any sequence derived from each sample… Indeed, metadata is so important that we create databases dedicated to organising it….Storing metadata in this way ensures that a specific sample is referred to consistently in several data resources" (www.ebi.ac.uk).

Although contemporary collections include metadata on the environmental context and origins, earlier records did not include this information, and not all contemporary records are complete, or follow the minimum information recommendations of the standards groups. As one database manager put it: "In the 1980s, researchers were sequencing laboratory organisms and were – for example – trying to understand a common virus. But the focus shifted once people started sequencing things from around the world, and wanted to know where things came from. Gradually, people are beginning to give this information, but we are still at the mercy of the data submitters. We don't have the ability to curate individual records, we get submissions on average every 6 minutes, so we can't have a great deal of communication with submitters. We are working on this, though, and hope to get the community as a whole to take responsibility for this."

4.2 Registries of Standard Parts

In addition to digital sequence information accessed through public databases, another common source of genetic sequence data is repositories like the Registry of Standard Biological Parts, and the Inventory of Composable Elements as part of Open Science networks. The standard parts used in iGem and elsewhere are called BioBricks, and are DNA sequences that encode for a specific biological function (iGEM.org). DNA parts are a mix of natural and synthetic; they might be identical to sequences found in nature, either cloned from an organism or synthesized from information held in a public database or private collections, or they may be designed, mutated or degenerated parts (Patron, Earlham Institute, in Scott and Berry, 2017).

Assembly Standards, like the BioBrick Standard, ensure compatibility between parts and define how part samples will be assembled together by an engineer. Part samples that belong to the same Assembly Standard can be combined to create new, longer, and more complex parts. Parts might include coding sequences, promoters, ribosomal binding sites, protein domains, protein coding sequences, plasmids, primers, and terminators. The Registry of Standard Biological Parts creates a library of standard parts that have been tested, characterized and organized (each with an identification code), making it easier for researchers to share parts and collaborate (http://parts.igem.org). The Registry is available for use by researchers around the world, who may also contribute their own parts following validation. The Registry currently holds over 20,000 documented parts.

5. Generation of "New" Digital Sequence Information from Physical Samples

Most digital sequence information is accessed through databases or parts registries, but some groups seek out physical samples through field collections, citizen science sourcing programs, and many acquire samples and digital sequence information through *ex situ* collections. These physical samples are then sequenced, and the sequence information subsequently loaded onto databases.

5.1 Field collections and citizen science

Field collections of physical samples are a much smaller part of research strategies in higher technology industries than in the early years of the CBD, and few companies undertake regular and systematic collections, although there are exceptions. Academic groups continue to have an interest in physical samples, with a recent surge of interest resulting from the wide diversity of microbial species that can now be studied in environmental samples using metagenomic sequencing technologies. Interest persists

in areas with high species diversity, extreme environments, and unique ecological niches. As one researcher described this: "There are environmentally selected strains of organisms that are so different you still need to go out and collect them. We are still discovering new genes, and do not know what they do... This is mainly driven by academics and smaller companies."

Another researcher made the point that although the science is moving away from physical material, it is still necessary in most cases: "We still need to work with the physical material… yes, more and more we will be able to use digital sequences alone, but it is still very difficult if you don't have a living organism to deal with. I am hard pressed to come up with examples of true applications that were just pulled from a sequence unless you are talking about very modest metabolic engineering to produce stuff in *E. coli* or another production organism. Much past this is very tough still."

The J Craig Venter Institute (JCVI) Ocean Sampling collections have been one of the most extensive field collecting programs in recent years. The Global Ocean Sampling Expeditions involved circumnavigating Earth and collecting samples from dozens of countries in temperate and tropical regions, and extreme environments like Antarctica and deep sea vents, areas beyond national jurisdiction. JCVI sequenced and analyzed microbial life found in the marine water samples, and placed all resulting sequence data in the public databases (http://www.jcvi.org/cms/research/projects/gos/overview/; Slobodian *et al*. 2017).

A number of citizen science programs solicit samples from around the world as part of efforts to understand biological and genetic diversity, particularly of microorganisms. Citizen scientists and researchers share samples in exchange for data analysis, with no costly and time-consuming collecting expeditions required. As a result, the scope of these efforts can be enormous, generating quantities of data and covering geographic distances not otherwise possible.

Examples of citizen science projects include the Citizen Science Soil Collection Program at the University of Oklahoma. Started in 2010, this program collects soil samples from citizen scientists from around the United States as part of an effort to identify new drug-like molecules from fungi (http://whatsinyourbackyard.org). A similar project is Drugs from Dirt, run by Sean Brady of Rockefeller University, which receives soil samples ("the poor man's rainforest") from individuals across the US. The objective is to survey the metagenome for genes of potential value to drug development, which contain conserved motifs that can be amplified by PCR and NGS. From collections they made in the parks of New York City and elsewhere, they found an abundance of unfamiliar genes, with less than one percent of molecule-encoding sequences matching up to known genes. "Throughout the history of the field, there has been this idea that one travels to remote parts of the world to collect strange bacteria. But those environments are fragile and disappearing," Brady says. "Meanwhile, we're finding that by using modern sequencing approaches, it's possible to turn up all of the same potentially useful molecules in our own backyards." (Science Daily, 2016).

Other academic research projects undertaken by citizen scientists and researchers are global in reach. Examples include the Earth Microbiome Project founded in 2010 as a "systematic attempt to characterize global microbial taxonomic and functional diversity for the benefit of the planet and humankind" using DNA sequencing and mass spectrometry on crowd sourced samples (www.earthmicrobiome.org; Gilbert et al, 2014). The Earth Microbiome Project focuses on bacterial, archael, and eukaryotic microbial diversity. Samples have come from 7 continents, 43 countries, 21 biomes, 92 environmental features, and 17 environments; the Project "has now dwarfed by a hundred-fold the scale of both sampling and sequencing of meta-analysis efforts" (Thompson et al, 2017). The genetic sequence data they collect is loaded onto databases, and is available for public use. Likewise, the

Ocean Sampling Day is a citizen science project that collects samples from around the world to provide insights, describe microbial diversity and function, and contribute to 'ocean-derived biotechnology' (https://www.microb3.eu/osd.html).

5.2 Biological-to-Digital: Portable Sequencers

A significant technological advance is the MinION, the world's "first and only nanopore DNA sequencer". The MinION is a "portable, real time, long-read, low-cost device that has been designed to bring easy biological analyses to anyone, whether in scientific research, education or a range of real-world applications such as disease/pathogen surveillance, environmental monitoring, food-chain surveillance, self-quantification or even microgravity biology." The company, Oxford Nanopore Technologies, describes its goal as "to enable the analysis of any living thing, by any person, in any environment"(https://nanoporetech.com/). The day has arrived when individuals can easily and affordably sequence genes from physical material anywhere in the world, and send it via the internet to researchers, databases, foundries, and other institutions in regions far from the site of collection.

5.3 Digital-to-Biological Converters

Advances in greater automation are making it simpler and cheaper than ever before to make synthetic DNA parts. The cost of synthesizing DNA fell by 85% between 2009-2014, and synthesizers can now churn out strings of several thousand base pairs rather than a few hundred. Today, academic researchers and companies outsource synthesis to specialized synthesis companies like Gingko Bioworks, Gen9 and SGI-DNA. As a participant in a 2016 workshop described Gingko Biowork's approach: "… when they want to look at a metabolic pathway, [they].. take 100 genes or so, synthesize all of them, and then modify them. They make use of computer evolutionary techniques to create optimized pathways, which is where the value is going to lie – using existing biodiversity as an inspiration… but they have broken the direct link between what they are creating and what they started from" (in Scott and Berry, 2017). In November 2016, an adenovirus with a genome of 34,000 nucleotides was synthesized, and the Synthetic Yeast Genome Project, an international collaboration, is synthesizing the 16 chromosomes of *Saccharomyces cerevisiae*, a total of 12 million base pairs (Hammond, 2017). The Synthetic Yeast Genome Project will culminate in the first eukaryotic cell with a fully synthetic genome (Sliva et al, 2015).

A digital-to-biological-converter has been developed to produce functional biologics in an automated fashion from digitally transmitted DNA sequences, in particular DNA templates, RNA molecules, proteins and viral particles (Boles et al, 2017). This is not widely practiced today, and Boles et al (2017) note that "manufacturing processes for biological molecules in the research laboratory have failed to keep pace with the rapid advances in automation and parallelization", but the trend is towards affordable and widespread synthesis.

The technology is moving so quickly that it will soon be possible for a researcher to inexpensively synthesize DNA on their lab, whether 10,000 or a million base pairs (Eisenstein, 2016). Indeed, SGI-DNA, a synthetic genomics company, has introduced the world's first DNA printer, a machine that will allow any company or academic laboratory to create genes, genetic elements, and molecular tools starting with digital sequence information (Welch et al, 2017).

This trend has significant implications for the identification and monitoring of samples, but even with advances in field sequencer and DNA synthesis technology, researchers are still interested in physical

samples, and the origins of the material they use. The original material can provide information a sequence alone cannot, including the relationship of genotype to phenotype, and interactions between organisms and their environment. As one molecular biologist put it: "Once you have the genomic sequence you do not necessarily need the living organisms to do something, because you can mine the genome for something of value. We are getting there – a number of companies are doing this research. But discovering things that are completely unknown from a genome alone is off in the future a bit, and it is important to remember that when one looks at the genome, there are extra chromosomal elements you have to worry about, epigenetic steps that go on in the cytoplasm that turn off or on genes, post-translation and post-transcriptional process that you may not fully understand, whether or not molecules are active or inactive…we are not there yet. How many years off is this? It is incredibly hard to predict, because the field is moving so quickly." Another researcher explained: "What most of the academic researchers don't realize is that there are/were compound collections within industry derived from decades of screening, and often the organisms from which various leads were derived. Being able to reverse engineer a pathway from a compound is the missing knowledge at this point in time."

5.4 Ex situ Collections

There are a wide and varied range of *ex situ* collections held by public entities, non-profits, scientific research institutions like botanical gardens and natural history museums, culture collections, universities, companies and others. Below we look at three *ex situ* collections and the increasing move to digitize their collections. These efforts are primarily focused on producing digital images and sharing data about specimens like location and date of collection, however some digital sequence information is also shared with the public via databases: the Royal Botanic Gardens, Kew; the Natural History Museum, London; and the World Federation for Culture Collections. [17]

The Royal Botanic Gardens, Kew has around 7 million botanical specimens in the herbarium and 1.25 million fungi; 50,000 living specimens in the gardens and 35,000 in the seed bank; with laboratory based collections including a DNA tissue bank and genetic sequence collections. Around 26,000 accessions, linked to appropriate permits, come into Kew every year, roughly 25% of these collected by Kew staff and project partners, and the remainder sent to Kew from other botanical institutions. *Ex situ* collections typically share physical materials to provide a level of redundancy and safety to collections and materials are also provided for research purposes. Most of Kew's collections have digital analogues and they are working to make these freely available via their website. When collecting, Kew seeks to acquire permission on whether material may be digitized, and whether results might be disseminated in publications or databases; if the donor is unsure, Kew includes on the permission forms what they will do with the material, so they have a record. Specimens can move between collections, duplicates may be made and sent to other herbaria, DNA extracted and the genetic information passed to international databases (eg GenBank), seeds are taken and cryopreserved elsewhere, and so on. Archival collections are increasingly included in the DNA database. Kew notes "a rapid change over the past 18 months or so, as devices such as minIONs are becoming much more accessible and practical, and there is increased demand for DNA samples from Kew's collections" (Paton, A. in Scott and Berry, 2017).

The Natural History Museum, London, has large collections of around 80 million objects including animals, plants and microorganisms from all regions of the world including areas beyond national jurisdiction, and is still acquiring material from these places. Within the Museum, researchers use DNA sequencing, genomics, and biochemistry techniques, but much of the collections may never have its DNA examined. DNA of specimens 100 years old is now routinely being examined, although to a much

lower extent and with less success than modern specimens. Given the number of specimens in the collection, their disparate origins, and range of possible conditions attached to them, data management to implement ABS conditions is challenging, but this is an important first step for identifying provenance for any sequence information. The Museum staff have identified ABS decision points in their workflow and, using the Consortium of European Taxonomic Facilities Code of Conduct and Best Practices (http://tinyurl.com/hmon7ff) have developed policies and procedures to manage ABS compliance. They are working with the Global Genome Biodiversity Network (GGBN) to develop the use of data standards for permit information (https://terms.tdwg.org/wiki/GGBN_Permit_Vocabulary) which will allow for the transfer of information on permits alongside sequences placed in public databases (C. Lyal in Scott and Berry, 2017).

Microbial research has undergone significant changes over the last few decades (see Reichmann et al, 2016 and Table 2) and is central to the use of digital sequence information today.

Table 2. The Changing Characteristics of Contemporary Microbial Research

Pre-1990s	**Recent past and future trends**
Phenotype-based inquiry	Genotype-based inquiry
Primary focus on single organisms and subsystems	Increasing focus on interdependence and complex systems
Mostly single discipline	More inter-disciplinary
Atomistic/insular/local	Integrative/collaborative/global
In vitro resources	*In silico* resources
Print communication	Networked digital communication
Data limited	"Big data", especially genomics
"Small" science organizations	"Big science" organizations
Public/basic research largely separated from the private applied research	Distinction between basic research and applications frequently collapsed

Source: Reichmann et al, 2016

Ex situ collections for microorganisms are spread across the globe, with the World Federation for Culture Collections (WFCC), based at the Institute of Microbiology, Chinese Academy of Sciences (IMCAS), representing 728 microbial resource centers in over 75 countries. The WFCC is concerned with the collection, authentication, maintenance and distribution of cultures of microorganisms and cultured cells, and represents a vast array of academic, public, and industry collections (www.wfcc.info; Reichmann et al, 2016). Most WFCC collections belong to academic or government public entities, with 8% semi-governmental, 4% private non-profit, and 1% industry (Dedeurwaerdere et al, 2012).

The World Federation for Culture Collections manages the Global Catalogue of Microorganisms and is continuing to build this system. To date, the Catalogue contains information on 48,335 bacterial, fungal and archaea species from 112 collections in 43 countries and regions. The growth of biotechnology has increased demands "for authenticated, reliable biological material and associated information…" (http://gcm.wfcc.info). [18]

More than 200,000 new samples of microorganisms are deposited each year in the WFCC collections, but this represents only a small fraction of newly discovered organisms referred to in published research (less than 1% were deposited in 2008) (Dedeurwaerdere et al, 2012; 2016). Deposits in culture collections are primarily from *in situ* environments in the country of the culture collection, however a substantial portion of deposits continues to come from the countries where collections are found. The distribution of strains from collections is primarily to academic and public institutions, but 23% goes to the private sector (Dedeurwaerdere et al, 2012).

Active very early in the ABS policy process, WFCC introduced the MOSAIC code of conduct in 1993, which included a standard MTA. In 2005, the WFCC developed a Global Unique Identifier (GUID) for microbes, which is a permanent persistent label, linked to the internet, that allows for up to date tracking of microorganisms. More recently, in 2015, the WFCC began to develop the TRUST system, which combines tracking with a search engine and data on the outcomes of research. TRUST works with the Global Catalogue of Microorganisms, merging administrative and legal data with scientific and technical data. At present, 108 culture collections, from around 43 countries, have signed on to the Global Catalogue, and are merged into a single portal (Desmeth, P in Scott and Berry, 2017). Other culture collections and microbial projects have also addressed ABS, in ways relevant to digital sequence information, including MTAs and codes of conduct.[19]

6. Tools to Manage Digital Sequence Information: Conditions of Use Notices and User Agreements

A range of approaches attach conditions to the use of digital sequence information, some specifically addressing benefit sharing, and others serving as possible templates for consideration. Below, we review a few of these approaches including notifications on databases and websites, conditions of use notices, click through agreements, open source MTAs, and user agreements. In most cases, negotiation of an agreement between a commercial user and a contributor of sequence information is envisioned at some point in the future, once commercial interest or use has been established.

6.1 Conditions of use notices

A number of websites and databases include Conditions of Use Notices in an attempt to identify the obligations of those using material or information, and in order to protect the rights of the country of origin. This includes, for example, the JCVI online database and computational resource, CAMERA (A Community Resource for Metagenomics), which was later absorbed into the iMicrobe data resource. In their user notice, JCVI sought to protect the country of origin's interests by asserting that the downloaded digital sequence information was the patrimony of the country of collection, and stating that users of this information agreed to acknowledge the country of origin in any publication and contact the CBD focal point of that country if they intended to use the genetic information for commercial purposes (www.jcvi.org). It is not clear to what extent these notices are legally binding, however (Slobodian et al, 2017).

Other groups more generally place notices about the Nagoya Protocol, and information on their practices, on their websites. The Drugs from Dirt Project (www.drugsfromdirt.org), for example, links to the Nagoya Protocol. In a section called "What Will Be Done with Your Samples & Biodiversity Best Practices" they describe how they will sequence and then destroy physical samples (no organisms will be cultured); the sequence data will be added to their database; DNA from samples will not be cloned, and no samples will be used directly for drug discovery.

One step beyond a condition of use notice, is a click-wrap, or click-through, agreement that requires users to click their assent to certain terms in order to gain access to the website or database. These are commonly used by software companies, and if the user does not click "ok" or "agree", they are not granted access to the database/website/software. These steps might be taken by a user to assent to ABS provisions, and are being explored by a number of researchers for use with databases. Skeptics argue, however, that the value of click through agreements is limited, as one said: "Who actually reads these and fully understands their obligations?".

6.2 Open source and user agreements

Open source agreements

Open source agreements grew from a desire within the scientific community to facilitate the exchange of methods and materials that underpin basic research without the costs associated with traditional material transfer agreements (MTAs) or other forms of licensing agreements. MTAs are seen as overly burdensome, costly, time-consuming, and restrictive, resulting in delays for research (www.biobricks.org). MTAs might be manageable for larger research institutions and companies, but are considered out of reach for smaller research institutions and individuals. Based on experiences in the open software movement, open MTAs are intended to support both "freedom to operate and freedom to cooperate". In open source software agreements, the source code for computer programs is available under the terms of the license to others who agree to these terms, so that the program can rapidly evolve with many users involved in debugging and modifying it to develop other products and improvements (www.bios.net).

Developments in scientific research methods mean researchers are more networked and collaborative than ever before, and require the use of automated knowledge discovery tools that depend on unfettered access and re-use conditions, and widely shared information in databases and publications. "Thickets" of patents resulting from disjointed legislative initiatives impede this process and lead to high transaction and litigation costs and growing anti-commons effects (Reichman and Okediji, 2012). Open source agreements seek to make the free exchange of materials possible without overturning existing intellectual property laws. As the BioBricks Foundation mission notes: "*Today, it is difficult to share and reuse genetically encoded functions due to high transaction costs associated with patent-based licensing (i.e., time and money). We aren't against patents per se. But we believe that biotechnology must move towards a free-to-use "dictionary" of biological functions that allow many people to benefit from all the potential creative and constructive uses of biology*…"(www.biobricks.org).

Following are three examples of initiatives to develop open source agreements: BiOS, Open Source Drug Discovery, and a collaboration between BioBricks and Open Plant:

Bios has developed different licenses and MTAs, including a detailed version adaptable for genetic resources; a simple version for seeds and plasmids that can also be adapted for other materials; a sample DNA Transfer Agreement that can be adapted; as well as shrink wrap MTAs for use of CAMBIA materials (www.BiOS.net; www.cambia.org). With these open MTAs, everyone knows the conditions under which material is transferred, there is legal clarity and recognition and attribution of material, but the process is flexible and easy, allowing materials and methods to spread through a "dynamically expanding group of those who agree to the same principles of responsible sharing" (BiOS, 2017).[20] Technology is available royalty-free to anyone in any country, for commercial or non-commercial applications. All agreements are non-exclusive; all licensees covenant to share improvements, making them available for use, even though they may be patented, to all other licensees; and participants share biosafety data and any other information needed to meet regulatory requirements for use in commercial products (www.bios.net).

This contrasts with typical MTAs which usually impose the condition that materials or technology be used only for certain purposes, often not allowing the development of commercial products without further negotiation. Open source MTAs also do not involve fees or royalties for the use of material or methods, which they consider to work against innovation. Instead, the user must agree to conditions that encourage cooperation and the development of technology – they cannot appropriate the fundamental "kernel" of the technology and improvements exclusively for themselves, and while the base technology remains the property of the entity that developed it, improvements must be shared as part of the protected commons. *"To maintain legal access to the technology, you must agree not to prevent others who have agreed to the same terms from using the technology and any improvements in the development of different products"* (BiOS, 2017).

Another example of an open source agreement is the Open Source Drug Discovery (OSDD) platform in India, which "addresses the potential problem of third parties acquiring proprietary rights based on the information available on its Portal, either pre-existing or generated by the OSDD community, or during the drug discovery process or otherwise, without contributing the improvements made thereon by them back to OSDD" (Bhadwarj et al, 2011;2011).

The BioBricks Foundation[21] and OpenPlant have also developed an open MTA, which incorporates many of the protections of traditional MTAs, such as protection from liability and no warranties, but also includes provisions that reflect the values of "open communities" including access, attribution, reuse, redistribution and non-discrimination. They describe these provisions as follows (www.biobricks.org):

- Access – materials available under the Open MTA are free of any royalty or fees, other than appropriate and nominal fees for preparation and distribution;
- Attribution – providers may request attribution and reporting for materials distributed under the OpenMTA, allowing researchers and their institutions to be credited for materials and data made openly available;
- Reuse – materials available under the OpenMTA may be modified or used to create new substances;
- Redistribution – the OpenMTA does not restrict any party from selling or giving away the materials, either as received or as part of a collection or derivative work; and

- Nondiscrimination – the OpenMTA supports the transfer of material between researchers at all types of institutions, including those at academic, industry, government, and community laboratories (www.biobricks.org/openmta).

User agreements

Similar in many respects to the open source agreements, user agreements are employed by some targeted databases and other groups. GISAID has developed an agreement, the Database Access Agreement (DAA) by issuing licenses on the use of data (https://www.gisaid.org/registration/terms-of-use/). While the GISAID database provides open access to the public, verified user identification is a requirement, providing the ability to enforce the license terms of the DAA. Under the agreement, users will: (1) share their own data and allow other users to access it; (2) not share or distribute data submitted to GISAID to other non-GISAID servers; (3) credit the use of others' data in publications; (4) make best efforts to collaborate with the originating laboratory; (5) analyze findings jointly; and (6) maintain common access to the technology derived from the data. GISAID users have the right to develop a commercial product based on data obtained, but should strive to collaborate with data contributors and may not impose any terms on the data itself. Intellectual property and other rights associated with the data are not forfeited when they are shared, but others may develop commercial products on the basis of data obtained (Elbe and Buckland-Merrett, 2017).

In addition to the condition of use notice on CAMERA noted above, the J Craig Venter Institute (JVCI) has negotiated more involved MOUs that address digital sequence information as part of its seawater collections of marine microbes, some of which were collected inside the territorial waters of other countries (http://www.jcvi.org/cms/research/projects/gos/collaborative-agreements/). In the MOUs, JVCI agreed to acknowledge the source country of origin in publications and elsewhere, and will "publish or publicly disclose genomic sequence data including a limited and reasonable description of the material consistent with generally accepted database curation standards." On their website they include sample agreements from Australia, Ecuador, French Polynesia, Mexico, New Caledonia, Seychelles, Tanzania, and Vanuatu (http://www.jcvi.org/cms/research/projects/gos/collaborative-agreements/). As part of their work, they collaborate with researchers from the countries of collection, who co-author publications, and they deposit data in GenBank and other databases that make the digital sequence information publicly available to any researcher worldwide.

The JCVI MOUs typically include five fundamental principles: 1. A purpose statement regarding advancing scientific knowledge of microbial biodiversity and humankind's basic understanding of oceanic biology, yielding insights into the complex interplay between groups of microorganisms that may affect environmental processes; 2. a clear commitment to making genomic sequence data from the study publicly available to scientists worldwide; 3. confirmation that intellectual property rights will not be sought by the Venter Institute on these genomic sequence data; 4. JCVI and its research collaborators will coauthor one (or more) scientific journal articles that describe and evaluate these genomic sequence data; and 5. JCVI will offer training opportunities to scientists and students in the countries where sampling is conducted (www.jcvi.org).

7. Digital Sequence Information, Biodiversity Conservation, and Sustainable Use

The use of digital sequence information supports biodiversity conservation and the sustainable use of its components in a range of ways. These extend from deepening our knowledge about biodiversity, identifying and mitigating risks to threatened species, enhancing our ability to track illegal trade in seafood, wildlife, timber and other products, through to identifying species and geographic origins of products and so allowing governments and consumers to make informed decisions about what they use and buy, and enabling more effective biodiversity planning. We review the use of digital sequence information in biodiversity conservation and sustainable use below, followed by a brief review of potential conservation and sustainable use impacts of technologies that make use of digital sequence information.

7. 1 Biodiversity Conservation

7.1.1 Identification and characterization of biodiversity

Understanding the Earth's biodiversity and its dynamic changes relies heavily on access to appropriate information, yet our knowledge of some of the most basic aspects of biodiversity remains inadequate, especially for microbes and invertebrates. Increasingly, cost-effective DNA-based diagnostic techniques are part of the toolkit of biodiversity researchers. DNA 'barcodes', for example, are now used extensively as an accurate means to identify species (Laiou et al, 2013). With the goal of genetically 'fingerprinting' five million specimens from 500,000 species in five years, initiatives such as the International Barcode of Life Project (ibol.org) and the Barcode of Life Data System (BOLD) have made important contributions towards supporting the Global Taxonomy Initiative (XI/29), the Strategic Plan for Biodiversity 2011-2020 and scientific and technical needs related to implementation of the Plan (XIII/31), and goals C (Target 13) and E (Target 19), of the Aichi biodiversity targets. These focus respectively on improving the status of biodiversity by safeguarding ecosystems, species and genetic diversity, and enhancing implementation through participatory planning, knowledge management and capacity building. Many governments increasingly rely on genetic sequence data to characterize their national biodiversity, including genetic, species, and ecosystem level diversity of a wide range of plants, animals, and microorganisms. In Brazil, for example, the use of genetic sequence data has helped to characterize the germplasm of fungi and plants, while in Canada, it has been used to manage natural forests and plantations (FAO, 2017; Canada submission, 2017).

The use of genetic sequence data has contributed significantly to the process of taxonomy and improving our understanding and knowledge of biodiversity, especially in cases where morphological identification is difficult. For example, comparisons of sequence data from specimens maintained in the world's museums and *ex-situ* collections allow the identification of cryptic and new species, while reducing the need to take samples from the wild. As a case in point, the use of such data and associated molecular techniques has proven to be a non-invasive approach to study the ecology, behavior, and conservation of mammalian carnivores (Palomares and Adrados, 2014). Genetic sequence data can also be used to identify fragmented samples, such as the remains of birds recovered from airplane engines, helping to inform mitigation strategies at airports and the design process for aircraft engines.

As noted above, advances in technology, reduced sequencing costs, and increased research and commercial interest in the genetic sequences of microorganisms has meant a rapid expansion of genetic sequence data available for taxonomists. As a result, there has been a significant advance in

phylogenetic studies and in the biological characterization and ecology of microbial species. The advances in metagenomics that make possible complete or near complete sequences from previously unknown uncultured microorganisms have also created an astonishing expansion of our understanding about the diversity, biology, ecology and function of microbial communities (Thompson et al, forthcoming; Eloe-Fadrosh et al, 2016; Garza & Dutih, 2015; Gilbert et al, 2014).

7.1.2 Conservation genetics and genomics: understanding genetic variability in populations

Whole genome sequencing is increasingly used as a tool to understand genetic variability in populations and to analyze relationships between populations. This helps to plan measures to minimize further genetic loss in endangered populations, or to identify invasive alien species or pests. Genomic studies of the critically endangered California condor, for example, are providing a model system for avian conservation genomics, enabling empirical evaluation of basic facets of transmission genetics, including segregation, linkage, recombination and mutation (Ryder et al, 2014). In the Democratic Republic of Congo, genetic sequence data has been used to develop a conservation strategy for a highly endangered population of eastern lowland gorillas which no longer had sufficient genetic variability for the colony to continue (Xue et al, 2015). Coral restoration strategies are drawing on genetic sequence data by comparing the genetic characteristics of different coral populations as potential candidates for reintroduction (e.g. Drury et al, 2016). In the United States, genetic sequence data has been used to identify, understand and mitigate factors that threaten populations of vulnerable plant and animal species such as manatees, the Western White Pine, and the Sierra Nevada Yellow-Legged Frog. Captive breeding programs for animals such as the black footed ferret, giant pandas and golden lion tamarin have also relied on genetic sequence data to reintroduce stable and healthy individuals to their natural habitats (US submission, 2017).

Understanding of evolutionary processes has also been enhanced by sequence data. It is now considered feasible to study entire genomes at a population scale, using thousands of samples, allowing for a much better understanding of how genetic diversity varies across the genome of an organism and how this diversity is shaped by evolutionary processes such as natural selection, genetic drift and recombination. Pan-genome studies allow hundreds or thousands of genomes from the same species to be analyzed simultaneously. The science is still evolving (Shafer et al, 2015), but sequence information already plays an important role in measuring the genetic diversity of different populations, and helping to identify how diversity can be conserved in different ecosystems.

7.1.3 Invasive species

Invasive alien species, including pests and pathogenic agents, are well recognized as a central threat to biodiversity as well as to agriculture, with Aichi Target 9 explicitly targeting their control or eradication as a priority and COP decision IX/22 recognizing the significance of DNA barcodes to facilitate identification of alien species and for agricultural border inspections[22]. Genetic sequences, and eDNA, provide important diagnostic tools for early detection, surveillance and management of invasive and agricultural pest species, and to distinguish those that are harmful from those that are beneficial and part of natural ecosystems (e.g. Ball and Armstrong, 2006; Hand et al, 2015). For example, using sequence information, researchers can calculate the likelihood of a non-native species becoming invasive in an ecosystem by determining their source populations, and thus their introduction pathways. Genetic sequence information, and the global availability of sequence data, are especially useful in the

context of alien species, given that they are typically not native to the country and are thus less likely to be known. No country holds sequence data for all of its biota and Parties can only obtain sequence data for these efforts through international databases (Natural History Museum, RBG Kew and RBG Edinburgh, 2017).

7.1.4 Understanding pollinators

Genetic information can also help with understanding pollinators, which support at least 35% of global crop production and most fruits (López-Uribe et al, 2017 and related articles in this special issue). For example, alarming declines in both commercial and wild bee populations due to pesticide and herbicide use, land use changes and pathogens have led to increased fears that current agricultural productivity may be jeopardized without concerted conservation efforts (Vanbergen and the Insect Pollinators Initiative, 2013). Research underway uses online databases and global research networks to compare genetic sequence information on bees and their pathogens.

7.1.5 Monitoring environmental change

Genetic sequence information is playing an increasingly prominent role in helping to monitor environmental change and develop models about the impacts of climate change on species and their distribution (Bacon et al 2015; Global Genome Biodiversity Network, 2017). In Canada, this approach has been used to identify threats to genetic diversity and changes in the distribution of forest tree species, enabling timely precautions to be set in place for species conservation. Metagenomics projects such as the EcoBiomics in Canada characterize aquatic microbiomes, soil microbiomes and invertebrate zoobiomes to test hypotheses to enhance environmental monitoring, assessment and remediation. Increasingly, genetic sequence information is used in agriculture to understand the role of genes that control plant growth, development and stress tolerance in different climates and their responses to environmental change (Canada submission, 2017).

Environmental DNA (or eDNA) provides an unprecedented ability to identify species present in different areas and biomes, and a powerful new tool for biomonitoring (Thomsen and Willserlev, 2015). Information on fish species can be found by analyzing eDNA from sea water, on soil organisms by analyzing soil, and on aquatic species by analyzing freshwater samples. The potential of such studies to estimate population size and genetic diversity and to aid environmental monitoring is significant (Bohmann et al, 2014; Barnes and Turners, 2016; Gilbert et al, 2014), along with its ability to provide insights to the study of ancient environments.

7.1.6 Ex situ conservation

Ex situ collections held for conservation purposes typically distinguish between the physical genetic materials that are stored and the sequence data found in digital databases. Digital sequence data is used as a comparison tool to define, differentiate, classify and explore biodiversity, and can also help to identify threats to biodiversity. Through molecular characterization, genetic sequence data plays an important role in *ex-situ* collections by eliminating duplicates in collections, reducing the costs of field collection, and ensuring that collected and conserved material provides a genetically representative picture of diversity (Pessoa-Filho et al, 2007). By enabling comparisons of representativeness across genebanks in which accessions are living (in contrast to collections of dead organisms held for

taxonomic purposes), it is also possible to identify accessions that may be at risk through inadequate representation.

The leverage of conservation benefits through the use of genetic sequence information is considered most effective when backed up by as many sequences as possible, in as accessible a manner as possible. Because no Party has the capacity to manage information on all of its biota, Parties typically rely on information generated and held elsewhere, with conservation supported by greater genomic and geographical coverage. Because of a reduced reliance on expensive fieldwork, it may in some cases also provide a cost-effective tool for conservation research.

7. 2 Sustainable Use

7.2.1 Tracking trade and wildlife trafficking

Genetic sequence analysis is a powerful tool for implementation of CITES and related agreements and supports the fight against illegal logging and seafood 'fraud', including the mislabeling of products. Databases containing sequence data have been used extensively to track illegal harvesting and trade; they comprise "reference libraries" for comparing specimens and samples that are confiscated by law enforcement officials (Manel et al., 2002; Degen et al, 2013). Using DNA sequence markers, it is possible to distinguish between wild and cultivated species; to identify the source of samples thought to be derived from threatened or endangered species; and to monitor processed products, which otherwise might be difficult to identify.

7.2.2 Developing new crops, and minimizing genetic erosion

Identifying and characterizing genetic resources is important not only for conservation, but also for the development of new foods, crops and other resources, especially in the context of climate change. A number of groups are developing crops that are resilient in the face of global threats like climate change, but also new pathogens, soil degradation, salinity and drought (www.openplant.org).

Aichi Target 13 has an explicit focus on maintaining the genetic diversity of cultivated plants and their wild relatives, and farmed and domesticated animals, and promotes strategies for minimizing genetic erosion and safeguarding this genetic diversity. With only a small fraction of food crop diversity characterized, genetic sequence data can expand our knowledge base to preserve the genetic diversity of wild relatives of cultivars and domesticated livestock, and minimize genetic erosion.

7.2.3 Pathogens and health emergencies

The application of digital sequence information is also invaluable in molecular epidemiology and tracing the phylogeny of the pathogens causing disease outbreaks. Tracing the origin of pathogens in emergency situations often includes sequence information which has, for example, been used in the Zika and Ebola outbreaks (Tyler, 2017), and on an ongoing basis through the EpiFlu database which includes genetic sequence data (www.gsaid.org). Genetic information is also increasingly used to manage disease outbreaks among livestock, such as foot and mouth disease, and in supplying appropriate vaccines for contingency planning in virus-free areas (Japan submission, 2017).

7.3 Conservation and sustainable use implications of technologies that use digital sequence information

In addition to its valuable role in conservation science, planning and management, digital sequence information is also integral to technologies and applications that have potentially positive and adverse effects on biodiversity conservation and sustainable use, some of which we review below.[23] Those growing from synthetic biology will be reviewed in greater detail by the SynBio AHTEG.

7.3.1 Potential positive impacts of technologies associated with digital sequence information

Proponents of the conservation benefits of technologies associated with digital sequence information, including synthetic biology, argue that they can reduce consumption of fossil fuels by relying on biological processes that use renewable raw materials to produce biofuels, and so can mitigate climate change. New technologies have produced cleaner, more efficient manufacturing processes that pollute less and reduce waste; microorganisms designed for bioremediation and biosensors to clean up pollution; and new manufacturing processes to produce chemicals, plastics, and drug-precursors currently extracted unsustainably from natural resources or synthesized from petrochemicals. As noted above, DNA sequencing has also contributed to agricultural breeding processes to identify key agronomic traits that are potentially useful for climate change adaptation, or for tolerance to certain environmental factors through the identification of functional molecular markers in plants or genomic selection in livestock (Scott et al, 2015; UNEP/CBD/SBTTA/20/8 March 2016; Piaggio et al, 2016).

Biotech applications might also increase farm productivity per acre and reduce the environmental impact of agriculture in some cases (The One Acre Study, www.novozymes.com). Synthetic biology could potentially be used to control invasive species, tackle threats to endangered species, and restore habitats through modification of genomes; it can reintroduce extinct alleles; and synthetic biology tools could be used to recreate extinct species - the controversial concept of species “de-extinction” (Kaebnick and Jennings, 2017; Redford et al, 2014; Redford et al, 2013; Desalle and Amato, 2017).

7.3.2 Potential negative impacts of technologies associated with digital sequence information

Concerns raised by these technologies for biodiversity conservation and sustainable use include potential unsustainable production of biomass and feedstocks that provide sugar to ‘biological factories’ producing biofuels, chemicals, plastics, pharmaceuticals, and other useful products. Concerns also center around the removal of ‘waste’ from forest and other areas to be used as feedstocks, since these are important organic matter for soils and ecosystem functioning. The clearing of so-called ‘marginal’ or ‘degraded’ lands for biomass production, with no definition of what constitutes ‘marginal’ or ‘degraded’, and for whom, has also resulted in reduced biodiversity in some areas. The enormous quantities of biomass required has placed pressure on land, in some cases displacing food crops, and the overall benefits for climate mitigation remain unclear, with some claiming these systems do not result in net reduction in CO^2 emissions (Webb and Coates, 2012; ETC, 2010; Laird, 2012). In some regions, demand for land for feedstocks has also displaced indigenous and local communities through land grabs, and resulted in social and economic disruption (Bagley, 2017; Scott et al, 2015; ETC, 2010). There are also concerns that the use of the label “natural” on synbio products like vanillin, saffron, artemisin and stevia could displace small farmer-grown products, rather than the petrochemical-produced products they are intended to supplant, thereby damaging local livelihoods (Bagley, 2017; TWN submission, 2017).

Other concerns include the fear that organisms comprised of genetic sequences or parts from diverse organisms, if released into the environment might become invasive species, toxic to other non-target organisms, or damage native genetic diversity as a result of their potential for survival, persistence and transfer of genetic material to other organisms (Scott et al, 2015; Sliva et al, 2015). Gene drives that spread traits aimed at suppressing or extirpating populations of disease vectors might also produce unknown effects on local species and ecologies. Fundamental to all risks identified is the unpredictability of both the positive and negative impacts of synthetic biology on biodiversity, and the limited exploration of social, economic, and cultural impacts (Scott et al, 2015; UNEP/CBD/SBSTTA/20/ 8 March 2016; ETC Group, 2016).

8. Digital Sequence Information, Fair and Equitable Benefit Sharing, and the Nagoya Protocol

The use of digital sequence information presents opportunities and challenges for benefit sharing. Awareness of ABS within industry and academic research communities is obviously a critical first step, and although awareness of ABS has grown since the Nagoya Protocol came into force, significant gaps remain (Laird and Wynberg, 2013; 2015). Challenges for benefit sharing also grow from very different views of the public goods that can be derived from digital sequence information. This includes different approaches to access, with some promoting the wide and free exchange of knowledge, materials and technologies to achieve public benefits, and others seeking to restrict access in order to "capture some of those benefits for a narrow and defined public" (Lawson and Rourke, 2016).

The nature of benefits has also shifted, with research collaborations, capacity-building, and technology transfer taking new forms through virtual sharing of software and technologies, genetic sequence data and parts, and cloud laboratories. New research arrangements, referred to by some as a "protected commons" (www.BiOS.org) or "contractually constructed research commons" (Reichman and Okedji, 2012), retain attribution and co-authorship as benefits, and in some cases more involved research collaborations, but eschew monetary benefits.

Across scientific disciplines that use digital sequence information there is widespread interest in seeking out data, samples, and metagenomes from around the world, and providing sequences and analysis to in-country counterparts, and the global community, in return. Sequencing and analysis of the genetic diversity of countries lacking capacity is seen as a form of benefit sharing. However, researchers providing samples or data sometimes have limited control over its use, and the amount of real capacity-building that emerges from research collaborations is varied.

In a departure from previous forms of high tech research and development, however, the capacity to undertake research using digital sequence information, including synthetic biology, is far more geographically dispersed than previously. The technology needed to engage at an advanced level is cheaper and more accessible than ever before, and research approaches more fluid and flexible. North America, Europe, and Asia still dominate these technologies, but there are many emerging research powerhouses like Brazil, South Africa, and Singapore, that can work as equal partners in synthetic biology and other research programs.

It is difficult to generalize about benefits that might result from the use of digital sequence information given the rapid and transformative nature of the science and technology associated with sequences.

However, a number of potential benefits, as well as challenges to benefit sharing, have emerged over the course of this research. In addition to more speculative monetary benefits that might accrue from the system that manages, disseminates, and uses digital sequence information, below we discuss new forms of non-monetary benefit sharing, in keeping with those identified in the Annex to the Nagoya Protocol. These include wider access to databases, knowledge and technology; technology transfer, capacity-building, and collaboration; and research directed at priority public needs. We also review some of the challenges to benefit sharing growing from digital sequence information use, including difficulties determining value; identifying providers and users; determining provenance of material; monitoring use; and distinguishing between commercial and non-commercial research.

8.1 Non-monetary benefits

8.1.1 Wider accessibility of databases, knowledge, and technology

An important form of benefit sharing is access to publicly available databases. Tax payers in the countries and regions that undertake the bulk of research using digital sequence information (the US, Europe and Japan), provide funds, expertise, and technological capacity to store, analyze and manage data within the public databases. Most countries do not have the funds or capacity to manage comparable systems, and so the INSDC databases serve as a resource for the global community. Every contributor of data or research results from around the world adds value to a shared global system, and in return gains access to the greater value of the collection. In addition, these databases house information, and provide analyses, on global biodiversity, and serve as an important resource for biodiversity conservation and sustainable use.

The Natural History Museum, the Royal Botanic Garden, Kew and Royal Botanic Garden, Edinburgh (2017) describe this view as follows: "…any modification of the current model of use of digital sequence information would risk limiting the non-monetary benefits currently available to Parties, and consequently the implementation of the Convention. The financial equivalence of these benefits has not been assessed, but before any action is taken it would be helpful to make this calculation and compare it (plus the implementation costs) to the revenues that might be generated by alternative models". This view is shared widely among researchers and database managers, who express concern that efforts to change the existing system in order to achieve benefits for a few would endanger greater benefits for the many.

However, others consider access to databases and technology an insufficient benefit since countries rich in biodiversity may lack sufficient molecular research capacity or biotechnology infrastructure to make use of global database systems, and some feel they lose control over national patrimony when DNA is sent overseas for more affordable sequencing and loaded onto public databases. A few have even found that the samples they share for analysis are presented at international meetings without advance notification, or without including them as authors (Elbe and Buckland-Merrett, 2017). Other researchers engage in successful collaborations with institutions based in higher income countries, but ironically this can create fragmentation within their own country, with researchers networking with the global scientific community, but not with each other.

Open access and open source databases

Benefit sharing associated with databases is impacted by the different approaches taken to access bulk sequence information. Approaches to database access exist along a gradient from the fully *open access*

public databases, through *open source* approaches, to systems that require fees and subscriptions for access and restrict the use of data. The two main approaches are *open access* (or *gratis, public domain*) and *open source* (or *formalized, libre, controlled access*) databases and registries of parts. The open access approach allows the free and unencumbered use of digital sequence information to fuel innovation and scientific research. The open source approach ensures smaller groups and individuals are not locked out of these innovations and technologies, can attach conditions to the use of data to ensure wider forms of benefit sharing, and might involve user agreements or MTAs. Although proponents of these approaches to access differ in their view of how to ensure the 'greatest good', both support making as much data publicly available as possible, for easy use by a wide range of researchers across the globe.

A number of researchers and policy-makers have explored these approaches using different language and conceptual frameworks. Lawson and Rourke (2016) distinguish between the supply of data and information from databases that is free of charge and without restrictions (*gratis)* and those that are subject to some limits on how the data and information might be used (*libre*). They suggest this distinction might provide "a possible avenue to the apparent conflict between open access to DNA, RNA and amino acid sequence data and ABS regulatory schemes" (Lawson and Rourke, 2016). Slobodian et al (2017) distinguish between a *public domain approach* – eg the INSDC databases – and an *open source approach* – eg BioBricks and BiOS - that can include conditions (eg materials must be available for multiple generations of users) while still permitting patenting and commercialization. Welch et al (2017) refer to *open access*, meaning once genetic sequence data is released it is unencumbered, in contrast to *formalized access*, which refers to the open source approach; they note, however, that neither case requires identification of provenance of the digital sequence information.

The PIP Framework uses *public domain* and *public access* databases to describe, on the one hand GenBank, a member of the INSDC system, and on the other hand GISAID (PIP Framework, 2011). INSDC databases do not require a data access or use agreement, nor registration or log-in. This contrasts with the identified user access required by GISAID and other specialized databases in which users register, explicitly accept terms of data access via a user agreement, and sign in. Elbe and Buckland-Merrett (2017) describe *public domain* databases as, rather, *anonymous access* databases, since a major difference between them and open source or specialized databases like GISAID is the identification of contributors and users. Clearly in cases of virus and pathogen data, biosecurity concerns would dictate the need to not only identify contributors and users, but to track use, with the added benefit of allowing researchers to acknowledge and potentially collaborate with genetic sequence data providers. GISAID was launched in 2008 in part as an alternative to public domain databases, so that data providers would have a choice in how they shared their data with the public. Although the sequence data of human viruses is not accessed from biological diversity in the same way as that associated with microorganisms, plants, and other organisms, the arrangements of moderately restricted access used by GISAID might provide useful lessons and insight to ABS discussions.

The open source community has developed a variant on open access that allows rapid and easy exchange of materials and sequences but requires users to join a community via a user agreement. This approach is more cumbersome in that users and contributors identify themselves, but it provides legal certainty to users, which open access does not. Open source also allows for a form of technology-transfer within the community, distinguishing between the tools of innovation (which should be freely available) and products (which can be patented). The open source approach grows from the idea that what open access and public domain approaches mean in practice is that larger companies and research

institutions can patent applications over genetic markers, targets, specific genotypes, and so on (with variation by country in what is patentable subject matter), while smaller groups that lack capacity are locked out (www.bios.net). As Reichman and Okedji (2012) describe it, these research groups are in effect creating a "contractually constructed research commons" that make it possible for genetic sequence-related research – which relies on exchange, collaboration, and the free flow of information – to flourish in an otherwise highly protectionist intellectual property environment.

Interestingly, much like the open source movement today, the original motivation behind the open access approach for databases was to maximize benefits to society and resolve the "moral tensions between different conceptions of credit attribution, data access, and knowledge ownership" of that time (Strasser in Lawson and Rourke, 2016; see Lawson and Rourke, 2016 for a valuable review of the history of open access and databases). The concept of serving science, society and humankind by making scientific data and information available "free of charge and without restriction" has been affirmed repeatedly over the years (eg The Bermuda Principles, 1996; the Fort Lauderdale Agreement 2003; the Toronto International Data Release Workshop, 2009), although some researchers have suggested revisiting these principles in light of recent advances in technology.[24]

Scientific publications

Journals increasingly require that genetic sequence data be deposited in the public international databanks, with an INSDC accession number, as a condition of publication. Databases work with publishers to ensure a flow of data into repositories for release before, or at the time of, publication, often creating embargo periods prior to publication during which data remains confidential.[25] This approach allows scientists to access these records to plan experiments and analyze published findings to support or refute their hypotheses, while ensuring that authors receive appropriate credit, and that this context remains linked to the underlying data (Cochrane et al, 2016; see, too: http://ncbi.nlm.nih.gov/genbank/submit).

This means that any researcher wishing to publish in their field for a global audience must lodge data with the international, open-access system, including researchers from high biodiversity countries working on domestic species. In this way, academic research places even domestic research using digital sequence information in the global public domain. However, if governments restrict this practice, and researchers could not publish as a result, it is feared that international researchers would stop working on the biodiversity of that country (NHM, RBG Kew, and RBG Edinburgh, 2017). One of the ironies noted by a number of researchers is that, should publication or use of digital sequence information be restricted by governments, or if industry cannot acquire legal certainty to use digital sequence information, research will shift (and already has in some cases) to countries that do not have ABS measures, or to non-Parties to the CBD or the Nagoya Protocol.

8.1.2 Technology transfer, capacity-building, and collaboration

Capacity development and research collaborations present a significant opportunity for benefit sharing. In a similar way to conventional biodiscovery, such benefits growing from the use of digital sequence information likely outweigh any potential financial benefits over time, and this is particularly important with sequence use, which is difficult to monitor. The nature of research collaborations associated with sequence information can be quite different from those undertaken for biodiscovery, however. They might occur through cloud labs, involve the sharing of software, materials and technology, the provision of samples in exchange for sequencing and analysis, and other exchanges that do not include bi-lateral

agreements, or perhaps even direct interaction between groups and individuals – much as with other manifestations of digital technology in our lives today. Technology transfer and collaboration might also include contests like iGEM, small grants, and other efforts to extend new technologies as widely as possible.

Many international researchers seek to help countries with less capacity participate in publishing their sequence data, and see this as an important form of benefit sharing. As the University of Guelph BIO (2017) remarks: "…more needs to be done to support developing country researchers in generating and publishing digital sequence information from their respective national genetic resources." However, a microbiologist makes the point that "publishing the data merely to benefit 'the larger community' has the potential to destroy its potential value as IP", and that there needs to be "some built-in protection for the provider to legally guard against the misappropriation of the rights to use the data for something more than another academic publication or sequences added to those amassing in the databases of developed nations".

In some cases, significant capacity exists to undertake advanced research on digital sequence information within developing countries and the limitation can be resources rather than expertise - something developed country research institutions, even those seeking to share benefits, do not always understand. For example, Massarani and Deighton (2017) describe concerns about a recently launched program working to involve more Latin American researchers in international bioinformatics projects, but which appeared to sideline them (despite 2,119 papers published on bioinformatics and computational biology between 1991-2016 from researchers in 19 countries in the region).

The GISAID User Agreement emphasizes research collaboration and attribution as a central form of benefit sharing. By identifying contributors and users of data, researchers can discover and properly acknowledge contributors, and any biosecurity considerations arising from the data can be addressed. GISAID's policy is to ensure all data contributors benefit, including the Originating Laboratory where the clinical specimen or virus isolate was first obtained, and the Submitting Laboratory where sequence data have been generated and submitted to the EpiFlu Database (Elbe and Buckland-Merrett, 2017). In many ways this approach – while more explicitly emphasizing research collaboration – resembles that of the open source groups which also include attribution, and emphasize benefits growing from data and technology sharing.

8.1.3 Research directed at priority public needs

Open science non-profit networks that share knowledge, technology and materials see the provision of these benefits as significant, but also view the broader research collaborations they spawn as contributing benefits to humankind. These collaborations address critical healthcare, environmental, food security and other challenges we face today. Much of this research is also intended to address the needs of marginalized or under-served communities around the world.

For example, as BioS describes it, biological innovation is at the very heart of sustainable and socially equitable development, and the problems and needs of "those most neglected in the high capital world can be served by the tools of informatics, communications and transformative biological understanding and technologies" (www.bios.net, Biological Innovation). The Open Source Drug Discovery (OSDD) Project in India uses global collaboration and exchanges to develop a new model of drug discovery that better serves developing country populations, and ensures the availability of drugs through lower cost

community drug discovery processes (Bhardwa et al, 2011; Bhardwaj et al, 2011). The model of benefit-sharing that emerges from this type of open source science is very different from that envisioned in ABS agreements to date. Rather than bi-lateral negotiation of agreements, with identified providers and users, these new models involve a global web of collaborators contributing in iterative ways to a final product that is openly available for use. In the case of OSDD, this includes research on developing country diseases that receive limited attention from large companies at present.

Additionally, the open science model envisions reciprocal benefit sharing in which everyone is a provider and a user. As one researcher describes the open source practice of returning results from work on accessed materials: "If a small lab uses a PhD student to modify a gene and contributes that modified gene back to Biobricks then that is a sharing of a benefit proportional to what they took and their capacity. If a company takes one hundred genes and carries out one hundred years' worth of work and contributes the results of that work back to Biobricks, then that is proportional to what they took in the first place and allows others to do more with the products they generated."

8.2 Monetary benefits

Monetary benefits growing from the use of digital sequence information are largely speculative to date, and are potentially complex due to challenges in identifying provenance and the value of any given sequence or part. The negotiation of monetary benefits through database and registry conditions of use notices, MTAs, licenses and user agreements, is generally deferred to a point in the future when a commercial product has been developed, although open source agreements usually eschew monetary benefits altogether. The practicalities of implementation remain undeveloped, however.

As a result of the uncertainties associated with monetary benefits from bi-lateral agreements, many have suggested the establishment of a global fund to address benefit sharing from public databases (e.g. Bagley, 2015 and 2017). Experience from funds established under the ITPGRFA and the WHO PIP Framework may provide relevant lessons in this regard.

Potential funding sources for a global fund include a kind of fair trade label that certifies companies' contribution to biodiversity conservation and benefit sharing (Jaspars, 2017). Some have suggested a standard access fee, or subscription, in which users pay a small charge for accessing a sequence, or an annual subscription. Given the blurring boundaries between commercial and non-commercial users, all might gain access on the same terms. Others have proposed that contributors of data pay to publish the data and monitor its use, thereby covering the additional costs incurred by monitoring.

Most database managers and researchers are opposed to any fee-based approach, however, given the significant cost and potential bureaucracy associated with creating a payment system and monitoring use. There is also concern that a fee-based system might isolate data or reduce the effectiveness of databases. Others from industry and research oppose a fee-based system because it would create a financial burden on users of digital sequence information although the value of the information from any one sequence may be limited (IFPMA submission, 2017). As the Natural History Museum and partners argue: "The 100 million search jobs run annually are not generating 100 million finance-generating outputs. Putting even a very small financial penalty on reading a sequence (were it to be possible) would outweigh the benefits generated and, given the number of sequences being seen, be unduly costly both for users and to implement" (NHM et al, 2017).

8.2.1 Determining the value of digital sequence information

Central to monetary benefit sharing is determining the value of digital sequence information. The intrinsic value of genetic data has increased alongside advances in science and technology (Nussbeck et al, 2016), but placing a monetary value on data is challenging. For example, products, processes and technologies growing from digital sequence information might involve genes from multiple countries and organisms combined together to create new biosynthetic pathways. Additionally, homologous, or identical, sequences vital to life, and in which natural selection has eliminated mutations, might be found in different organisms around the world. This means that if companies cannot acquire legal certainty for a sequence of interest in one country, they can search for, and often find, the sequence in another country. Further complicating matters is that sequence information is regularly modified and can be re-used indefinitely, raising questions about whether benefits attach to each transaction, or if there is a cut-off point after which benefit sharing does not apply.

Additionally, the value of digital sequence information is often found in the aggregate, rather than an individual sequence, when it is part of a larger collection of sequences within databases against which searches and analyses are run. As Welch et al. (2017) describe: "… an individual sequence or 'part' has more value in a library where it can be screened with other sequences to find the connections between a particular trait and its function and use in other things…As a result, the value of an individual sequence from a species may be very difficult to quantify".

The role and value of sequences within R&D are also very different from those of genetic resource samples in earlier forms of biodiscovery. As one researcher explains: "A sequence on its own does not have real value. Value begins with identification of a valuable trait, a characteristic of an organism that is of interest like drought-resistance, fungal resistance, or a slug whose slime's stickiness helps close surgical wounds… DNA sequences work in the opposite direction of observing these characteristics and then trying to find what produces the useful trait. With sequences, we have an enormous amount of material, but we do not know what it does… With bioprospecting we had big collections and screened everything looking for active compounds, but with DNA sequences we are one step further back in the process, because we don't know what compounds they generate…"

Finally, determining the value of digital sequence information is challenging because the type of data and information used varies significantly[26], and the relationship between a sequence or single piece of genetic material, and final products and processes, is complex[27]. The commercial applications of sequence information are also so enormously varied, and so rapidly changing, it is extremely difficult to characterize utilization of sequences, and their commercial value. Digital sequence information might contribute to the development of a commercial product, but it is also used to develop new industrial processes, research tools, or improved technologies that are not sold, and are shared freely.

Below we review in greater detail three core challenges to both identifying and determining the value of sequence information: the combination of genetic materials from many sources; homologous, or identical, genes; and the indefinite nature of sequence information and use.

<u>Combinations of genetic material from many sources</u>

Genetic material from diverse organisms, from around the world, is commonly combined in the development of new products, processes and technologies. A strain of yeast, for example, was engineered to produce thebaine, an opiate closely related to morphine, by engineering "the microbes to express a medley of 21 genes, some from yeast themselves, as well as others from plants, bacteria, and

even a rodent" (Service, 2015). Burgess and Berry (2016) describe mixing 12 enzymes from three spheres of life, including plants, humans and microbes, to create a new biosynthetic pathway that is more efficient at fixing carbon dioxide than plants.

The International Chamber of Commerce (2017) provides a case study on the development of a new consensus phytase to improve the nutritional value of animal feed, and notes that: "…in state-of-the-art bioinformatics projects, hundreds of thousands of (amino acid or nucleic acid) sequences may be used to develop a particular commercial product. The final product has a sequence that represents an 'average' of all input sequences; as a consequence, it is virtually impossible to determine the relative value of each individual input sequence."

In a final example, Jaspars (2017) describes a hypothetical case to illustrate both that genetic material is combined from many sources around the world, and that the relationship between what is used and the original organism varies significantly – both of which have significant implications for identification of provenance, and difficulties determining the value of each contribution. In this case, synthetic biology is used to combine genes from a number of different organisms, from around the world, into a vector, which is then incorporated into the host organism. The original molecule comes from a marine invertebrate from the Australian Great Barrier Reef. Genes are collected from:

- Brazilian reef organism - cloned without further modification, with the gene taken from the unmodified organism and inserted in the vector;
- Indian marine Cyanobacteria (blue-green algae) – in this case, the organism was collected by an Indian scientist, sequenced, and the whole genome is deposited online in a public database. The gene was then synthesized without modifications and incorporated in the vector;
- Canadian marine sponge – here the genome was sequenced and deposited in an online public database; the gene was synthesized with major modifications and incorporated in the vector.
- Gene from marine microorganisms were isolated from sediment obtained from a deep sea trench located in areas beyond national jurisdiction.

In considering the difficulties of benefit-sharing in such scenarios, Dutfield (in Scott and Berry, 2017) explores the concepts of "cognitive and material distance" of a resource from a final product, as well as "quantitative proportionality" in which benefit-sharing is determined based on the contributions of respective knowledge and resources. As a researcher asked: "If a small percentage of a sequence is used in the creation of a synthetic product or protein, how does one value that and share benefits? This is complicated. What percentage do I recognize as part of the original organism – 5%, 15%, 70%? If I take 2% of a particular sequence, is it treated the same as if I used the whole sequence?" Furthermore, do companies using sequence information from multiple countries to develop a single product negotiate dozens of ABS agreements for its use? [28]

Benefit sharing under the Nagoya Protocol is based on a bi-lateral model in which a genetic resource links directly, in a relatively short amount of time, within a simple institutional framework, with identified providers and users, to a commercial product.[29] Tvedt et al (2016) explore changes in research and development in recent decades that have implications for determining the value of digital sequence information, and benefit sharing. They examine the case of Cyclosporine A (Sandimmun), developed from a sample collected in Hardangervidda National Park in Norway in 1969, and the more contemporary case of hydrolytic enzymes used in the advanced bioprocessing of lignocellulosic biomasses such as wood and by-products from the fish industry. The latter case involves the use of

digital sequence information in industrial biotechnology, and demonstrates the complex arrangements typical to new forms of research and development, in contrast to the earlier Cyclosporine A case. [30] For the hydrolytic enzyme research, multiple public and private institutions were involved, contributing innovation, investment and resources in different ways. Diverse sources of genetic information contributed to the research process, including sequences from public databases, libraries, and collections. Resources were modified and 'optimized' through protein engineering, and genetic information from different organisms was combined to, for example, increase stability at high temperatures or salinity, conferring traits that complement those contributed by other organisms (Tvedt et al, 2016).

Identical genes found around the world

Conserved – or homologous or identical – sequences are similar or identical sequences in DNA, RNA, proteins or polysaccharides occurring across species, or within different molecules produced by the same organism. These sequences are vital to life, and so natural selection has eliminated mutations, meaning the same sequence can be found in different organisms around the world. For example, a study comparing bacterial strains from different habitats in different hemispheres found up to 93% of the gene content was similar, and secondary metabolites identical (Thole et al, 2012; VBIO, 2017). In another study, researchers examining soil samples from New York City parks found ties to genes from many other parts of the world: "…a set of 11 representative compounds discovered elsewhere around the world -- such as the antibiotic erythromycin from the Philippines and the antifungal agent natamycin from South Africa -- are encoded by gene clusters that were observed within the city parks' soil" (Drugs from Dirt Project, www.drugsfromdirt.org; Science Daily, 2016).

Conserved sequences create complexities for benefit-sharing, including whether a single country should benefit from utilization of a sequence shared by many. Another challenge is that research might move to countries in which sequence data is most easily accessed. Remarked a molecular biologist: "There is a massive influx of data already – for example, roughly 115,000 bacterial genomes are stored in GenBank, and more all the time. Things are moving at an incredible pace. If I can't find pathways or genes from organisms from one country, I will move to another country – from one genetic background to another. Genetic material is shared across organisms, kingdoms, and countries, so it is harder to claim it is owned by a particular country. Geopolitical boundaries are human constructs. Just because India or Brazil or some other country wants to place restrictions on the material they hold doesn't mean I can't find something similar and just as useful in some other geographic area."

Additionally, users might seek 'favorable' jurisdictions where they can have legal certainty over resources (PSEL, May 2017; Vogel et al, forthcoming). As one industry representative said: "Homologous sequences can be from any part of the world, and we take the position that if you decide up front you want to be certain that what you are doing is legal, then you know there are only a very few places where you can have that certainty up front, where you know you have accessed that sample the right way, and that includes the US, as well as our home country." Another manager from industry echoed this: "If governments make it too complicated to use their genetic resources, and they do not have clear cut and simple ABS laws in place, then we will use sequences from another country where we can get legal certainty. If you don't know where something originates from, or if it is a place with unclear ABS laws, don't use it." This approach is not workable for agriculture companies, however. As a representative of the industry explains: "If you are looking at a crop, you must go to particular places where it is found. For example, for coffee you need to go to Ethiopia. If a certain pest is found in certain

countries, you must go there. And to begin with, you want the diversity from the megadiverse countries..."

Outside of agriculture, homologous or conserved sequences mean the value of any given sequence, or collection of sequences from a particular country, is likely to be diminished, since many sequences might be found in other regions, including countries that are not Parties to the CBD. As a result, companies are unlikely to invest significant resources to gain access to raw digital sequence information from a particular country.

Digital sequence information can be reused and shared indefinitely

A further complication for identifying and valuing sequence information arises with its reuse in perpetuity. Unlike physical samples, digital sequence information survives 'utilization' intact, and the public databases make all records permanently accessible as part of the scientific record (Cochrane et al, 2016). Synthetic or modified digital sequence information may also be created from long-standing, publicly-available genetic sequence data, much of which may not have links to the original genetic resource or country of origin.

This raises the question of how benefit sharing attaches to digital sequence information over time. For example, does each further transfer require "… additional permission and documentation resulting in long term and increasing litigation burden, [and] financial and time delays to research and innovation"? (IFPMA, 2017; VBIO, 2017). Slobodian et al (2017) ask: "Is there ever a point where the original genetic material has passed through so many stages of transformation that ABS requirements attached to the original material no longer apply?" Given that synthetic biology products can involve "multiple cycles of modification, transformation, and combinations of different components of DNA," Slobodian et al (2017) raise the option of "cut-off-points"[31]. Tvedt et al (2016) also suggest cut-off points after a maximum number of transfers, which over time obscure the product's origin or mean that other inputs in innovation would far outweigh the contribution of the original genetic resource. If each transfer triggered benefit sharing, it "could end up creating an exorbitant total sum of aggregated obligations through the value chain".

Each additional modification to a sequence would also be increasingly difficult to value in relation to previous modifications. As a researcher asked: "What percentage similarity of a gene sequence requires you to consider benefit sharing? Small introduced changes can have massive effects on the genes being used, turning them from unusable to very valuable. How would this be accounted for?"

8.3 Challenges to benefit sharing

Core elements of benefit sharing under the Nagoya Protocol are challenged by the emergence of digital sequence information, and the 'dematerialization' of genetic resources. These include tensions between open access and controlled access to data, and difficulties determining the value of digital sequence information, discussed above. Below, we review additional challenges to benefit sharing arising from the use of digital sequence information, including: identification of contributors, users, and the provenance of sequence information; monitoring utilization; and distinguishing between non-commercial and commercial research.

8.3.1 Identification challenges

Identification is the first step in monitoring and establishing an effective benefit sharing system (Garrity et al, 2009). In their study for the ITPGRFA, Welch et al (2017) describe "identification logic" as one of three key ABS principles that are challenged by the emergence of sequence information and synthetic biology. They note that ABS policies are based on the principle that control over access to resources grows from identification of sources, providers and users in order to establish agreements, but digital sequence information and "the proliferation of data, multiplication of users, varied importance of information about provenance and other factors" will mean that the underlying "ABS logic of identification will be subject to erosion over time" (Welch et al, 2017). Below we review the two primary identification challenges with regards to digital sequence information: identification of contributors and users, and identification of provenance.

Identification of contributors and users of digital sequence information.

The bulk of digital sequence information is accessed through public databases, which do not require contributors or users to register or log in, agree to terms and conditions, or sign user agreements. Internal policies, and the governments that fund the databases, require that such databases do not erect barriers to free access, or apply conditions to their use; this might be understood to include ABS conditions, and user and contributor identifications. The INSDC databases do not take responsibility for assessing the ownership and conditions of use, and explicitly avoid placing any legal or other restrictions on the use of data; they instead require submitters of sequences to receive any necessary consent or authorizations prior to submitting sequences, and ensure the accuracy and quality of data submitted (Cochrane et al, 2016; see INSDC policy Annex 2). [32]

However, many of the hundreds of specialized sequence databases directed to particular organisms, gene groupings, or diseases have developed policies and regulations, including the protection of personal privacy and confidentiality. These might indicate ways that, even when open access is a priority, limitations on the release of data might serve ABS objectives (Lawson and Rourke, 2016). [33] An example is the Global Initiative on Sharing All Influenza Data (GISAID)[34], which promotes the international sharing of genetic sequences and associated data, including virological, clinical, epidemiological and demographic information (if available) about the influenza virus. In this case, the identification of contributors and users of genetic sequence data serves multiple goals, and does not interfere with the timely sharing of data during health crises. The GISAID Database Access Agreement (DAA) that governs the database retains the principle of public accessibility, meaning that access to EpiFlu is free and open to anyone who positively identifies themselves and agrees to respect the rights of contributors.[35] Open source agreements similarly require that contributors and users identify themselves as part of joining the community of researchers, including providing usernames and passwords (www.biobricks.org).

Unique identifiers for researchers have also been proposed as a way to support ABS; these follow researchers through their careers, and link to publications. Unique identifiers could also potentially link to sequence data that is deposited in or accessed from databases. Funders might require unique identifiers for research projects they support, and journals for publications. An example of a persistent digital identifier already in practice is ORCID (https://ORCID.org/). As one manager put it: "In this day and age, when everyone is encouraged to reveal their identity through Facebook and what not, why must access to genetic sequence data be anonymous? With bioterrorism and other threats, it certainly seems time to track access."

Identification of the provenance of digital sequence information.

There are increasing efforts to better link original physical material with digital sequence information, including metadata on the location of specimen collections[36]. Many in the database and research community support inclusion of the provenance of digital sequence information, which is important for science, and might also support benefit sharing. As Petra ten Hoopen of EBI reported in a workshop held in November 2016, accurately and consistently recording provenance is extremely important. EMBL-EBI is involved in data standards development and collaborations to encourage best practice for provenance reporting, ideally beginning at the point of collection in a way that can follow samples through as sequences (Scott and Berry, 2017).

Explained a researcher: "The scientific community is well positioned to say where things come from – not because of ABS, but because of the need for scientific reproducibility… If a publication does not say where a sample came from, which collection or institution, then it is outside the norms of scientific behavior. Scientific integrity demands a linkage between a specimen and a digital sequence." Some journals already require this type of data (eg *Journal of Natural Products*), and curated databases that include metadata often missing from the larger public databases, are of significant value to researchers.

A number of groups holding specimens are working to link sources, physical samples, and international databases. Some have proposed adding GPS positions to sequence data as part of correlating genomes with organisms, and points of collection. A number of groups holding specimens are working to link sources and samples with international databases. One example is the Global Genome Biodiversity Network, with 65 members from 22 countries. GGBN aims to increase the number of sequence data that are vouchered, since voucher specimens in collections are "crucial for all molecular research to enable verification and transparency of taxon identification." (GGBN, 2017; Annex). The Consortium for the Barcode of Life (CBOL) and the International Barcode of Life Project (iBOL) make available a massive barcode library for more than 10 million species to support identification of species and strains globally, and in ways that might be helpful with identification of digital sequence information (www.barcodeoflife.org; NHM et al, 2017; US government, 2017). [37] The CIESM Charter also emphasizes the need for provenance recording and reporting in marine research (http://www.ciesm.org/marine/charter/CIESMCharter.pdf).

However, there are concerns about how effectively identification can work for sequence information, since taxonomic names are not unique or persistent, and they change over time (Garrity et al, 2009).[38] Microorganisms are difficult to consistently categorize at phenotypic and genotypic levels, and the definition of a unique genetic sequence for the purposes of ABS is thus fraught with complexity. At the same time, sequences of the same species, from the same habitat, might differ due to natural mutations, and these might occur very often and in a short time. If a sequence does not have a 100% match in the public databases, would it then be considered unique? (RSB submission, 2017; VBIO submission, 2017).

An additional identification challenge is that, unlike other digitally shared resources like music, movies and computer code, digital sequence information is not immediately recognizable as belonging to a particular source, and this problem increases as it undergoes modification (Slobodian et al, 2017). As a researcher commented, "It's easy to hide where your sequence came from. I can take a natural sequence and have it codon optimized in such a way that one could not determine the original gene sequence again". A final challenge relates to digital sequence information already in the public domain.

As Robert Friedman of JCVI put it: "Once information goes into the public domain, to keep saying 'that's mine, and so you're bound by some rule' seems a very difficult thing to pull off" (in Servick, 2016). However, although there remain significant challenges to comprehensively identifying the provenance of sequences, there appears widespread agreement within the database and research community that, going forward, inclusion of the origin of digital sequence information is critical.

8.3.2 Monitoring the Utilization of Digital Sequence Information

Monitoring is critical for effective benefit sharing, yet genetic sequences are far more difficult to monitor than physical genetic resources. These challenges increase over time as sequences pass through multiple hands, are modified, and the unique identity of a sequence erodes. Challenges include, as noted, that much of the data currently held in databases lacks identification and origin, important parts of monitoring (Garrity et al, 2009). Additionally, annotated information that accompanies sequences published in international databases is not verified, which means it might not be reliable. The modification of digital sequence information over time in ways that make it unrecognizable also create enormous challenges for monitoring. Indeed, a single researcher in one step can fully erode sequence identity: "If I codon optimize a gene to produce a protein, the DNA will be unrecognizable and untraceable, but the product (protein) will be the same as before." Slobodian et al (2017) describe the challenges as follows: a "genetic resource may be sequenced, split into parts, shared in different registries and databases with different levels of reporting, modified, and combined with different genetic resources…". The most important step for digital sequence information monitoring is the inclusion of origin information in databases and registries, as is done by biorepositories, which is supported by international databases. These databases are not supportive, however, of calls for them to monitor data usage, which poses technical challenges, isolates data, and requires structures to pool information (ten Hoopen, EBI in Scott and Berry, 2017).

A number of groups are working to attach information on origin to sequences, and to include stronger links between physical samples and sequences. These include the INSDC and other databases, ontology and standards organizations, and some governments. A variety of approaches have been proposed, including 'watermarking' a DNA sequence in a non-coding region of DNA. The JC Venter Institute experimented with watermarking when developing Synthia, the first cell controlled by a synthetic genome. Watermarking has limitations, however, including difficulties scaling up to large quantities of sequences, susceptibility to degradation (eg through mutation), and removal of the watermark by third parties (Bagley, 2017; Yamamoto et al, 2014; Slobodian et al, 2017). The Global Genome Biodiversity Network (GGBN) Data Standard has been working on ways to share and use genomic sample material and associated specimen information in a consistent and open manner, including a vocabulary for permits and loans according to the requirements of the Nagoya Protocol. They are building a system that promotes transparency and accountability around ABS permits, including within the INSDC system (GGBN, 2017). [39]

Garrity et al (2009) describe the elements required in a monitoring system, and issues to consider, including: the need to accurately reflect current knowledge; regularly incorporate new knowledge; the legally required granularity of identification; the need to transcend existing institutional tracking systems but also ensure that the system is based somewhere that is stable and well-resourced over time; and the possible role of a trusted third party to manage ABS monitoring systems (based in an existing institution or system, rather than creating one from scratch). ABS systems could link materials

and information to relevant documents that provide PIC and MAT, and to other documents like MTAs. [40] Paul Oldham and others are developing a model to assist countries with adapting national permit systems to facilitate ABS monitoring. This will typically involve the creation of a coordinated system where separate national authorities continue to own their parts of a permit database (normally under statute) but join them together so that users access a single portal that would allow them to get whatever permit they need (http://abspermits.net/). This would give each permit a unique identifier, and possibly a QR code that, if used by the relevant researchers, collections and others potentially links the permit to collection specimens, vouchers, digital sequence information, and so on, and could also link to publications and authors (Oldham in Scott and Berry, 2017).

The WHO PIP Framework has taken the approach of monitoring and tracing the use of sequence data through end products. When the Framework was negotiated, there was an awareness that digital sequence information might be used independently of the physical sample to synthesize candidate vaccine viruses, virus proteins, and antibodies. Therefore, there is a means to monitor physical samples through the Influenza Virus Traceability Mechanism, but this has been difficult when no physical sample is used. As a result, the PIP Framework set up a Technical Expert Working Group (TEWG) on Genetic Sequence Data. Areas the TEWG considered include the potential to meet benefit-sharing by monitoring and tracing the use of genetic sequence data from commercial products and including technical mechanisms to trace or monitor downloading of genetic sequence data from databases. They also are exploring the use of influenza-related products like regulatory approval files and patent applications (PIP Expert Working Group, 2015; see, too, discussion of tools used to mine patent literature and reveal uses of patent strains by companies over time in Parker and Garrity, 2010).

By working back from commercial products and utilization, the PIP Framework seeks to ensure benefit-sharing results from open-access regimes, but there are significant questions about how this will work in practice, and it is unlikely to be as effective as linking digital sequence information to provenance early in the research process (TWN, 2017). Additionally, the scope of products the PIP Framework must evaluate – those using influenza virus sequence information – are modest in size compared with the vast and sprawling applications of digital sequence information in all other fields, and the value of this model for monitoring use of sequences in other sectors is likely limited.

Some are skeptical of the potential to monitor digital sequence information in any meaningful way, and express concern about the management, bureaucracy and expense involved in adding layers of legal documents and information to databases. With assembled and annotated sequence datasets doubling every few years, and jobs run being more than a 100 million a year, one researcher asked “what methodology would allow you to check all of those permissions?” The University of Guelph, Biodiversity Institute of Ontario (2017), considers it “computationally impossible” to implement a mechanism to monitor the transfer of digital sequence information. The separation of legal and scientific databases has been suggested to address this concern. For example, scientific databases that hold sequence information could be separate from, but linked to, legal databases that are managed by governments and which contain permits and agreements associated with data.

<u>Information included in patent applications</u>

Information included in patent applications has received attention as a way to monitor the use of genetic resources. Some governments require patent applications for an invention based on or using biological material to include the origin of the material. Oldham et al (2013) used informatics techniques to mine patent databases for key data under the ITPGRFA, including varieties, accession codes, and

UPOV determination names, and concluded these techniques can identify patent applications in need of further scrutiny. In a number of countries, intellectual property offices are the official Nagoya Protocol checkpoints, which could assist in this monitoring approach.

While acknowledging the value of patent search engines in revealing the focus of commercial research more broadly, others are skeptical of their role in monitoring digital sequence information. Most sequence information never makes its way into patent applications or databases, and naturally-occurring, unmodified sequence data is neither eligible for patent protection or subject to other legal obligations in most countries. The US Supreme Court in 2013 found that isolated but otherwise unmodified gene sequences are not patentable subject matter because they are a product of nature, but cDNA might be patented; the Australian High Court found both unmodified isolated DNA and cDNA ineligible for patent protection (Slobodian et al, 2017). The primary data associated with biological sequences that is provided to national patent offices is also not yet comprehensive, standardized, timely and meaningful (Jefferson et al, 2015).[41]

8.3.3 Distinguishing between non-commercial and commercial research

Commercial and non-commercial research have very different implications for benefit sharing. However, the lines between them have grown indistinct in recent decades, as academic and government researchers increasingly partner with industry. Additionally, sequences move fluidly between commercial and non-commercial institutions, and once uploaded to public databases are available for all to use. When genetic resources or digital sequence information are accessed, it is also not always clear how the material and information will be used in the future. For example, samples or sequences might be accessed under academic research terms, uploaded onto databases, and eventually used commercially, potentially by multiple different users, without the original providers aware of or involved in this process.

As one microbiologist explained: “There is a lot of blurring of academic and commercial research. It is not at all clear that academic collections, ten years later, won’t be used for commercial purposes. Academic institutions are extraordinarily leaky in that way, things are shared across labs, material moves around. It is not clear how you would even control that in a university setting”.

Genetic resources or digital sequence information might be accessed under academic research terms, uploaded onto databases, and end up being used commercially, potentially by multiple different users, without the original providers aware of or involved in this process (Dedeurwaerdere et al, 2012). The World Federation of Culture Collections developed an MTA in 1993 alongside its MOSAICC code of conduct, which distinguished between commercial and non-commercial research, but in subsequent years realized that one rule for all users was more effective. The main issue, they determined, was monitoring use; anyone receiving collections could use them for any purpose. If that use became commercial, however, then the user must report back to the collection, which then contacts the original depositor (Scott and Berry, 2017).

The case of DivSeek illustrates the difficulties of drawing clear boundaries between commercial and academic research today, particularly when the objective of open source research groups is to promote all research, both commercial and academic. DivSeek’s mission is to accelerate crop improvement by building networks and facilitating the use, sharing, better characterization and tracking of plant genetic resources; it “advocates the application of state-of-the-art genomic, phenotyping and bioinformatics

technologies to enhance the quality, efficiency, and cost-effectiveness of germplasm conservation, provision and utilization for breeding…" (www.divseek.org). Proposed partnerships with the commercial companies DuPont and Syngenta to share access to sequences and patenting opportunities, however, have raised concerns about DivSeek's role as broker of data, information, and technologies, and the nature of the research they facilitate (Hammond, 2016). The trend over the last decade towards 'open innovation' and the free sharing of data is likely to create further blurring of the lines between academic and commercial research.

9. Conclusion

Digital sequence information is clearly a critical resource and tool for the conservation and sustainable use of biodiversity. The use of this information through transformative science and technologies also creates significant opportunities for non-monetary, and possibly monetary, forms of benefit sharing. There are, however, a range of challenges to realizing many of these benefits, linked in part to the difficulties of monitoring and identifying contributors, users and the provenance of sequences; the problems of determining value; and the increasingly grey area between non-commercial and commercial research.

It behooves ABS policy makers to stay abreast of the profound developments shaping research today. Sequencing platforms have become faster, cheaper and more accurate in recent years, producing massive quantities of sequence information. Researchers can now edit and synthesize genes. In the last year, new affordable and portable devices allow researchers to sequence physical samples, and then upload them to the internet or databases. Physical samples are still of interest to researchers, but their role in the research and commercialization process is changing, and the future is unclear.

Paralleling dramatic changes in science and technology are developments in the institutional, legal and social context of research. These include new, open and multi-party collaborations and diffuse research networks. Such collaborations are typically underpinned by a philosophy supporting unencumbered and free exchange of materials and technology, often as a way of serving the greatest public good, and to avoid intellectual property and transaction costs. New and significant benefits result from these innovative approaches, but use novel forms of benefit sharing that have not traditionally featured in ABS agreements. It might be that the strengths of ABS, open science, and other approaches could be combined in pioneering and inventive ways to develop flexible and adaptive policies that ensure benefits for the global community from the use of digital sequence information, including the important role it plays in the conservation and sustainable use of biodiversity.

ANNEXES

ANNEX 1: Ontology Projects

Ontology projects grew alongside the exponential growth of genomic data, and the need to capture these data electronically in a standard format. This has led in recent years to the inclusion of environmental data which helps to link sequences to their country of origin, something not commonly done previously. As Field et al (2008) put it: "…given the vast number of uncultivated microbes, it may be that a DNA-centric approach, in which genes are linked to habitats (locations), is more useful than the species-centered view".

The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products across databases. Founded in 1998, the project began as a collaboration between three model organism databases, FlyBase (*Drosophila*), the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD). The GO Consortium (GOC) has since grown to incorporate many databases, including several of the world's major repositories for plant, animal, and microbial genomes.

The GO project has developed three structured ontologies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. There are three separate aspects to this effort: first, the development and maintenance of the ontologies themselves; second, the annotation of gene products, which entails making associations between the ontologies and the genes and gene products in the collaborating databases; and third, the development of tools that facilitate the creation, maintenance and use of ontologies.

The use of GO terms by collaborating databases facilitates uniform queries across all of them. Shared vocabularies are an important step towards unifying biological databases, but additional work is still necessary as knowledge changes, updates lag behind, and individual curators evaluate data differently. The GO aims to serve as a platform where curators can agree on stating **how** and **why** a specific term is used, and how to consistently apply it, for example, to establish relationships between gene products. http://www.geneontology.org/

Growing from the Gene Ontology Consortium, **the Sequence Ontology (SO)** is a collaborative ontology project for the definition of sequence features used in biological sequence annotation. Contributors include the GMOD community, model organism database groups such as WormBase, FlyBase, Mouse Genome Informatics group, the Sanger Institute and EBI. SO is also part of the Open Biomedical Ontologies library (OBO), and has links to other ontology projects like RNAo Consortium and the Biosapiens polypeptide features. The SO describes its objectives are follows:

- To provide for a structured controlled vocabulary for the description of primary annotations of nucleic acid sequence, e.g. the annotations shared by a DAS server (BioDAS, Biosapiens DAS), or annotations encoded by GFF3.
- To provide for a structured representation of these annotations within databases. Were genes within model organism databases to be annotated with these terms then it would be possible to query all these databases for, for example, all genes whose transcripts are edited, or trans-spliced, or are bound by a particular protein. One such genomic database is Chado.

- To provide a structured controlled vocabulary for the description of mutations at both the sequence and more gross level in the context of genomic databases.www.sequenceontology.org

The Genomic Standards Consortium was founded in 2005 to promote mechanisms that standardize the description of genomes and the exchange and integration of genomic data, and the setting of minimum information about a genome sequence (MIGS) in order to promote participation in its development and discuss resources to improve mechanisms to capture and exchange metadata. They describe their aim as "making genomic data discoverable" by enabling genomic data integration, discovery and comparison through international community-driven standards. GSC brings together: 1) evolutionists, ecologists, molecular biologists and other researchers analyzing collections of genomes; 2) bioinformaticians producing genomic databases, 3) those who sequence genomes and 4) computer scientists, ontology experts and members of other standardization initiatives like the INSDC. As part of this effort, the Consortium sought to keep the process through which "minimal information" is supplied streamlined and 'minimal' in order "to encourage its adoption." (Field et al, 2008). www.gensc.org

A number of groups have been formed to standardize the design, documentation and assembly of synthetic-biology parts across academic institutions and industry. These include the US National Institute of Standards and Technology (NIST), launched by the Synthetic Biology Standards Consortium in March 2015; the Digital Imaging and Communications in medicine (DICOM) standard for sharing medical information, which is expanding to include synthetic biology; and the international Synthetic Biology Open Language (SBOL) which was established to provide researchers with a standardized vocabulary to describe genetic parts and circuits (Eisenstein, 2016).

http://sbolstandard.org/

https://www.nist.gov/property-fieldsection/engineered-biology-ensuring-quality-and-predictability-synthetic-biological

http://synbis.bg.ic.ac.uk/dicomsb/

ANNEX 2: International Nucleotide Sequence Database Collaboration Policy

Soren Brunak, Antoine Danchin, Masahira Hattori, Haruki Nakamura, Kazuo Shinozaki, Tara Matise, Daphne Preuss (2002) Nucleotide Sequence Database Policies**,** *Science 298 (5597): 1333 15 Nov 2002*

1. The INSD has a uniform policy of free and unrestricted access to all of the data records their databases contain. Scientists worldwide can access these records to plan experiments or publish any analysis or critique. Appropriate credit is given by citing the original submission, following the practices of scientists utilizing published scientific literature.

2. The INSD will not attach statements to records that restrict access to the data, limit the use of the information in these records, or prohibit certain types of publications based on these records. Specifically, no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party.

3. All database records submitted to the INSD will remain permanently accessible as part of the scientific record. Corrections of errors and update of the records by authors are welcome and erroneous records may be removed from the next database release, but all will remain permanently accessible by accession number.

4. Submitters are advised that the information displayed on the Web sites maintained by the INSD is fully disclosed to the public. It is the responsibility of the submitters to ascertain that they have the right to submit the data.

5. Beyond limited editorial control and some internal integrity checks (for example, proper use of INSD formats and translation of coding regions specified in CDS entries are verified), the quality and accuracy of the record are the responsibility of the submitting author, not of the database. The databases will work with submitters and users of the database to achieve the best quality resource possible.

Annex 3: Tracking Digital Sequence Information: Persistent Identification Schemes

(Garrity et al, 2009)

Concerns to address prior to implementation

- What will the identifier be identifying — the object, an abstract representation, or a physical object with associated metadata? How will the referent (the object which is identified) be precisely defined in such a way as to be understood by other users outside the control of the assigner? What metadata scheme will be used to do so?
- What will the identifier be required to resolve to: location, metadata, services?
- How can we avoid conflating —referent of the identifier‖ with —what the identifier resolves to‖ (not necessarily the same thing at all - though that may be intended!) – this conflation often arises due to the case with URL referencing.
- Does the identifier need to be globally or locally unique?
- What level of granularity is needed and will opaque or semantic identifiers be assigned?
- Are there legacy naming systems that need to be incorporated? If so, how will interoperability between naming systems be handled?
- At what point does an object change enough that it becomes a separate, new object for the purposes of an application (and so requires its own identifier?
- How will metadata be stored and bound to the identified object?
- Will the identification scheme of today be able to meet future needs?
- When is an identifier applied to an object and who will manage the identifiers over time?
- How will the assignment and long-term management of identifiers be financed?

BIBLIOGRAPHY – Digital Sequence Information

African Centre for Biodiversity 2017. *Submission of Views and Relevant Information on Potential Implications of the Use of Digital Sequence Information on Genetic Resources*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal, August 30.

Allied Market Research 2016. *Synthetic Biology Market By-Products (Synthetic DNA, Synthetic Genes, Software tools, Synthetic cells, Chassis organisms), and Technology (Genetic Engineering, Bioinformatics, Microfluidics). Global Opportunity Analysis and Industry Forecast, 2014 – 2020.*

Angerer, K. 2011. Frog tales – on poison dart frogs, epibatidine, and the sharing of biodiversity. *Innovation – The European Journal of Social Science Research*, 24(3): 353-369.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K. et al. 2000. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1): 25-29. doi:10.1038/75556.

Aswad, A. 2017. DNA sequencing and big data open a new frontier in the hunt for new viruses. *The Conversation*, August 4.

Australian Government 2017. *Digital Sequence Information on Genetic Resources.* Submission to the Secretariat of the Convention on Biological Diversity, Montreal, Notification 2017-037.

Bacon, C.D., Silvestro, D., Jaramillo, C., Tilston Smith, B., Chakrabarty, P. and Antonelli, A. 2015. Biological evidence supports an early and complex emergence of the Isthmus of Panama. *Proceedings of the National Academy of Sciences*, 112(19): 6110-6115.

Bagley M.A. and Rai A.K. 2014. *The Nagoya Protocol and synthetic biology research: A look at the potential impacts*. Virginia Public Law and Legal Theory Research Paper (2014-05).

Bagley, M.A. 2015. Digital DNA: The Nagoya Protocol, Intellectual Property Treaties, and Synthetic Biology. Virginia Public Law and Legal Theory Research Paper No. 11. http://dx.doi.org/10.2139/ssrn.2725986

Bagley, M. 2017. Towering wave or tempest in a teapot? Synthetic biology, Access and Benefit Sharing, and economic development. In: Frankel, S. and Gervais, D. (eds) *The Internet and Intellectual Property: The Nexus with Human and Economic Development.* Victoria University Press, Wellington.

Ball, S.L. and Armstrong, K.F. 2006. DNA barcodes for insect pest identification: A test case with tussock moths (Lepidoptera: Lymantriidae). *Canadian Journal of Forest Research,* 36(2): 337-350.

Bhardwaj, A., Scaria, V., Patra, D. and Open Source Drug Discovery Consortium 2011. *Science and Culture*, January – February 2011.

Bhardwaj, A., Scaria, V., Raghava, G.P.S., Lynn, A.M., Chandra, N., Banerjee, S., Raghunandanan, M.V., Pandey, V. et al. 2011. Open source drug discovery – A new paradigm of collaborative research in tuberculosis drug development. *Tuberculosis* 91: 479-486.

BIO, IFPMA and Cracknell, W. 2017. *Guidance on the sharing of influenza viruses with pandemic potential, genetic sequence data and information under the PIP Framework*. PIP Advisory Group Consultation with Stakeholders.

Biodiversity Institute of Ontario 2017. *Digital Sequence Information*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal.

Bohmann, K., Evans, A., Gilbert, M.T.P., Carvalho, G.R., Creer, S., Knapp, M., Yu, D.W. and de Bruyn, M. 2014. Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution*, 29(6): 358-367.

Boles, K.S., Kannan, K., Gill, J., Felderman, M., Gouvis, H., Hubby, B., Kamrud, K.I., Venter, J.C., and Gibson, D.G. 2017. Digital-to-biological converter for on-demand production of biologics. *Nature Biotechnology* 35: 672-675.

Bolser, D.M., Chibon, P-Y., Palopoli, N., Gong, S., Jacob, D., Del Angel, V.D., Swan, D., Bassi, S. et al. 2012. MetaBase—the wiki-database of biological databases. *Nucleic Acids Research*, 40 (Database issue): D1250-D1254. doi:10.1093/nar/gkr1099.

Broggiato, A., Vanagt, T., Lallier, L.E., Jaspars, M., Burton, G. and Muyldermans, D. (forthcoming). Mare Geneticum: Balancing Governance of Marine Genetic Resources in International Waters. Submitted to the *International Journal of Marine and Coastal Law*.

Burgess, S. and Berry, D. 2016. Regulating the use of genetic sequence data. *PLOS Synbio Community,* December 15, 2016. http://blogs.plos.org/synbio/2016/12/15/regulating-the-use-of-genetic-sequence-data/.

Carlson, R. 2014. *How Did the US Bioeconomy Perform in 2012*? https://synbiobeta.com/u-s-bioeconomy-perform-2012-rob-carlson/

Chadwick, L.H. 2012. The NIH roadmap epigenomics program data resource. *Epigenomics*, 4(3): 317-324. doi:10.2217/epi.12.18.

Church, D.M. and Hillier, L.W. 2009. Back to Bermuda: How is science best served? *Genome Biology* 10: 105. April 24.

Cochrane, G. and Galperin, M.Y. 2010. The 2010 nucleic acids research database issue and inline database collection: A community of data resources. *Nucleic Acids Research,* 38 (Database issue): 1-4.

Cochrane, G., Karsch-Mizrachi, I., Takagi, T. and INSDC. 2016. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, 44 (Database Issue): 48-50. doi:10.1093/nar/gkv1323.

Consortium of European Taxonomic Facilities (CETAF) 2017. *Digital Sequence Information on Genetic Resources – Benefits of their Use and their Public Availability for the Three Objectives of the CBD and Ramifications of Restricting Access to DSI.* Submission to the Secretariat of the Convention on Biological Diversity, Montreal.

Cressey, D. 2014. Biopiracy ban stirs red-tape fears. Nature News, *Nature* 514: 14-15, October 2.

Cressey, D. 2017. Treaty to stop biopiracy threatens to delay flu vaccines. Nature News, *Nature*, February 8.

Davis, K., Holanda, P., Lyal, C., da Silva, M., Fontes, E.M.G. 2016. *Implementation of the Nagoya Protocol on ABS: Dialogue between Brazil and the European Union*. European Union and Federal Republic of Brazil.

Deloitte, 2016. 2016 *Global life sciences outlook: Moving forward with cautious optimism.* https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Life-Sciences-Health-Care/gx-lshc-2016-life-sciences-outlook.pdf.

DeSalle, R. and Amata, G. 2017. Conservation genetics, precision conservation, and de-extinction. In: *Recreating the Wild: De-Extinction, Technology, and the Ethics of Conservation*. Special Report, Hastings Center 47(4), July-August.

Dedeurwaerdere, T., Melindi-Ghidi, P. and Broggiato, A. 2016. Global scientific research commons under the Nagoya Protocol: Towards a collaborative economy model for the sharing of basic research assets. *Environmental Science & Policy*, 55: 1-10.

Dedeurwaerdere, T., Broggiato, A., Louafi, S. Welch, E. and Batur, F. 2012. Governing Global Scientific Research Commons under the Nagoya Protocol. Chapter 15.

Droege, G., Barker, K., Seberg, O., Coddington, J., Benson, E., Berendsohn, W.G., Bunk, B., Butler, C. et al. 2017. GGBN - Strategies for Standardized Exchange of Genetic Resources on a Global Scale. In: Löhne, C., Zippel, E., Rohkemper, M. and Gardt, S. (eds) *Genetische Ressourcen, Gesetze & Gute Praxis: Wege zur Umsetzung des Nagoya-Protokolls in Deutschland*. Projektbericht. BfN-Skripten.

Drury, C., Dale, K.E., Panlilio, J.M., Miller, S.V., Lirman, D., Larson, E.A., Bartels, E. Crawford, D.L. and Oleksiak, M.F. 2016. Genomic variation among populations of threatened coral: *Acropora cervicornis*. *BMC Genomics,* 17: 286.

Eilbeck K. and Lewis S.E. 2004. Sequence Ontology Annotation Guide. *Comparative and Functional Genomics*, 5(8): 642-647. doi:10.1002/cfg.446.

Eilbeck K., Lewis S.E., Mungall C.J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. 2005. The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology*, 6(5): 44. doi:10.1186/gb-2005-6-5-r44.

Eisenstein, M. 2016. Living factories of the future. *Nature*, Technology Feature 531(7594): 401-403, March 16.

Eloe-Fadrosh, E.A., Paez-Espino, D., Jarett, J., Dunfield, P.F., Hedlund, B.P., Dekas, A.E., Grasby, S.E., Brady, A.L., et al. 2015. Global metagenomics survey reveals a new bacterial candidate phylum in geothermal springs. *Nature Communications*, 7(10476) (2016).

Elbe, S. and Buckland-Merrett, G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1: 33–46. doi: 10.1002/gch2.1018.

Environment Canada 2017. *Response to Request for Submission of Views on DSI.* Submission to the Secretariat of the Convention on Biological Diversity, Montreal.

Ernst and Young 2017. *Beyond Borders: Staying the Course. Biotechnology Report*, www.ey.com

Escalante, E., Barbolla, L.J., Ramírez-Barahona, S. and Eguiarte, L.E. 2014. The study of biodiversity in the era of massive sequencing. *Revista Mexicana de Biodiversidad*, 85 (4): 1249–1264.

ETC Group 2010. *Synthetic Biology: Creating Artificial Life Forms*. Briefing and Recommendations for CBD Delegates to COP 10.

ETC Group 2016. Four Steps Forward, One Leap Back on Global Governance of Synthetic Biology.

FAO 2017. Submission of information received by the Commission on Genetic Resources for Food and Agriculture on the use of "digital sequence information on genetic resources for food and agriculture" and potential implications for the conservation and sustainable use of genetic resources for food and agriculture, including exchange, access and the fair and equitable sharing of the benefits arising from their use. Submissions from the Brazilian and Canadian governments.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., et al. 2008. The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, 26(5), May.

Garrity, G.M. et al 1993. Genetic relationships among actinomycetes that produce the immunosuppressant macrolides FK506, FK520/FK523 and rapamycin. J Ind Microbiology 12(1): 42-47

Garrity, G.M, Thompson, L.M., Ussery, D.W., Paskin, N., Baker, D., Desmeth, P., Schindel, D.E. and Ong, S.S. 2009. Studies on Monitoring and Tracking of Genetic Resources. UNEP/CBD/WG-ABS/7/INF/2, March 2.

Garrity, G.M. and Parker, C.T. 2010. Recent trends in US patent grants and issues to be considered. Nature Precedings. http://precedings.nature.com/documents/4998/version/l/f/iles/npre20104998-1.pdf.

Garza, D.R. and Dutilh, B.E. 2015. From cultured to uncultured genome sequences: Metagenomics and modeling microbial ecosystems. *Cellular and Molecular Life Sciences,* 72: 4287-4308.

Gene Ontology Consortium 2008. The Gene Ontology project in 2008. *Nucleic Acids Research*, 36 (Database issue): D440-D444. doi:10.1093/nar/gkm883.

Global Genome Biodiversity Network (GGBN) 2017. *Digital Sequence Information*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal.

Gilbert, J.A., Jansson, J.K. and Knight, R. 2014. The Earth Microbiome project: Successes and aspirations. *BMC Biology*, 12: 69.

Grand View Research. 2017. *Biotechnology Market Analysis by Application (Health, Food & Agriculture, Natural Resources and Environment, Industrial Processing Bioinformatics), By Technology and Segment Forecasts, 2014-2025*. August, www.grandviewresearch.co

Hammond, E. 2016. *Digital genebankers plan to ignore UN request on the impact of genomics and synthetic biology on access and benefit sharing*. Third World Network, 4 April.

Hammond, E. 2017. *Gene sequences and biopiracy: Protecting benefit-sharing as synthetic biology changes access to genetic resources*. Third World Network Briefing Paper, August.

Hammond, E. 2017. Sequence Information: A Pressing Concern for the Seed Treaty. Third World Network Briefing for GB7, October, 2017.

Hand, B.K., Hether, T.D., Kovach, R.P., Muhlfeld, C.C., Amish, S.J., Boyers, M.C., O'Rourke, S.M., Miller, M.R. et al. 2015. Genomics and introgression: Discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Current Zoology*, 61(1): 146–154.

Handelsman J. 2009. Metagenetics: Spending our inheritance on the future. *Microbial Biotechnology*, 2(2): 138-139.

Heather, J.M. and Chain, B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 107: 1-8.

Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R. 2003. Proceedings of the Royal Society London B. Biological identifications through DNA barcodes. 270: 313-321. DOI: 10.1098/rspb.2002.2218. Published 7 February.

IFPMA 2017. *IFPMA Views on the Potential Implications of the Use of DSI on the Objectives of the Nagoya Protocol.* Submission to the Secretariat of the Convention on Biological Diversity, Montreal, September 7.

India, government of. 2017. *India's submission on Digital Sequence Information on Genetic Resources in responses to CBD Notification no. 865000*.

Intellectual Property Owners Association (IPO) 2017. *Proposed Application of Digital Genetic Sequence Information under the Nagoya Protocol*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal.

International Chamber of Commerce (ICC) 2017. *Digital Sequence Information and the Nagoya Protocol*. Prepared by the ICC Task Force on Access and Benefit Sharing. Submission to the Secretariat of the Convention on Biological Diversity, Montreal..

Japanese Bioindustry Association 2017. *Views of the Japan Bioindustry Association (JBA) on the Issues of Digital Sequence Information on Genetic Resources*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal, September 8.

Japan Pharmaceutical Manufacturers Association 2017. *JPMA's Comments on Digital Sequence Information on Genetic Resources*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal, September 7.

Jaspars, M. 2017. Categories of information and types of data incorporating different levels of processing and analysis, Presentation at IUCN Workshop entitled "Exchange of views on building a consensus on benefit sharing" at New York University Law School, New York, USA, 1 April 2017.

Jaspars, M. 2017. *Mare-Geneticum – The Science*. PHARMASEA, presentation.

Jefferson, O.A., Kollhofer, D., Ajiikuttira, P. and Jefferson, R.A. 2015. Public disclosure of biological sequences in global patent practice. World Patent Information, 43: 12-24.

Jefferson, O.A., Kollhofer, D., Ehrich, T.H., and Jefferson, R.A. 2015. The Ownership Question of Plant Gene and Genome Intellectual Properties. *Nature Biotechnology* 33(11): 1138-1143.

Kaebnick, G.E. and Jennings, B. 2017. De-Extinction and Conservation. In: *Recreating the Wild: De-Extinction, Technology, and the Ethics of Conservation*. Special Report, Hastings Center 47(4), July-August.

Laiou, A., Mandolini, L.A., Piredda, R., Bellarosa, R. and Simeone, M.C. 2013. DNA barcoding as a complementary tool for conservation and valorisation of forest resources. *Zookeys,* 365: 197-213.

Laird, S.A. and Wynberg, R.P. 2016. Locating Responsible Research and Innovation within Access and Benefit Sharing Space of the Convention on Biological Diversity: The Challenge of Emerging Technologies. *Nanoethics: Studies of New and Emerging Technologies*, 10(2), June.

Laird, SA. 2015. *Access and Benefit-Sharing: Key Points for Policy-Makers: Industrial Biotechnology.* www.abs-initiative.info

Laird, S.A. 2013. *Bioscience at a Crossroads: Access and Benefit Sharing in a Time of Scientific, Technological and Industry Change: Industrial Biotechnology*. Convention on Biological Diversity, Montreal.

Lawson, C. and Rourke, M. 2016. *Open Access DNA, RNA and Amino Acid Sequences: The Consequences and Solutions for the International Regulation of Access and Benefit-Sharing*. Griffith Law School Research Paper No. 16-12, Griffith University Law School, Australia. October 5.

Ledford, H. 2017. Artificial intelligence identifies plant species for science. *Nature News and Comment*, 11 August.

López-Uribe, M.M., Soro, A. and Jha, S. 2017. Conservation genetics of bees: Advances in the application of molecular tools to guide bee pollinator conservation. *Conservation Genetics*, 18(3): 501-506.

Manel, S., Berthier, P. and Luikart, G. 2002. Detecting wildlife poaching: Identifying the origin of individuals with Bayesian assignment tests and multilocus genotypes. *Conservation Biology,* 16: 650-659.

Mannheim, B. 2016. Regulation of synthetic biology under the Nagoya Protocol. *Nature Biotechnology*, 34(11), November.

Manzella, D. 2016. The Global Information System and Genomic Information: Transparency of Rights and Obligations. ITPGR for FAO, 24-25 November, Rome.

Martyniuk, E. et al 2017. Digital Sequence Information on Animal Genetic Resources for Food and Agriculture. Submission to the Secretariat of the Convention on Biological Diversity, Montreal.

Massarani, L. and Deighton, B. 2017. Latin American bioinformatics project attracts criticism. *SciDevNet*, August 4.

Ministry of Foreign Affairs, Brazil 2017. *Digital Sequence Information*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal.

National Academy of Sciences, Engineering and Medicine 2017. *Proposed Framework for Identifying Potential Biodefense Vulnerabilities Posed by Synthetic Biology: Interim Report.* Washington, DC, The National Academy Press.

Natural History Museum, Royal Botanic Garden Edinburgh, Royal Botanic Garden Kew 2017. *Potential Implications of the Use of Digital Sequence Information on Genetic Resources for the Three Objectives of the Convention*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal, 25 April.

NCBI 2017. *GenBank and WGS Statistics*. Accessed September 22, 2017.

NCBI Resource Coordinators 2013. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. 41(Database issue): 8-20. doi:10.1093/nar/gks1189.

Nussbeck, S.Y., Rabone, M., Benson, E.E., Droege, G., MacKenzie-Dodds, J. and Lawlor, R.T. 2016. "Life in Data" – Outcome of a multi-disciplinary, interactive Biobanking conference session on sample data. *Biopreservation and Biobanking* 14(1): 56-64.

Oldham P., Hall S. and Burton, G. 2012. Synthetic Biology: Mapping the Scientific Landscape. *PLoS ONE,* 7(4): e34368. https://doi.org/10.1371/journal.pone.0034368

Oldham P., Hall S., Forero, O. 2013. Biological Diversity in the Patent System. *PLoS ONE,* 8(11): e78737. https://doi.org/10.1371/journal.pone.0078737

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G.A., Papanikolaou, N, Kotoulas,G., Arvanitidis, C. and Iliopoulos, I. 2015. Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and Biology Insights*, 9: 75-88. doi:10.4137/BBI.S12462.

Palomares, F. and Adrados, B. 2014. The use of molecular tools in ecological studies of mammalian carnivores. In: Verdade L., Lyra-Jorge M., Piña C. (eds) *Applied Ecology and Human Dimensions in Biological Conservation*. Berlin, Springer.

Pauchard, N. 2017. Access and benefit sharing under the Convention on Biological Diversity and its protocol. What can some numbers tell us about the effectiveness of the regulatory regime? Resources 6(11), February 19.

Peruvian Society of Environmental Law 2017. *Lawful Avoidance of ABS: Jurisdiction Shopping and Selection of Non-Genetic-Material Media for Transmission*, submission to SBSTTA-21 and COP-14, May 1.

Peruvian Society of Environmental Law 2017. *Unpacking 'Digital Sequence Information on Genetic Resources': Scaffolding Errors to Preserve a Category Mistake*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal, July 30.

Pessoa-Filho, M., Belo, A., Alcochete, A.A., Rangel, P.H. and Ferreira, M.E. 2007. A set of multiplex panels of microsatellite markers for rapid molecular characterization of rice accessions. *BMC Plant Biology*, 7(23) doi:10.1186/1471-2229-7-23.

Pevsner, J. 2015. *Bioinformatics and functional genomics*. Wiley Blackwell, Chichester.

Piaggio, A.J. Segelbacher, G., Seddon, P.J., Alphey, L., Bennet, E.L., Carlson, R.H., Friedman, R.M. and Kanavy, D. et al. 2017. Is it time for synthetic biodiversity conservation? *Trends in Ecology and Evolution,* 32(2): 97-107.

Redford, K.H., Adams, W., Carlson, R., Mace, G.M. and Ceccarelli, B. 2014. Synthetic biology and the conservation of biodiversity. Fauna and Flora International, *Oryx,* 48(3): 330-336.

Redford, K.H., Adams, W. and Mace, G.M. 2013. Synthetic Biology and Conservation of Nature: Wicked Problems and Wicked Solutions. *PLOS Biology* 11(4): e1001530.

Reed, J., Stephenson, M.J., Miettinen, K., Brouwer, B., Leveau, A., Brett, P., Goss, R.J.M., Goossens, A. et al. 2017. A translational synthetic biology platform for rapid access to gram-scale quantities of novel drug-like molecules. *Metabolic Engineering* 42: 185-193.

Reichman, J.H., Uhlir, 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.

Reichman, J.H. and Okediji, R.L. 2012. When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale. *Minnesota Law Review* 96: 1362-1480.

Rockefeller University 2016. Survey of New York City soil uncovers medicine-making microbes. *Science Daily*. November 28.

Royal Society of Biology 2017. *Response from the Royal Society of Biology to the UK Government Request for Views and Relevant Information on the Potential Implications of the Use of DSI on Genetic Resources*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal.

Ruiz Muller, M.R. 2015. *Genetic resources as natural information: Implications for the Convention on Biological Diversity and Nagoya Protocol*. Routledge, London.

Ryder, O., Chemnick, L.G., Thomas, S., Martin, J., Romanov, M.N., Ralls, K., Ballou, J.D., Mace, M., Ratan, A., Miller, W. and Schuster, S. 2014. Supporting Califonria Condor Conservation Management Through Analysis of Species-wide Whole Genome Sequence Variation. In: International Plant and Animal Genome XXII Conference, 11-16 January 2014, San Diego, CA, USA.

Schei, P.J. and Tvedt, M.W. 2010. The Concept of "Genetic Resources" in the Convention on Biological Diversity and how it Relates to a Functional International Regime on Access and Benefit Sharing, Fridtjof Nansen Institute, Oslo, UNEP/CBD/WG-ABS/9/INF/1, March 19.

Schiele, S., Scott, D., Abdelhakim, D., Garforth, K., Gomez Castro, B., Schmidt, L. and Cooper, H.D. 2015. *Possible gaps and overlaps with the applicable provisions of the Convention, its Protocols and other relevant agreements related to components, organisms and products resulting from synthetic biology techniques.* Technical Series, Part II: Synthetic Biology. Convention on Biological Diversity, Montreal.

Schindel, D.E., Bubela, T., Rosenthal, J., Castle, D., du Plessis, P. and Bye, R. 2015. The New Age of the Nagoya Protocol. *Nature Conservation* 12: 43-56.

Scott, D. and Berry, D. 2017. *Genetic resources in the age of the Nagoya Protocol and gene/genome synthesis*. November 18, 2016 Workshop Report, the Engineering Life Project of the University of Edinburgh, and OpenPlant of the University of Cambridge.

Scott, D., Abdelhakim, D., Miranda, M., Hoft, R. and Cooper, H.D. 2015. *Potential positive and negative impacts of components, organisms and products resulting from synthetic biology techniques on the conservation and sustainable use of biodiversity and associated social, economic and cultural considerations. Technical Series, Part I: Synthetic Biology*. SCDB, Montreal, Canada.

Service, RF. 2015. Modified Yeast Produce Opiates from Sugar. *Science*, 349: 677, August 14.

Servick, K. 2016. Rise of digital DNA raises biopiracy fears. *Science*, AAAS, November 17.

Shafer, A.B.A., Wolf, J.B.W., Alves, P.C., Bergström, L., Bruford, M.W., Brännström, I., Colling, G., Dalén, L. et al. 2015. Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution*, 30(2): 78-87.

Singh, S. 2008. India takes and open source approach to drug discovery. Leading Edge Analysis, *Cell,* 133: 201-203.

Sliva, A., Yang, H., Boeke, J.D. and Mathews, D.J.H. 2015. Freedom and responsibility in synthetic genomics: The Synthetic Yeast Project. *Genetics,* 200: 1021-1028, August.

Slobodian, L.N., Lloyd-Evans, M. and Broggiato, A. 2017. The traceability of MGRs and genomic tech/synthetic biology. PharmaSea Milestones.

Society for Applied Microbiology 2017. *SFAM Position on Digital Sequence Information and the Nagoya Protocol*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal.

Solomon, D. 2013. *Industrial Views on Synthetic Biology*. Agilent Technologies, ACS Science.

Soren B., Danchin, A., Hattori, M., Nakamura, H., Shinozaki, K., Matise, T. and Preuss, D. 2002. Nucleotide Sequence Database Policies. *Science,* 298 (5597): 1333, 15 November.

Strasser, B.J. 2011. The experimenter's museum: GenBank, natural history and the moral economies of biomedicine. *Isis*, 102(1): 60-96.

Swetlitz, I. 2017. From chemicals to life: Scientists try to build cells from scratch. STATsnews.com.

SynbiCITE 2016. SynbiCITE: UK's First Commercial Synthetic Biology Foundry Goes into Production, April 2016, www.synbicite.com.

Third World Network 2017. *Potential Implications of the Use of Digital Sequence Information on Genetic Resources for the Three Objectives of the Convention*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal, September 6.

Thole, S. Kalhoefer, D., Voget, S., Berger, M., Engelhardt, T., Liesegang, H., Wollherr, A., Kjelleberg, S. et al. 2012. *Phaeobacter gallaeciensis* genomes from globally opposite locations reveal high similarity of adaptation to surface life. The *ISME Journal,* 6: 2229-2244.

Thomsen, P.F. and Willerslev, E. 2015. Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation* 183: 4-18.

Toribio, A.L. Alako, B., Amid, C., Cerdeño-Tarrága, A., Clarke, L., Cleland, I., Fairley, S., Gibson, R. et al. 2016. European nucleotide archive in 2016. *Nucleic Acids Research*, 29 November, (D1): 32-36.

Toronto International Data Release Workshop Authors 2009. Prepublication data sharing. *Nature* 461: 168-170.

Tvedt, M.W., Eijsink, V., Steen, I.H., Strand, R. and Rosendal, G.K. 2016. The missing link in ABS: The relationship between resource and product. *Environmental Policy and Law*, 46(3), 227-237.

UK Bioindustry Association 2017. *The BIA's Response to the CBD Secretariat's Consultation on the Impact of DSI Regulation in the Nagoya Protocol*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal.

US Submission 2017. United States submission on Digital Sequence Information on Genetic Resources, 18 August.

UWE 2016. *Science for Environment Policy - Synthetic Biology. Science Communication*. Future Brief 15, European Commission DG Environment. http://ec.europa.eu/science-environment-policy

Xia, L.C., Cram, J.A., Chen, T., Fuhrman, J.A. and Sun, F. 2011. Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One*, 6(12): e27992.

Vanbergen, A.J. and the Insect Pollinators Initiative 2013. Threats to an ecosystem service: Pressures on pollinators. *Frontiers in Ecology and the Environment*, 11(5): 251-259.

VBIO, German Life Sciences Association 2017. *Inclusion of Digital Sequence Information under the Scope of the Nagoya Protocol.* Submission to the Secretariat of the Convention on Biological Diversity, Montreal.

Vogel, J.H., Angerer, K., Ruiz Muller, M. and Oduardo-Sierra, O. (forthcoming). Bounded openness as the global multilateral benefit-sharing mechanism for the Nagoya Protocol, pp. 377-394, In: McManis, C.R. and Ong, B. (eds) *Routledge Handbook on Biodiversity and the Law.* Routledge, London.

Webb, A. and Coates, D. 2012. *Biofuels and Biodiversity*, Convention on Biological Diversity, Montreal. Technical Series No 65: 69 pp.

Welch, E.W., Bagley, M., Kuiken, T. and Louafi, S. 2017. *Potential implications of new synthetic biology and genomic research trajectories on the International Treaty for Plant Genetic Resources for Food and Agriculture (ITPGRFA or 'Treaty").* Draft study prepared for Special Event on Genomics Information, 28 October 2017.

Wellcome Trust and Wellcome Trust Sanger Institute 2017. *CBD – Call for Information: The Use of Digital Sequence Information on Genetic Resources*. Submission to the Secretariat of the Convention on Biological Diversity, Montreal, September 8.

World Health Organization 2011. *Pandemic Influenza Preparedness Framework for the Sharing of Influenza Viruses and Access to Vaccines and Other Benefits*, http://www.who.int/influenza/resources/pip_framework/en/

World Health Organization 2013. *Global Solidarity: Addressing our Health Responsibilities for Pandemic Influenza Preparedness*, http://www.who.int/influenza/pip/WHO_PIP_brochure.pdf?ua=1

World Health Organization, PIP Framework Advisory Group 2014. Technical Expert Working Group (TEWG) on Genetic Sequence Data. *Final Report to the PIP Advisory Group*, 10 October.

World Health Organization 2016. *Implementation of the Nagoya Protocol and Pathogen Sharing: Public Health Implications*.

World Health Organization, PIP Preparedness Framework 2017. *Meeting of the Pandemic Influenza Preparedness Framework Advisory Group, 28-31 March 2017, Report to Director General WHO*, Geneva.

Wynberg, R. and Laird, S. 2017. Fast Science and Sluggish Policy: The Herculean Task of Regulating Biodiscovery. *Trends in Biotechnology,* https://doi.org/10.1016/j.tibtech.2017.09.002.

Xue, Y., Prado-Martinez, J., Sudmant, P.H., Narasimhan, V., Ayub, Q., Szpak, M., Frandsen, P., Chen, Y. et al. 2015. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, 10 April, 348(6231): 242-245.

Yamamoto, N., Kajiura, H., Takeno, S., Suzuki, N. and Nakazawa, Y. 2014. A Watermarking System for Labeling Genomic DNA. *Plant Biotechnology*, 31(3): 241-248.

Zhi-Liang, H., Park, C.A. and Reecy, J.M. 2015. Developmental progress and current status of the animal QTLdb. *Nucleic Acids Research* 44 (D1): D827-D833.

1

ENDNOTES

[1] The outbreaks of H5N1 avian flu in 2006, and the reluctance of Indonesia to send samples of the virus to the World Health Organization (WHO) on the grounds that it required a more equitable system of access to vaccines for developing countries, catalyzed the development of a new global mechanism for virus sharing. After four years of negotiation, the Pandemic Influenza Preparedness (PIP) Framework was unanimously adopted on 24 May 2011 by the World Health Assembly. The PIP Framework aims "to improve pandemic influenza preparedness and response, and strengthen the protection against the pandemic influenza by improving and strengthening the WHO global influenza surveillance and response system ("WHO GISRS"), with the objective of a fair, transparent, equitable, efficient, effective system for, on an equal footing: (i) the sharing of H5N1 and other influenza viruses with human pandemic potential; and (ii) access to vaccines and sharing of other benefits" (WHO, 2011). It thus recognizes the need for the sharing of viruses and information about influenza, along with the benefits resulting from the sharing of that information. The Framework establishes some of the principles and rules for how this should be done and provides tools such as a Virus Traceability Mechanism, an electronic, internet-based system that records transfers of PIP biological materials into and within GISRS and from GISRS to parties outside. This system allows users to see where materials have been sent and allows them access to the results of analyses and tests on these materials. Standard Material Transfer Agreements (sMTAs) are used to cover all transfers of PIP biological materials within the WHO GISRS (WHO, 2011). In 2016, the Executive Board of WHO asked the Secretariat to prepare a study on how implementation of the Nagoya Protocol might affect sharing of pathogens, and the potential public health implications. Findings included potential enhanced benefit-sharing for Member States given that the Nagoya Protocol reinforces principles of fairness and equity by providing an opportunity to establish clear, pre-arranged benefit-sharing expectations arising from access to pathogens that will in turn enhance public health responses to infectious disease outbreaks (WHO Secretariat, 2016). Ongoing discussions recognize the relevance of the Nagoya Protocol in these deliberations and continue to explore the landscape for genetic sequence data (GSD) sharing and benefit sharing. This has included work in 2016 by the PIP Advisory Group Technical Working Group to document the "Optimal Characteristics of an influenza genetic sequence data sharing system under the PIP Framework" (http://www.who.int/influenza/pip/advisory_group/twg_doc.pdf?ua=1).

[2] A wet lab is a laboratory where chemicals, drugs, or other biological matter are tested and analyzed using water/liquids; in a dry lab, computers or computer-generated models are used for analysis.

[3] The main tasks involved in *NGS* data analysis include pre-analysis processing and quality control, genome assembly, de-novo genome assembly, short read mapping to a reference genome, variant calling, variant classification and annotation, genome wide association study (GWAS), and gene expression analysis (Griffiths-Jones, 2010).

[4] Metagenomics uses two approaches to prepare samples and generate digital sequence information (Oulas et al, 2015). In the first approach, environmental samples are sequenced directly (without the extra step of preparing clonal cultures prior to sequencing); this is known as "full shotgun metagenomics" (Xia et al, 2011). In the second approach, polymerase chain reaction (PCR) is used to amplify specific genes of interest before the sample is sequenced, thus ensuring that these specified genes will be sequenced and identified in the sequencing run. This second approach is termed "marker gene amplification metagenomics" (Handelsman, 2009).

[5] In paragraph 4 of decision XIII/17, the Conference of the Parties to the Convention acknowledged the definition developed by the AHTEG and considered it useful as a starting point for the purpose of facilitating scientific and technical deliberations under the Convention and its Protocols.

[6] In announcing the launch of the UK's first synthetic biology foundry, SynbiCITE, in April 2016, the CEO Dr Stephen Chambers, described its mission as follows: "To accelerate the translation of synthetic biology R&D into the marketplace. The Foundry has been created and built to operate as a 'cloud lab' to support synthetic biologists across the UK and is for everyone in the business of synthetic biology and who can use synthetic biology – the engineering of biology – in their business. These remote users send their biodesigns to the Foundry, which executes the work and delivers the data or prototype to the biodesigner once the work is complete. The Foundry provides a 'maker space' for entrepreneurial scientists looking to commercialize their research, ready access to state-of-the-art automation for SMEs, and is a facility for large and small companies to explore the enormous potential of synthetic biology."

[7] Nicola Patron of the Earlham Institute offers useful background on this process: Many parts with different origins, and different intellectual property claims, are used in combination, and are assembled in foundries using design software. A collection of plasmids housing the DNA parts (which are also produced from an automated process) are used, and might come from a collaborator or a synthesis company while others might be on hand in the freezer. The plasmids are used in the assembly reaction and this is then transformed into a chassis (organism), which is usually a bacteria that acts as an intermediary before the construct is delivered to the final cell or organism. At this stage a series of validation and characterization experiments are carried out to determine whether the circuit has assembled correctly. Ideally, all information from these characterization experiments will be returned to a Registry, informing future and new users about the specific functions of DNA parts – thus contributing to both understanding of organisms and potential commercial products. The synthetic circuit may be created from a mix of natural and synthetic genes. The chassis organism the circuit is put within might also have multiple benefit claims attached to it (Nicola Patron, Earlham Institute, in Scott and Berry, 2017).

[8] A rough estimate of the value of these activities in the US alone in 2012 came to $125 billion, with the bulk from chemicals and biofuels (Solomon, 2013; Carlson, 2014).

[9] Biotech drug sales – vaccines and biologics – were worth an estimated $289 billion in 2014 and are predicted to grow to $445 billion in 2019, totaling 26% of all prescription and over the counter sales by 2019. The majority of the top ten pharmaceutical products by sales in 2014 were biotech drugs, including monoclonal antibodies and recombinant products (Deloitte, 2016).

[10] A view of selected mergers and acquisitions (M&As) from 2016 provides a glimpse of the global and networked nature of the biotechnology industry, for example Shire of Ireland's acquisition of the US company Baxalta; the UK's Mylan acquisition of Sweden's Meda; and Astellas Pharma of Japan's acquisition of the German company Ganmymed Pharmaceuticals). (Debra Yu, Putting China's capital to work in the West, Ernst and Young, 2017; Grandview Research, 2017).

[11] A more complete review of the use of DSI in agriculture can be found in the ITPGRA study by Welch et al (2017) and the upcoming study by the Commission.

[12] Manzella (2016) summarizes applications in agriculture as follows: "Mapping the genetic variation of a crop onto the geographic landscape allows for prioritized collection. Genomic information allows for pedigrees and relatedness of germplasm in collections to be analysed, thus leading to informed genebank management (Wartmann, 2014). Genomic information guides selection for phenotypic evaluations for pre-breeding and development of introgression lines. Genomic information enables targeted breeding through advanced genotype and phenotype data analysis, to target agronomically significant genes by establishing causality between a particular trait and one or several loci in the genome and by providing molecular markers to detect trait inheritance. Having established that a given gene controls a given target trait, the breeder can select the gene directly, which is faster, less expensive and more reliable than the traditional approach of measuring the target trait." Examples of agriculture-related technologies associated with DSI include transgenesis; cisgenesis;

intragenesis; and targeted gene-editing (Welch et al, 2017; UWE, 2016). An additional summary of the use of DSI in agriculture today can be found in Welch et al (2017, pages 7-9).

[13] iGEM teams order genetic parts from the Registry as physical DNA samples, use them in their inventions, and contribute any modifications back to the registry, but as DNA synthesis becomes cheaper it is probable that users will synthesize the parts themselves from DSI (Slobodian et al, 2017).

[14] Databases are so central to genomic technologies that the journal *Nucleic Acids Research* has annual special issues on biological databases (published since 1993) and biological web servers (published since 2003) (http://academic.oup.com/nar).

[15] "INSDC partners have developed submission systems that guide users through the deposition of sequences, annotations and contextual data. These systems incorporate validations to ensure that deposited data is of high quality." INSDC supports "…standardization efforts driven by the expert communities for which sequence data is an essential tool. This includes the 'Minimum Information about any (x) Sequence standard (MIxS), which is developed by the Genomic Standards Consortium and the MINimal Contextual Data Checklist for pathogen surveillance data, which is developed by the Global Microbial Identifier (GMI) initiative. The MIxS relates to reporting on biological material provenance and experimentation procedure associated with genomes, metagenomes and marker gene sequences and has particular importance in environmental genomics. The GMI checklist relates to instructions for genome-scale pathogen sequence submissions, enabling the global identification of microorganisms and, ultimately, detection of outbreaks and new pathogens." (Cochrane et al, 2016)

[16] Ontology in computer science and bioinformatics means 'a formal naming and definition of the types, properties and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse' Other initiatives to manage and standardize the massive amounts of data generated by next generation sequencing include Amigo (http://amigo.geneontology.org/amigo), Biomedical Resource Ontology (http://bioportal.bioontology.org/ontologies/BRO), and Drug Target Ontology (http://drugtargetontology.org/). Other important biological ontologies can be found at: http://info.slis.indiana.edu/~dingying/Teaching/S604/OntologyList.html.

[17] Additional examples include MarBank, a marine genetic resource repository based in Norway comprised of marine organisms from field collections kept alive or processed and conserved in the biobank; and the NCI Open Repository, which contains extracts from 80,000 plants, and 20,000 marine organisms, all collected with the NCI letter of collection which addresses ABS issues, and access to which requires signing the MTA (Jaspars, 2017).

[18] Information associated with biospecimens that enhances their value includes that used to describe, annotate, and authenticate the biospecimen, and the processing and pre-analytical variables to which it has been exposed; permissions, including the ethical and regulatory documentation needed to acquire, transfer or collect biospecimens; associated data related to environmental or clinical information; and data that enable standardized access and exchange of information, like genetic sequencing data (Nussbeck et al, 2016).

[19] For example, the European Culture Collection Organisation has developed a standard MTA (www.eccosite.org) and the EU MICRO B3 (marine microbial biodiversity, bioinformatics and biotechnology) Consortium has adopted a model agreement for marine microbial research (https://www.microb3.eu/) (Dedeurwaerdere et al, 2016).

[20] BiOS refers to this as a "protected commons" in which exchanges are confidential and so protect future patent applications, but misappropriation by larger and better resourced groups is avoided. Patenting is still possible, and products and services can be developed for both profit and public good, but licensees and those who have used the technology under MTAs may not assert rights to exclude from use improvements (patented or not) by other licensees within the protected common. What is provided is not necessarily the product solution, but the enabling

technology that allow products to be developed by a range of individuals and groups. Unlike materials in the public domain, which can be patented by those with greater resources and so made unavailable for use by others, protected commons defers to the legal framework of patenting - "owners of improvements may wish to patent them, so we provide a space for confidential, non-public disclosure of improvements to all licensees." – but any improvements must be accessible to all other licensees, "so there is an incentive to protect the technology for open use." (www.bios.net).

[21] The BioBrick Foundation has also developed the BioBrick Public Agreement (BPA) for sharing the uses of standardized genetically encoded functions – eg BioBrick parts – or any genetically encoded function that contributors might own or make anew (www.biobricks.org/bpa/). The BPA is a contract between contributors and users, which - like the BiOS agreements - provides immunity from the assertion of IP, provides attribution for use of materials, requires respect for biosafety and other laws, and ensures contributors can't claim future rights against users who develop a new material or product. Users must provide usernames and passwords, disclose any IP associated with the parts, and get sign off from their employer in some cases if required to release materials into the public domain. There is no "give back" clause as with open software or the Open MTAs, so future parts and products are not required to be contributed to BioBricks.

[22] Also see Decision XIII/31 para 6g that encourages Parties to support the international barcode of life network, and applications of barcodes.

[23] In Decision XIII/17, the AHTEG on Synthetic Biology will continue to "analyze evidence of benefits and adverse effects of organisms, components and products of synthetic biology vis-à-vis the three objectives of the Convention, and gather information on risk management measures, safe use, and best practices for safe handling of organisms, components and products of synthetic biology."

[24] A number of principles and findings have been issued over the years affirming researchers' interest in releasing genetic sequence data as quickly as possible into the public domain in order to maximize benefits to society. The Bermuda Principles in 1996 address data from the Human Genome Project, and in 2003 the Fort Lauderdale Agreement affirmed the need for free and unrestricted use of genetic sequence data in biomedicine. The Toronto International Data Release Workshop in 2009 found that the rapid release of prepublication data has served the field of genomics well and recommended extending this practice to other biological data sets (Toronto International Data Release Workshop Authors, 2009).

[25] The withholding of data prior to publication has created a number of challenges to the international sharing of virus data critical to protecting populations against lethal infectious disease outbreaks. Researchers concerned that their scientific contributions would not be properly acknowledged and recognized, and whose professional standing and careers is tied to their publications and citations, have been unwilling to share data until their articles are published. The pressures do not lessen during outbreaks, when during such 'high-profile' times being 'first' matters more, and when researchers in this field also have an increased workload as part of assisting with control programs (Elbe and Buckland-Merrett, 2017).

[26] The nature and amount of DSI transferred or used can vary greatly, depending on the needs of the end-user, ranging from a few base pairs (eg data generated from a single Sanger sequencing run), to a large dataset with millions of base pair reads that was generated by NGS platforms. For example, researchers studying large-scale epigenetic effects on an organism under different conditions would run their large sequence data sets against existing databases to generate meaningful conclusions, but those interested in one gene with two polymorphs may only need to look at single Sanger sequence reads.

[27] As Bagley and Rai (2015) describe: ..."as biological science, including synthetic biology, moves away from a focus on individual full gene sequences towards a focus on parts of genes as well as the full genome and proteome, it is unclear how the notion of a 'functional unit of heredity' will map onto the new science".

[28] Brazil has focused on improving its ABS legal framework in light of experiences in recent decades, shifting the focus of regulation from access control to the control of results and a system based on registration and notification, using economic exploitation as the point at which benefit sharing obligations are raised (Brazil, 2017, Davis et al, 2016). However, it is still not clear how this would work for users, potentially negotiating with a dozen such parties for use of sequences, and with the value of each *in silico* contribution unclear. And as one Brazilian researcher put it: "… the basic regulations are still designed to address very traditional research of going to the field, collecting samples, doing extracts, and so on, they have not kept up with the times." Others expressed concern that national regulations could discourage researchers from sharing genetic sequence data with the public databases, which would undermine scientific research.

[29] In truth, however, ABS arrangements have rarely been straight forward, and digital sequence information complicates what is already complex regulatory terrain. Angerer (2011) describes the use of Epibatidine, an alkaloid originally extracted in the 1970s from the skin of a poison dart frog, *Phyllobates terribilis*, in Ecuador. After decades of research, and changes in the frog's taxonomy over the course of 30 years, today epibatidine is an important research tool that has opened up new avenues of research on nicotinic analgesics, rather than a substance of commercial value that is sold, with direct revenues. Given that the skin of poison dart frogs is used by indigenous people, there is added complexity around issues of benefit-sharing and traditional knowledge. The absence of a linear or simple ecological, research, economic, cultural, or legal context in this case illustrates the challenges that benefit sharing has always faced, including in earlier forms of biodiscovery.

[30] Although even this case is not as straight forward as it may appear. Cyclosporins were discovered as part of an antifungal screening program; the compound is a low molecular weight non-ribosomal decapeptide. The immunosuppressive properties were subsequently picked up in a screen for immunological agents. Cyclosporine was isolated from the fermentation broths of *Tolypocladium inflatum* in 1971 at Sandoz (which became Novartis) and was first used in transplant surgery in 1983. Like many such natural products, it has subsequently been shown to be ubiquitous in nature and is widely distributed across a number of ascomycetes (Garrity, pers. comm., 2017).

[31] Slobodian et al (2017) note that under the CITES agreement, after four generations of hybridization with non-CITES listed species, CITES protections no longer apply; in aquaculture, species are considered domesticated after three generations of controlled breeding.

[32] "The INSDC will not attach statements to records that restrict access to the data, limit the use of the information in these records, or prohibit certain types of publications based on these records. Specifically, no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party." The databases have disclaimers stating that if data is protected in some way by copyright laws, the user must determine this, and receive written permission from the copyright owners (www.ncbi.nlm.nih.gov/home/about/policies.shtml).

[33] These include the 2007 US National Institute of Health *Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies* (GWAS Policy), the 2014 *National Institute of Health Genomic Data Sharing Policy,* and the 2007 Organisation for Economic Cooperation and Development (OECD) *Principles and Guidelines for Access to Research Data from Public Funding*. Within the realm of DNA, RNA and amino acid sequence databases, new rules and principles have been developed to address the massive release of data into the public domain, including the *Principles for Proteomic Data Release and Sharing* (the Amsterdam Principles, 2008) and the *Toronto 2009 Data Release Workshop* best practices. These include pre-publication guidelines for different project types (eg genome sequencing, polymorphism discovery, genetic association studies, somatic mutation discovery, microbiome studies, RNA profiling, proteomic studies, metabolomics studies, RNAi or chemical library screen, 3D structure elucidation) (Lawson and Rourke, 2016).

[34] The WHO Collaborating Centers for Influenza (WHO CCs) provide scientific oversight and most GISRS laboratories use GISAID. GISAID is based on an understanding that the timely international sharing of health data is critical for protecting populations against lethal infectious disease outbreaks, but that without access to such information it is difficult to assess health risks, and to develop appropriate responses. GISAID contributes to global health in five ways, by: collating the most complete repository of high-quality influenza data; facilitating the rapid sharing of potentially pandemic virus information; supporting WHO's biannual seasonal flu vaccine strain selection process; developing informal mechanisms for conflict resolution around the sharing of virus data; and building greater trust with countries key to global pandemic preparedness (Elbe and Buckland-Merrett, 2017). In 2010, Germany entered into a public-private partnership with GISAID and has since hosted the publicly-accessible EpiFlu database, employing a unique sharing mechanism which ensures that the inherent rights of contributors of GSD are not forfeited. Some 650,000 genetic sequences had been deposited as of 2016, as well as a range of metadata including the date of specimen collection and specimen source.

[35] Positively identifying contributors and users is considered to ensure fair and transparent sharing of GSD, with all users mutually respecting the rights of contributors and other users. This mechanism is believed to provide contributors with the necessary incentive to rapidly share GSD, in the interests of Global Public Health. Access to GSD in EpiFlu can also be traced, permitting audits and providing the basis for an enforceability mechanism, and recourse should the need arise. It also makes it easier for scientists to discover and properly acknowledge those who contributed the data. GISAID is believed to work well because the data access agreement is very simple and there is a high level of trust and confidence that GSD is shared fairly whilst following the scientific etiquette of acknowledgement of the source of data (Elbe and Buckland-Merrett, 2017). GISAID is also exploring unique identifiers for their new database; for viruses, provenance is crucial, and the more than 1,000 institutions they work with have willingly identified themselves.

[36] For example, Metagenomes Online, a 'manually annotated resource of predicted proteins identified in viral and microbial shotgun metagenomes" (www.metagenomesonline.org), and the European Consortium of Taxonomic Facilities (CETAF) called for upgrading data management and curation systems to include or link to ABS legal documents (MTAs, licenses, etc.) and track sequence data in the large public databases to the original physical material (Manzella, 20116).

[37] The Barcode of Life Database (BOLD), based at the Biodiversity Institute of Ontario at the University of Guelph, coordinates on-going and international efforts to maintain and expand the global reference library of DNA barcodes as an open access online resource for DNA-based identification of living organisms. BOLD currently holds 1.3 million public records of the COI gene (www.boldsystems.org; University of Guelph, 2017).

[38] As Garrity et al (2009) note, there are significant challenges involved in linking sequences to taxonomic names, since earlier taxonomic identifications are not always accurate and undergo revisions. As they describe it: "… taxonomic names are commonly used in the scientific, technical and medical literature as well as in numerous laws and regulations governing commerce, agriculture, public safety and public health. But taxonomic names are not suitable for use as they are not unique, not persistent and do not exist in a one-to-one relationship with the abstract or concrete objects they identify." Efforts to address this challenge include the World Register of Marine Species (WoRMS), which makes synonyms and any changes to taxonomy or nomenclature easily discoverable (e.g. http://www.marinespecies.org/porifera/porifera.php?p=taxdetails&id=605442).

[39] The Global Genome Biodiversity Network (GGBN) Data Standard: "GGBN has developed the GGBN Data Standard (Droege et al. 2016) to complement existing biodiversity standards such as Darwin Core or ABCD. 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 (Droege et al. in press) 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, we are working on automated submission pipelines to INSDC, which includes permit information. This is an example of transparency and accountability regarding permits."

[40] They review a range of Persistent Identifiers schemes that would survive across the long time frames of genetic resource use (>20 years) including: Uniform Resource Name (URN); Persistent Uniform Resource Locator (PURL); Archival Resource Key (ARK); Life Science Identifiers (LSID); Handle System (Handle); Digital Object Identifier System (DOI). (See Annex for a list of the issues they identify as needing resolution prior to implementation of a Persistent Identification scheme; and see the original document Garrity et al, 2009).

[41] Another IP approach suggested by Lawson and Rourke (2016; Stemmer, 2002) might be a "copyright and database rights model" in which copyright subsists in the written representation of a sequence under copyright laws and database laws. They suggest this is an uncertain approach for sequence data which is scientific facts and findings, but suggest changing the form of expression, such as adopting a music format, or including watermarking (Lee, 2014). Where a copyright or database right exists another cannot copy without the permission of the rights holder subject to some exceptions, and this could be a restriction on commercial use without seeking specific ABS permissions (Lawson and Rourke, 2016).