UNEP

CBD

Convention on Biological Diversity

Distr.
GENERAL

CBD/DSI/AHTEG/2020/1/3
29 January 2020

ENGLISH ONLY

AD HOC TECHNICAL EXPERT GROUP ON
DIGITAL SEQUENCE INFORMATION ON
GENETIC RESOURCES
Montreal, Canada, 17-20 March 2020

DIGITAL SEQUENCE INFORMATION ON GENETIC RESOURCES: CONCEPT, SCOPE AND CURRENT USE

Note by the Executive Secretary

1. At its fourteenth meeting, the Conference of the Parties to the Convention on Biological Diversity requested the Executive Secretary "to commission a science-based peer-reviewed fact-finding study on the concept and scope of digital sequence information on genetic resources and how digital sequence information on genetic resources is currently used building on the existing fact-finding and scoping study[1]" (decision 14/20, para. 11 (b)).

2. Accordingly, and with the financial support from Norway and the European Union, the Executive Secretary commissioned a research team to carry out the study.

3. A draft of the study was made available online for peer review from 13 November to 11 December 2019.[2] The comments received in response have been made available online.[3] The research team revised the study in the light of the comments received and, in consultation with the Secretariat, prepared the final version as presented herein. Any views expressed in the study are those of the authors or the sources cited in the study and do not necessarily reflect the views of the Secretariat.

4. It should also be noted that this study is distinct but complementary to three other studies that the Executive Secretary was requested to commission pursuant to decision 14/20, paragraph 11(c), (d) and (e), and the synthesis of views prepared pursuant to decision 14/20, paragraph 11(a).

5. The executive summary of the study is presented below; the entire study is contained in the annex. The study is presented in the form and language in which it was received by the Secretariat.

EXECUTIVE SUMMARY

6. At the fourteenth meeting of the Conference of the Parties to the Convention on Biological Diversity, four studies related to digital sequence information on genetic resources were requested pursuant to decision 14/20, paragraph 11(b) to (e). This study is the first of those requested: "a science-based peer-reviewed fact-finding study on the concept and scope of digital sequence information on genetic resources and how digital sequence information on genetic resources is currently used building on the existing [Laird and Wynberg] fact-finding and scoping study".

[1] "Fact-finding and scoping study on digital sequence information on genetic resources in the context of the Convention on Biological Diversity and the Nagoya Protocol (CBD/DSI/AHTEG/2018/1/3).

[2] See notification 2019-094 of 22 October 2019.

[3] See https://www.cbd.int/dsi-gr/2019-2020/studies/#tab=0.

7. “Digital sequence information” (DSI) is widely acknowledged as a placeholder term for which no consensus on a replacement or precise definition exists to date. This study seeks, firstly, to ensure sufficient technical grounding with which to consider the concept of DSI by explaining the various types of information that can be understood to constitute DSI and providing context as to how such information is generated and used. The flow of information derived from genetic resources is shown in Figure 1, which is a key reference for the reader to understand the technical basis of this study. It builds on the “central dogma of molecular biology” (i.e. the processes by which DNA is transcribed into RNA, which, in turn, is translated into protein) to explain how the DNA of a genetic resource – whether obtained from a natural source or developed artificially – is used biologically. DNA, RNA, proteins and metabolites carry out the tasks and processes within organisms that we understand to be life. The figure also depicts different types of data that may be associated with a genetic resource and its derivatives, including genomic, transcriptomic, metabolomic, epigenomic data and metadata.

8. We also consider technical improvements and cost reductions in DNA sequencing which have led to “next generation technologies” that facilitate the sequencing of genomes from a single cell and entire ecosystems from environmental samples. As a result of these advances, DNA sequences and related information deposited in large open access databases on DSI continue to grow. Once a genome is deposited, its genes can be compared for similarities and differences to hundreds of other genes, thus helping understand its function and importance. Thus, building on half a century of scientific research, the utility of DSI is in the assembled data, not a single DNA sequence. DSI has many applications, including gene editing and synthetic biology.

9. This revolution in genomics has led to greater understanding of the tree of life and the function of genes and the metabolic processes with which they are associated. For example, epigenetics provides insights into hereditable changes without altering the DNA sequence. Transcriptomics informs on which genes are active in organisms and communities of organisms leading to greater understanding of interactions between organisms. Proteomics shows which proteins are expressed, and how they are modified. Metabolomics shows the complement of small molecules in organisms and provides a useful profile of metabolic activity and health status of organisms. These “omics” technologies, which are primarily aimed at the detection of genes (genomics), mRNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics) in biological and environmental samples, yield vast amounts of information associated with the underlying genetic resource, as depicted in Figure 1. Additional techniques, such as protein structure determination, codon optimization, and gene editing, rely on this information to enable modification of DNA, RNA and protein sequences in order to optimize expression, function or activity.

10. Technologies which are enabled by DSI are becoming ubiquitous in life-science-related research and industry. Understanding how the various types of information are generated and used in this context is essential in clarifying the concept of DSI, so we have chosen the following sectors to highlight for illustrative purposes: taxonomy and conservation (as indicative of basic research); agriculture and food security; industrial applications and synthetic biology; healthcare applications and discovery of pharmaceuticals. For each sector, we provide a brief overview of the sector accompanied by key trends and examples of the application of techniques technologies which are enabled by DSI in that sector. We show that each of these sectors is reliant on DSI, the disruptive nature of technologies/techniques enabled by DSI, and the significant economic footprint of several sectors. The same will be true for many other sectors in the life-sciences and these factors should be considered in the discussions regarding the scope of DSI and in assessing implications arising from the inclusion/exclusion of certain types of information from DSI subject matter.

11. Having established a technical grounding, the study seeks to clarify subject matter scope and terminology associated with DSI by proposing new logical groups to assist in evaluating the concept of DSI, and by identifying priority issues that will help determine whether certain types of information should be included or excluded from DSI subject matter. During the 2017-2018 intersessional period, Parties to the Convention and to the Nagoya Protocol took steps to attempt to clarify the concept of DSI. This process did not yield consensus on the appropriateness of the term “DSI” or what it refers to. These

challenges are not unique to the Convention or its Nagoya Protocol, as evidenced by comparable discussions under way in various other United Nations processes, such as the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA), the Pandemic Influenza Preparedness Framework (PIP) and the process concerning the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction (BBNJ). To help clarify the concept of DSI, we consider the flow of information from the utilization of a genetic resource, as depicted in Figure 1. It is evident that, at each step, the data/information it yields becomes progressively further removed from the original genetic resource. This proximity to the underlying genetic resource and additional information associated with each step provides a logical basis to group information that may constitute DSI. This gives rise to four alternative groups proposed to define the scope of DSI, summarized as follows (see Figures 6 and 7 in section 5):

- Group 1 - Narrow: concerning DNA and RNA
- Group 2 - Intermediate: concerning (DNA and RNA) + proteins
- Group 3 - Intermediate: concerning (DNA, RNA and proteins) + metabolites
- Group 4 - Broad: concerning (DNA, RNA, protein, metabolites) + traditional knowledge, ecological interactions, etc.

12. Group 1 has a narrow scope and proximity to the genetic resource and is limited to nucleotide sequence data associated with transcription. Group 2 has an intermediate scope and extends to protein sequences, thus comprising data and information associated with transcription and translation. Two interpretations for the scope of this group are possible: either subject matter is strictly limited to nucleotide and protein sequence data or it includes information associated with transcription and translation more broadly, for instance, functional annotations of genes, gene expression information, epigenetic data, and molecular structures of proteins. Group 3 has a wider intermediate scope and extends to metabolites and biochemical pathways, thus comprising information associated with transcription, translation and biosynthesis. Group 4 has the broadest scope and additionally includes information with the weakest proximity to the underlying genetic resource and extends to behavioural data, information on ecological relationships and traditional knowledge, thus comprising information associated with transcription, translation and biosynthesis, as well as downstream subsidiary information.

13. We use these four groups to evaluate a broad list of subject matters potentially comprising DSI as proposed in 2018 by the Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources. We also use these groups to evaluate a range of terms proposed to replace DSI, as shown in Table 4, which is a key reference for the reader to understand the different groups proposed to evaluate the concept of DSI in this study. It is evident from these evaluations that terminology is readily available to describe DSI with narrow subject matter as proposed in Group 1. These terms include Genetic Sequences (GS); Genetic Sequence Data/Information (GSD/GSI); and Nucleotide Sequence Data (NSD). Consequently, the terms Digital Sequence Data (DSD) or Genetic Resource Sequence Data and Information (GRSDI), could be used to describe subject matter of intermediate scope as proposed in Group 2. Depending on the interpretation, the term Genetic Resource Sequence Data (GRSD) could be used to describe either the narrow scope of Group 1 or an intermediate scope, as in Group 2. None of the terms proposed to date appear to adequately capture an intermediate range comprising information associated with the additional biosynthesis of a genetic resource as proposed by Group 3. Finally, terminology is also readily available to describe subject matter with broad scope as proposed in Group 4. Overall, the four logical groups proposed in this study provide a nuanced alternative to the 2018 list developed by the Ad Hoc Technical Expert Group, and so may better assist in clarifying the concept and scope of DSI; however, appropriate terminology will need to be evaluated, particularly for the intermediate groups.

14. The proximity of information to the underlying genetic resource is a useful proxy to determine whether it is possible to accurately identify or infer the source from which it is derived. This is possible to differing degrees in the case of RNA and protein sequences, however, it becomes much more challenging

with biosynthetic information and impossible with subsidiary information, which includes traditional knowledge, ecological relationships, sample metadata, taxonomy, biotic/abiotic environmental factors, phenotypic data, and behavioural data among other things. Accordingly, the proximity of data/information has significant implications for traceability to a genetic resource and also in identifying the source of information, including, subject to certain technical limitations, whether it has been generated through the utilization of a genetic resource or independently. In a system in which the traceability of DSI is important, a narrow scope of DSI subject matter appears better suited given the technical difficulties in identifying or inferring origin, whereas, if traceability is not important, a broader scope of subject matter may be able to be accommodated.

15. The study identified several key issues, as well as potential solutions which should be considered and resolved as a priority in order to help clarify the concept of DSI. The first two questions are: (a) how far along the flow from genetic resource onwards to DNA, RNA, protein sequences and metabolites "DSI" can be considered to extend; and (b) whether DSI includes both data and information and the extent to which data has been processed before it can be considered information. By using the four defined proposed groupings, the scope of DSI can adjusted for various contexts and clarity can be provided to inform further discussions. The third question is whether certain sequence information should be excluded from the scope of DSI subject matter, including sequences below a certain length, non-coding DNA, epigenetic heritable factors and modified DNA/RNA/proteins. A sequence below 30 nucleotides may not be unique, and this may provide a lower cut-off for sequence length. Non-coding DNA, epigenetic heritable factors and DNA/RNA/proteins modified naturally all have functions suggesting it might be logical to consider them for inclusion in the DSI subject matter. Conversely synthetically modified DNA, RNA or proteins cannot be said to have a natural functional role and so on this basis could be considered not to be an inherent part of the underlying genetic resource. Irrespective of whether the logical groups proposed in this study are adopted, it is anticipated that the priority issues identified in this study and illustrative insights regarding the extent to which DSI is used across a range of sectors in the life sciences, will assist the deliberations and the recommendations of the new Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources, which will consider the studies commissioned pursuant to decision 14/20 of the Conference of the Parties to the Convention on Biological Diversity.

16. Future deliberations concerning DSI may be aided by more comprehensive technical coverage regarding the use of DSI and technologies enabled by DSI, particularly in the context of commercially orientated research and development and including insights regarding the extent to which such uses of DSI are the subject of patent claims. Information of this nature would help further clarify the concept of DSI by facilitating more nuanced discussions concerning the possible implications of including or excluding particular types of information associated with an underlying genetic resource from the scope of DSI subject matter, within the context of the Convention and the Nagoya Protocol.

Annex

Digital Sequence Information on Genetic Resources: Concept, Scope and Current Use

Authors: *Wael Houssen[1,2], Rodrigo Sara[3], Marcel Jaspars[1]*

[1]Marine Biodiscovery Centre, Department of Chemistry, University of Aberdeen, Old Aberdeen AB24 3UE, Scotland, UK

[2]Institute of Medical Sciences, University of Aberdeen, Aberdeen AB25 2ZD, Scotland, UK

[3]Consultant to the Secretariat of the Convention on Biological Diversity

Table of contents

LIST OF FIGURES

Page

LIST OF TABLES

Page

LIST OF ABBREVIATIONS

BBNJ Process	Intergovernmental Conference on an international legally binding instrument under the United Nations Convention on the Law of the Sea on the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction
CBD	Convention on Biological Diversity
cDNA	Complementary DNA (Deoxyribonucleic acid)
CRISPR	Clustered regularly interspaced short palindromic repeats
DGR	Dematerialised genetic resources
DNA	Deoxyribonucleic acid
DSD	Digital sequence data
DSI	Digital sequence information
EVD	Ebola virus disease
GI	Genetic information
GMOs	Genetically modified organisms
GROs	Genomically recoded organisms
GRSD	Genetic resource sequence data
GS	Genetic sequences
GSD	Genetic sequence data
GSI	Genetic sequence information
ICC	International Chamber of Commerce
INSDC	International Nucleotide Sequence Database Collaboration
IPR	Intellectual property rights
ITPGRFA	International Treaty on Plant Genetic Resources for Food and Agriculture
IUCN	International Union for Conservation of Nature
LMO	Living Modified Organism
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
NP	Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity
NSD	Nucleotide sequence data
OED	Oxford English Dictionary
OTUs	Operational taxonomic units
PIP	Pandemic Influenza Preparedness Framework of the WHO
PNA	Peptide nucleic acids
POC	Point-of-care
RNA	Ribonucleic acid
SI	Subsidiary information
SMILES	Simplified molecular input line entry specification
WHO	World Health Organisation

xml	eXtensible Markup Language

1. INTRODUCTION

The 14th Conference of the Parties to the Convention on Biological Diversity requested four studies related to Digital Sequence Information on Genetic Resources.[4] "Digital Sequence Information" (DSI) is widely acknowledged as a placeholder term for which no consensus on a replacement or precise definition exists to-date. This study is the first of those requested: "a science-based peer-reviewed fact-finding study on the concept and scope of digital sequence information on genetic resources and how digital sequence information on genetic resources is currently used building on the existing fact-finding and scoping study".

The existing fact-finding study referred to in the decision is that by Laird and Wynberg published in 2018[1] and the Executive Secretary of the Convention on Biological Diversity commissioned the present study with the following two aims: 1) to explain in greater detail what types of information could be understood as DSI and how these are generated in order to help the process of determining what would be the most appropriate term and what it would cover; and 2) to explain how such information is used in different technological applications and life-sciences sectors in order to provide insights into how these might be affected by determinations regarding scope and the inclusion/exclusion of certain types of information from DSI subject matter. These inquiries regarding scope, terminology and the generation/application of different types of information that may potentially comprise DSI[5] will contribute to broader deliberations regarding different approaches for addressing DSI on genetic resources within the framework for access and benefit sharing established under the CBD/Nagoya Protocol.

This study is scientific in scope and does not cover associated policy implications. The work by the project team on this project took over 6 months and included a review of the primary and secondary literature, product documentation and websites belonging to institutes, research projects and companies.

To understand the subsequent discussion on scope and terminology, and to appreciate how this impacts sectors which use information that may comprise DSI, an understanding of the fundamentals of molecular biology and key developments associated with DNA sequencing and related technologies are essential. Accordingly, this study commences with a scientific background (Section 3) before evaluating different sectors in the life-sciences that rely on DSI and technologies/techniques enabled by DSI (Section 4). The study then considers the flow of data and information from a genetic resource and suggests new logical groupings for DSI subject matter, as well as evaluating alternative terminology to replace DSI and identifying priority questions/issues that need to be addressed in order to clarify the concept of DSI (Section 5). Implications for future discussions concerning scope and terminology which arise from this Study are considered (Section 6).

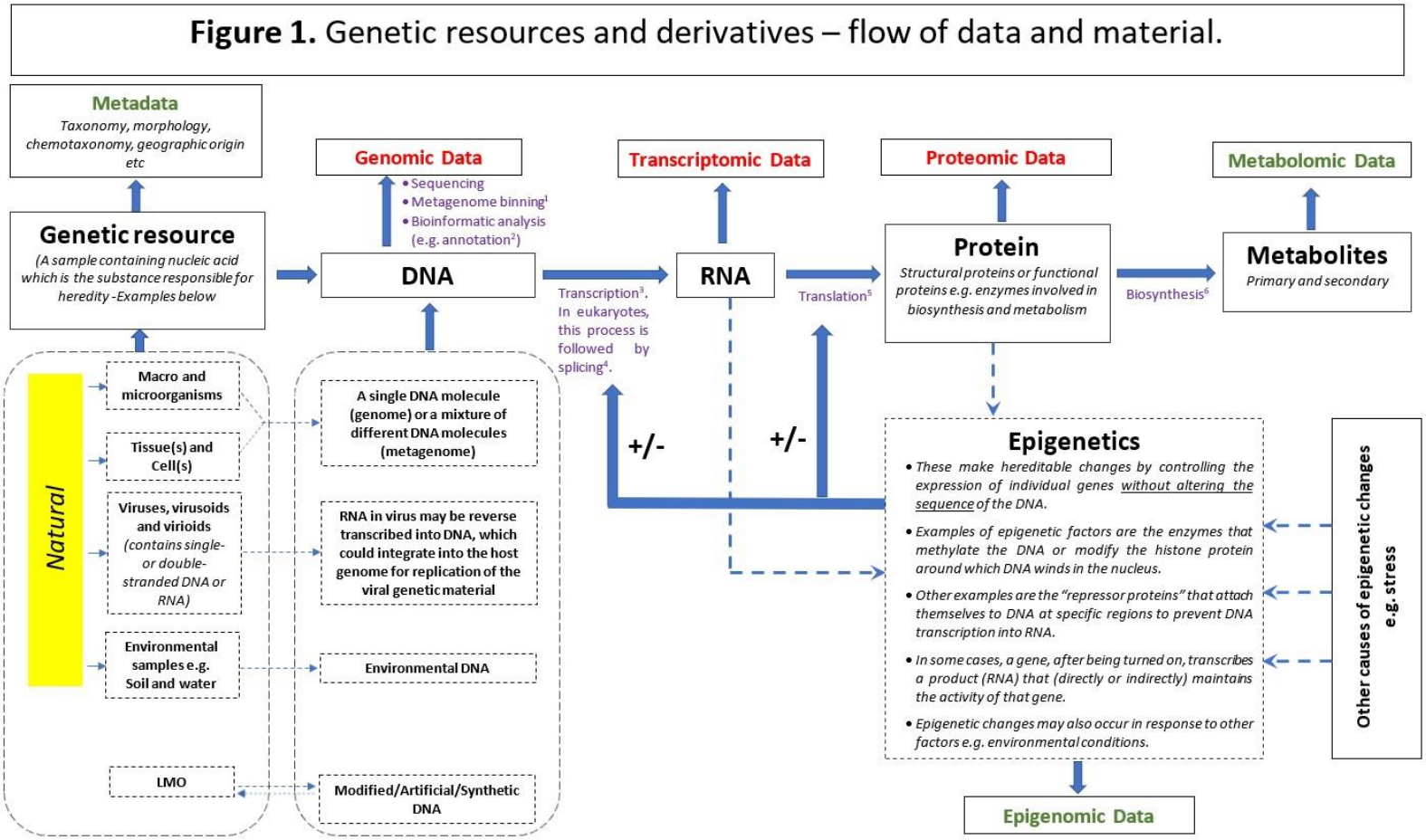
2. SCIENTIFIC BACKGROUND

The 'central dogma of molecular biology' represented in Figure 3 provides a basis for us to explain the structure of DNA and its copying mechanism, followed by the way in which DNA is translated into

[4] Decision 14/20, paragraph 11 (b) to (e), accessible at www.cbd.int/doc/decisions/cop-14/cop-14-dec-20-en.pdf
[5] The scope of "DSI" is of course yet to be determined, however, for convenience "information that may potentially comprise "DSI"" shall hereafter be used interchangeably with "DSI".

proteins and then biosynthesized into metabolites, including modifications that can occur at different stages of this process. We consider the relentless pace of technological advancement in this field in recent decades starting with DNA sequencing, followed by the ability to edit and engineer genes, the rise of synthetic biology, expansion of the genetic code and the emergence of 'omics' technologies, all of which generate or rely on information which may potentially comprise 'DSI'.

Throughout this section, the reader should refer to Figure 1 which provides a clear scheme showing the information flow from genetic resource to DNA, RNA, proteins and onwards to derivatives, using different techniques and approaches. This is used to explain how the DNA of a genetic resource – whether obtained from a natural source or developed artificially – is processed biologically, as well as the different types of information that may be associated with a genetic resource and its derivatives, including genomic, transcriptomic, metabolomic, epigenomic data and metadata. It builds on the 'central dogma of molecular biology' (as explained in Section 3.2 below) to depict the production of metabolites. The DNA, RNA, proteins and metabolites carry out the tasks and processes within organisms that we understand to be life.



[1]**Binning:** is the process of grouping sequencing reads of the metagenome and assigning them with certain groups of organisms.
[2]**Annotation:** is the process of identifying the locations of genes and the coding regions in a genome and determining what those genes do. This process is very research intensive.
[3]**Transcription:** is the first step in gene expression in which DNA is copied into RNA.
[4]**RNA Splicing:** is RNA processing step during which, introns (non-coding regions) are removed and exons (coding regions) are joined together.
[5]**Translation:** is another step in gene expression in which the base sequence information in RNA is converted into an amino acid sequence in proteins.
[6]**Biosynthesis:** is a multi-step, enzyme-catalysed process where substrates are converted into more complex products in living organisms.

Figure 1. Digital sequence information on genetic resources and derivatives. This figure shows the flow of information derived from genetic resources using different techniques and approaches

2.1 Discovery of DNA and its composition

The idea that a chemical structure could carry genetic information was suggested in 1944[2] and in the same year deoxyribonucleic acid (DNA) was identified as the substance responsible for heredity[3]. Subsequently it was discovered that: 1) DNA contains 4 nitrogenous bases, and in any double-stranded DNA, the number of guanine (G) bases is equal to the number of cytosine (C) bases and the number of adenine (A) bases is equal to that of thymine (T) bases; 2) the composition of DNA varies between species.[4] DNA is a double stranded helical structure (Figure 2). The two DNA strands are also known as polynucleotides as they are composed of simpler monomeric units called nucleotides. Each nucleotide is composed of a deoxyribose sugar, a phosphate group and one of the four nitrogenous bases (A, C, T, G). The deoxyribose sugar and phosphate groups form the backbone of each strand which resembles the sides of a ladder to which nitrogenous bases are connected. The bases face the center and each base is connected to and complements the base facing it in the opposite strand to constitute the "rungs" of the ladder: adenine in one strand is always paired with and complements thymine in the other whereas guanine is always paired with and complements cytosine as determined by DNA structural (hydrogen bonding) limitations.[5]

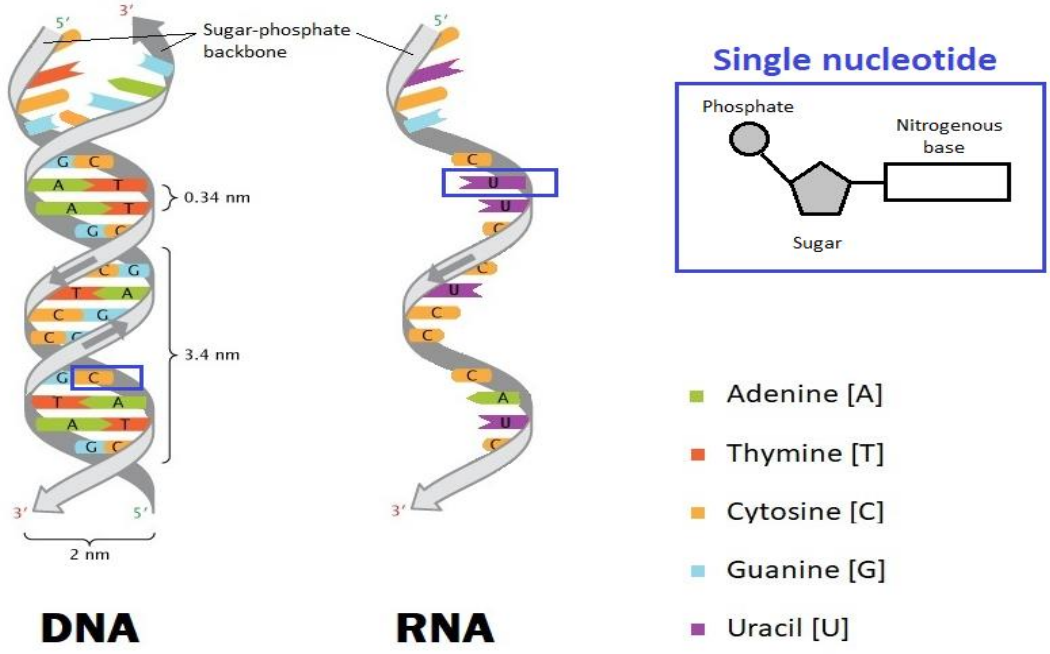


Figure 2. Structures of DNA, RNA and nucleotides (modified from reference 6).

2.2 RNA transcription, protein translation and biosynthesis of metabolites

The structure of ribonucleic acid (RNA, Figure 2) is similar to that of DNA but differs in that RNA is a single-stranded molecule, its sugar-phosphate backbone contains a ribose sugar instead of the deoxyribose, and it contains uracil [U] instead of thymine [T].[7] Pursuant to the 'central dogma of molecular biology'[8] (Figure 3) DNA directs the formation of RNA, which in turn directs the synthesis of proteins. Those proteins in turn carry out the tasks and processes within organisms that we understand to be life. Many proteins are involved in the production of metabolites.

Transcription. The process by which DNA is copied into RNA is called 'transcription' and is carried out by an enzyme called RNA polymerase. During transcription, a DNA sequence is read by RNA polymerase

which produces a complementary RNA strand (Figure 3). The amino acid sequence in a protein is correlated with the sequence of nitrogenous bases in RNA and each of the 20 natural amino acids is specified by a three-base sequence of the RNA called a 'codon'. For instance, the three-base codon (CCC) encodes the amino acid proline while the codon AAA produces the amino acid lysine.[9,10] Scientists deciphered the sequences of the 64 codons in nature as shown in Table 1. Transcription is the first step in gene expression and in eukaryotic cells (cells with a nucleus), it may be followed by splicing in which introns (non-coding regions) are removed and exons (coding regions) are joined together. In prokaryotic cells (cells without a membrane-bound nucleus), splicing does not occur. One field of study to determine which genes are expressed under given conditions in an organism is called 'transcriptomics' (Figure 1).

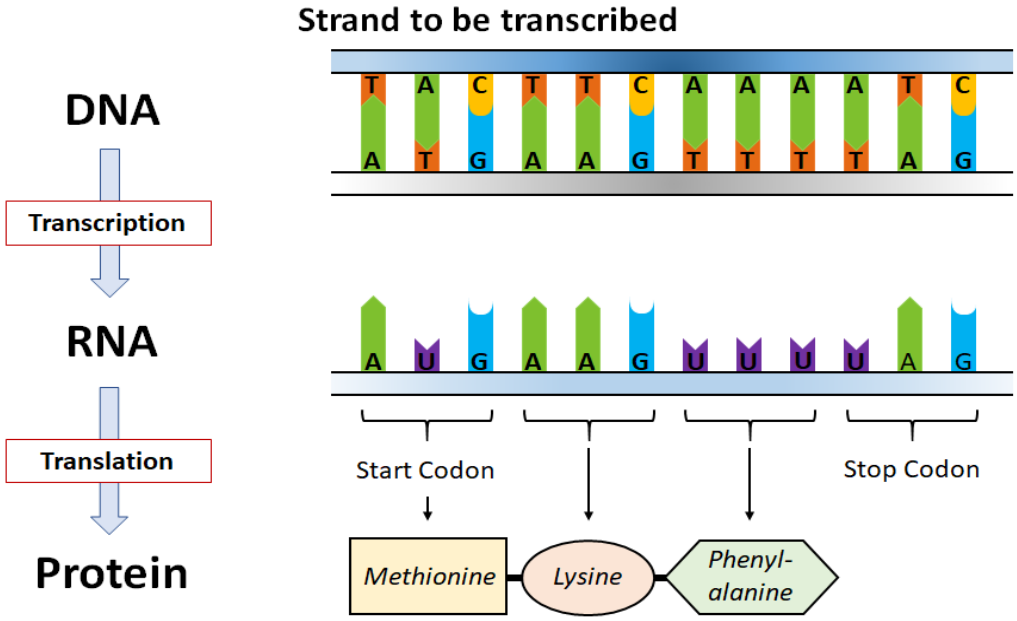


Figure 3. The 'Central Dogma of Molecular Biology' focuses on the processes of transcription and translation.

Translation. The process by which the base sequence information in RNA is converted into an amino acid sequence in proteins is called translation. This process takes place on the ribosomes which are large complexes of RNA molecules and proteins. Although the codon 'AUG' encodes the amino acid methionine, it also activates the ribosome to start the process of making a protein and is thus known as 'start codon'. Similarly, there are 'stop codons' which signal the termination of translation into proteins. Many amino acids can be encoded by different codons and because of such redundancy, the genetic code is described as degenerate. Translation of a DNA sequence to a protein sequence can be carried out automatically using the standard codon triplets, whereas the reverse process is not easily possible, thereby making it difficult or impossible to trace it back to the original DNA sequence.

As said, different codons frequently code for the same amino acid, so different DNA sequences could lead to the same protein. Different taxonomic groups have a 'preference' for the use of a particular codon for a specific amino acid. Researchers usually choose from the different options for codons the one that is preferred by the organism they are studying (or from which a desired substance was obtained) when they want to express a specific amino acid and this process called 'codon optimization'. Finally, it should be mentioned that some organisms do not use the standard triplet codons, for instance

Tetrahymena encodes the amino acid glutamine as TAA which in the 'universal code' is assigned as a 'stop' codon.[11]

Biosynthesis. The process by which proteins give rise to metabolites is called biosynthesis. Biosynthetic enzymes are protein catalysts directing the synthesis of 'primary metabolites' which are directly involved in the growth, development, and reproduction of all organisms (e.g. carbohydrates, proteins, lipids and nucleic acids), as well as 'secondary metabolites' or 'natural products' which are made by biosynthetic pathways specific to certain species (e.g. venoms, toxins and antibacterial agents). Metabolites can be simple (e.g. the commonly consumed alcohol ethanol, a primary metabolite) or complex (e.g. the plant derived anti-cancer agent paclitaxel, a secondary metabolite/natural product). In many cases, especially in microorganisms, genes encoding the biosynthetic enzymes of specific metabolites are clustered together to ensure the optimum production of a metabolite. Complementary to this process is biocatalysis/biocatabolism which catalyze the breakdown of macromolecules and small molecules and thus generate chemical diversity by a different mechanism from biosynthesis.

Table 1. The genetic code. Amino acids may have more than one triplet codon. Some codons are assigned 'start' and 'stop' functions which start/stop the process of translation by the ribosome to generate a protein from an RNA sequence.

		Second base									
		U		C		A		G			
First base	U	UUU	Phenylalanine	UCU	Serine	UAU	Tyrosine	UGU	Cysteine	U	Third base
		UUC		UCC		UAC		UGC		C	
		UUA	Leucine	UCA		UAA	Stop	UGA	Stop	A	
		UUG		UCG		UAG		UGG	Tryptophan	G	
	C	CUU	Leucine	CCU	Proline	CAU	Histidine	CGU	Arginine	U	
		CUC		CCC		CAC		CGC		C	
		CUA		CCA		CAA	Glutamine	CGA		A	
		CUG		CCG		CAG		CGG		G	
	A	AUU	Isoleucine	ACU	Threonine	AAU	Asparagine	AGU	Serine	U	
		AUC		ACC		AAC		AGC		C	
		AUA		ACA		AAA	Lysine	AGA	Arginine	A	
		AUG	Methionine / Start	ACG		AAG		AGG		G	
	G	GUU	Valine	GCU	Alanine	GAU	Aspartic acid	GGU	Glycine	U	
		GUC		GCC		GAC		GGC		C	
		GUA		GCA		GAA	Glutamic acid	GGA		A	
		GUG		GCG		GAG		GGG		G	

2.3 Natural and synthetic modifications to DNA, RNA and proteins

DNA, RNA and proteins are frequently modified in nature to allow them to carry out a range of different functions. In nature changes can be made using mutation, recombination or mutagenesis. Other significant types of possible natural and synthetic modifications are summarized in Figure 4. The first is that it is possible to substitute a DNA sequence with a new one but maintaining the same function. Secondly, the range of nucleotides and amino acids in DNA and proteins can be expanded to include non-natural ones and it is also possible for the subunits to be modified extensively so that they are no longer regarded as nucleotides or amino acids. Finally, modifications to DNA and proteins may be made after they have been formed, for example DNA that has been subjected to epigenetic modification. We will now consider each of these possibilities in turn.

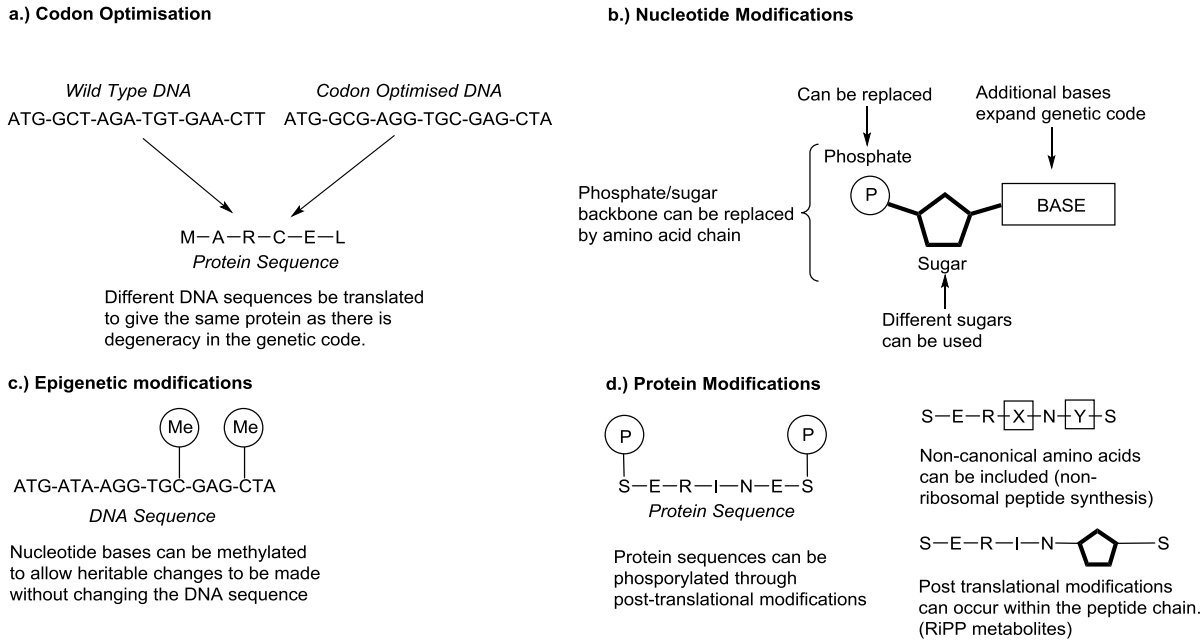


Figure 4. Different types of modification that can be made to DNA and protein sequences.

2.3.1 DNA sequence modifications

As discussed in Section 3.2, DNA sequences can be codon optimized to allow efficient protein expression in another organism (Figure 4 a.). A codon optimized DNA sequence can express a native protein despite differing from the 'wild-type' DNA on which it is based. The codon optimized DNA sequence is now non-native and tracing it back to the originating sequence will be complex, if not impossible as for each amino acid there are multiple different codons. However, the protein sequence derived from these will still be traceable as it remains the same, no matter which codons were used in its translation. In addition, DNA sequences can be designed and synthesized to generate wholly new proteins with known or novel functions.

2.3.2 Nucleotide modifications

A nucleotide is composed of 3 elements, a phosphate, a sugar and a base each of which can be modified (Figure 4 b.). There are several ways in which this has been achieved, for instance replacing the phosphate or sugar units with modified versions.[12] More radical is the complete replacement of the phosphate-sugar backbone with one made from an amino acid chain resulting in 'peptide nucleic acids' (PNA). These are no longer nucleotides but form a double helix and can carry and transfer the genetic code, like DNA.[13] It has also recently been possible to expand the number of bases from two pairs (G/C &

A/T) to four pairs of complementary bases using synthetic nucleotides, and to develop a test-tube system that allows these to be transcribed to RNA, thus expanding the density of information that DNA can encode.[14] The recent World Intellectual Property Organization Standard ST.26 'Recommended standard for the presentation of nucleotide and amino acid sequence listings using xml' uses a similar definition for nucleotides and their modifications.[6]

2.3.3 Epigenetic modifications

This allows the heritable changes to be made to DNA without altering the original sequence (Figure 4 c). There are multiple ways in which DNA can be prevented from being transcribed to RNA, such as via repressor proteins that attach themselves to specific regions of DNA or the modification of histones, around which DNA winds when packed in chromosomes, thus preventing the unwinding and expression of a gene. Most relevant here is the methylation of DNA, and if this happens in a gene promoter, it suppresses gene transcription. Epigenetic modifications can be context specific and epigenetic changes can have varying degrees of modification. In bacteria, epigenetic modifications may be transient and are not necessarily heritable. Methylation data is not straightforward to obtain, and controls aspects of gene expression and hence phenotype. As science advances it may be possible in the future to rapidly predict epigenetic methylation patterns. The study of all the epigenetic modifications in an organism is termed 'epigenomics' (Figure 1).

2.3.4 Protein modifications

Proteins are often modified through metabolic processes, with the simplest modification being the addition of phosphate ('phosphorylation') which acts as an energy source for a protein to enable it to function (although there is a list of others such as acetylation, glycosylation and several more) (Figure 4 d). Such 'post-translational' modifications are common and often happen at predictable sites – such as a particular amino acid, but at other times are hard to predict. These modifications often confer new functional properties to the peptide or protein such as adherence.

Smaller proteins – up to ~50 amino acids long are termed 'peptides' and these can contain heavily modified natural amino acids, with more than 200 reported compared to the 20 natural ('canonical') amino acids. There are two well understood processes by which such peptides can be formed, the so called RiPPs (ribosomally produced and post-translationally modified peptides) and NRPS (non-ribosomal peptide synthesis). The RiPPs rely on a series of enzymes that modify amino acid sequences that are produced by translation of RNA, whereas the NRPS are generated by a complex of enzymes that generate non-canonical amino acids and combine them into metabolites.

2.4 DNA sequencing technologies

Sanger sequencing, introduced in the 1970s, allowed stretches of DNA (100-1000 base pairs) to be accurately sequenced.[15] Longer strands of DNA are subdivided into fragments that are sequenced separately and these sequences are then assembled to give the overall sequence. Methods of DNA sequencing that involve randomly breaking up DNA into many small pieces and then reassembling the sequence by looking for regions of overlap using bioinformatics are sometimes referred to as 'shotgun sequencing'.[16]

[6] https://www.wipo.int/export/sites/www/standards/en/pdf/03-26-01.pdf

Sanger sequencing was successfully used in 1982 to sequence the genome of the bacterial virus, bacteriophage λ (48,502 base pairs).[17] The first commercial sequencer was introduced in 1987 which enabled the Human Genome Project to be launched in 1990. This project catalyzed the development of cheaper, high throughput and more accurate platforms known as the next generation sequencers. These new platforms have increased the speed of sequencing remarkably. They differ in read length, output data, quality and cost and Table 2 shows a comparison between the most used techniques today.

The error rate of DNA sequencing (Table 2) may mean that it is difficult to distinguish whether a change in a DNA sequence is due to an error in sequencing or a consequence of natural variation. In some cases, errors can be corrected using bioinformatic tools, and if carried out, the remaining differences are likely due to natural variation. However, errors in sequencing can also be reduced by ensuring adequate or increased coverage or depth[7].[18]

[7] Depth is the average number of times that a particular nucleotide location is represented in a collection of random raw sequences

Table 2. Characteristics, strengths and weaknesses of commonly used sequencing platforms[19]

***Platform* / Instrument**	**Throughput range (Gb)[a]**	**Read length (bp)**	**Strength**	**Weakness**
Sanger sequencing				
ABI 3500/3730	0.0003	Up to 1kb	Read accuracy and length	Cost and throughput
Illumina				
MiniSeq	1.7–7.5	1×75 to 2×150	Low initial investment	Run and read length
MiSeq	0.3–15	1×36 to 2×300	Read length, scalability	Run length
NextSeq	10–120	1×75 to 2×150	Throughput	Run and read length
HiSeq (2500)	10–1000	1×50 to 2×250	Read accuracy, throughput, low per sample cost	High initial investment, run length
NovaSeq 5000/6000	2000–6000	2×50 to 2×150	Read accuracy, throughput Low per sample cost	High initial investment, run and read length
Ion Torrent				
PGM	0.08–2	Up to 400	Read length, speed	Throughput, homopolymers[b]
S5	0.6–15	Up to 400	Read length, speed, scalability	Homopolymers[b]
Proton	10–15	Up to 200	Speed, throughput	Homopolymers[b]
Pacific BioSciences				
PacBio RSII	0.5–1[c]	Up to 60 kb (Average 10 kb, N50 20 kb)	Read length, speed	High error rate and initial investment, low throughput
Sequel	5–10[c]	Up to 60 kb (Average 10 kb, N50 20 k)	Read length, speed	High error rate
Oxford Nanopore				
MInION	0.1–1	Up to 100 kb	Read length, portability	High error rate, run length, low throughput

[a] The throughput ranges are determined by available kits and run modes on a per run basis. As an example of a 15 GB throughput, thirty-five 5 MB genomes can be sequenced to a minimum coverage of 40x on the Illumina MiSeq using the v3 600 cycle chemistry.

[b] Results in increased error rate (increased proportion of reads containing errors among all reads) which in turn results in false positive variant calling.

[c] Per one SMRTcell.

The new advances in sequencing technologies were also associated with a sharp decrease in the cost of sequencing. This can be seen clearly in the 'Cost per Genome' graph generated by the National Human Genome Research Institute (NHGRI) (https://www.genome.gov/) which has tracked the cost of genome sequencing at the sequencing centers it funds since 2001 (Figure 5). In this graph, two parameters were considered; 1) the size of the genome was assumed to be 3 billion base pairs (i.e. the size of the human genome) and 2) the required "sequence coverage" which is the number of reads that include a given base to overcome errors in the assembly of the genome. The latter differs among sequencing platforms depending on the average sequence read length for each platform.

This lowering of costs has made access to sequencers possible to many researchers, either in their own lab, or via larger-scale sequencing facilities. This has led to a large increase in the amount of sequence data available, leading to an increased necessity for interpreting this data using bioinformatics. The latter is an interdisciplinary field which uses computer programming to analyze biological data. Common uses including search for specific sequences/genes or alignment of homologous sequences to identify mutations and/or predict gene function. Bioinformatics is also used for comparative genomic studies in which the genomic features of different organisms are compared in order to trace the evolutionary processes responsible for the divergence of the genomes.

Gene sequences are often analyzed by reference to databases in which function has been ascribed to genes through laboratory-based research or bioinformatic predictions. Researchers can then compare sequences from different organisms and predict functions for genes which may be utilized commercially. The analysis of sequencing data is known as annotation in which researchers use different techniques to identify the locations and functions of genes and other coding regions in the genome.

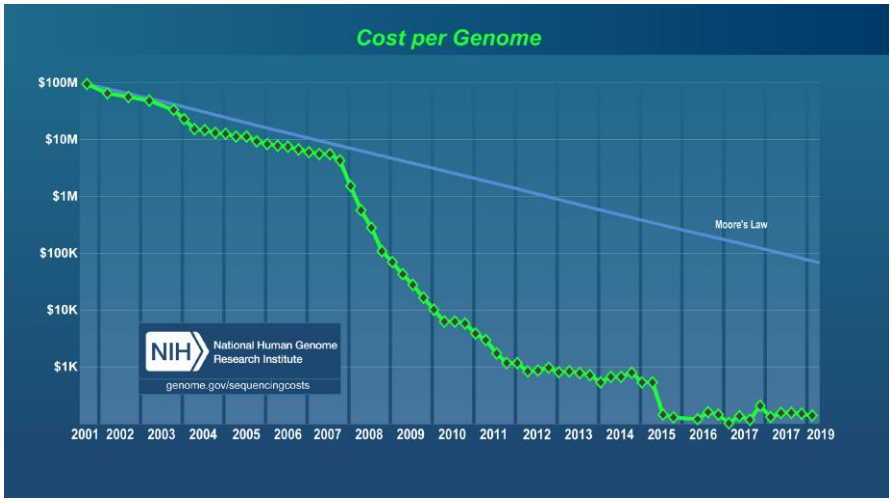


Figure 5. The significant reduction in the cost of genome sequencing over time. Moore's law is an observation and projection of a historical trend. It asserted that the number of transistors on a microchip doubles about every two years, though the cost of computers is halved. It is obvious that the decrease in the cost in sequencing is occurring at a much faster rate than that seen with computers.

(Source: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost).

2.5 DNA Sequencing

Continuous improvements in sequencing technology meant that it was possible to sequence whole genomes of increasingly more complex organisms starting with *Haemophilus influenza Rd* (1995, 1.8 megabase pairs)[20] followed by the fruit fly, *Drosophila melanogaster* (2000, 120 megabase pairs)[21], the first mammal, the mouse (2002, 2,700 megabase pairs)[22] and plant genomes such as rice *Oryza sativa indica* and *Oryza sativa japonica* (2002, 430 megabase pairs)[23,24] However, our knowledge of the genomes of the world's eukaryotic biodiversity is very limited, with in 2017 only 2,534 unique species in the NCBI database having sequenced genomes, representing less than 0.2% of the known species. Of these only 25 species have genomes at the highest level of quality proposed for reference genomes.[25]

Shorter DNA sequences termed 'DNA barcodes' are used to identify a given species through the comparison of nucleotide sequences in its DNA to that of the same regions/genes in other species. When barcoding is used to identify organisms from a sample containing DNA from more than one organism, the term DNA 'metabarcoding' is used.

In some cases, the sample from the organism may contain a heterogeneous mixture of cells as is the case with marine sponges that contain many uncultured microbial symbionts. In these cases, the DNA extracted is a mixture of the genomes derived from the sponge and all the symbionts and is called a 'metagenome'. Analysis of metagenomic data involves a process called binning in which sequencing reads are grouped and assigned to a group of organisms. Metagenomics is also applied to DNA samples directly recovered from environment which is known as environmental DNA or eDNA. eDNA is collected from a variety of environmental samples such as soil, seawater, snow or even air rather than directly sampled from an individual organism. As various organisms interact with the environment, DNA is expelled and accumulates in their surroundings. An example is the DNA fragments left behind by marine organisms in the sea water and this may be derived from living or dead organisms.[26,27] Over time DNA may degrade and give shorter stretches of DNA sequences, making it more difficult to relate these unambiguously to a particular species. An added problem is that DNA extraction from different organisms occurs with different efficiencies and this may skew any metagenome sequencing results. Recent technical improvements of next generation sequencing (NGS) technologies allowed the sequencing of the genome from a single cell thus providing access to previously inaccessible and extremely invaluable information about the function of an individual cell in the context of its microenvironment.[28] Single cell sequencing has been particularly useful to the field of metagenomics.[29]

The DNA sequences published continued to growat unprecedented pace. These data are deposited and maintained in large open access databases and collectively constitute what is understood to comprise DSI. For a better understanding of sequence databases, see the studies commissioned by the CBD Secretariat in parallel to this study which focuses on databases and traceability.

2.6 Genetic engineering and gene editing

Sequences deposited in databases are not limited to the information obtained from the different genome/metagenome sequencing projects. The possibility of engineering the genomes of organisms led to the inclusion of non-native sequences in these databases. The advent of genetic engineering in the 1970s allowed the transfer of genes within and across species boundaries and to introduce mutations in the DNA sequences to produce organisms with improved useful characteristics e.g. crops that tolerate herbicides or resist pests. The resulting entities are widely known as genetically modified organisms (GMOs) or Living Modified Organisms (LMOs). Genes can be readily copied using the polymerase chain

reaction, and editing can be achieved using techniques such as site-directed mutagenesis which modifies a single base in a sequence or by cutting and splicing larger DNA sequences using editing enzymes.

A recently discovered form of gene editing is provided by CRISPR (clustered regularly interspaced short palindromic repeats); a family of DNA sequences found within the genomes of bacteria and which represent part of the bacterial defense system against invading viruses.[30] The most commonly used variant is CRISPR/Cas9 which involves two critical components. The first is the "seeker" – an RNA encoded in the bacterial genome that matches and complements the DNA of the viruses and thus will be able to recognize and bind to the viruses' DNA during the attack. The second element is the "hitman". Once the viral DNA is recognized as foreign, a bacterial nuclease named Cas9 is deployed to cut the DNA of the virus. This system was found to be programmable and by substituting the recognition element, the system can be redirected to cut other genes and genomes at specific sites. The system has been further manipulated to make edits in the genome at the cut site. This was based on knowledge of how the gene repairs itself after being cut. Typically, a cut-open gene tries to recover any lost information from another copy of the gene in the cell. If the cell is given a DNA fragment that has a slightly different sequence from the gene, there is a high probability that the information written on this fragment is copied permanently into the genome.[31]

2.7 Synthetic biology

Synthetic biology[8] is 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. For example, synthetic biology can transform a biological cell into an industrial biofactory[32] using complex biological systems or circuits built from standard interchangeable DNA parts that have defined functions such as regulating transcription, regulating translation, binding small molecules, coding proteins etc. The BioBricks foundation (https://biobricks.org) maintains a registry of standard physical parts that can be freely used by synthetic biology researchers, who co-opt these parts and engineer them for use in applications outside of their natural settings.[32] In the near future this is likely to move from physical parts to DNA sequences of standard parts that can be made using DNA synthesis. Several enabling tools and technologies for synthetic biology have been identified including genomic databases, public and private registries of biological parts, methods for physical assembly of DNA sequences, commercial services for DNA synthesis and sequencing, and advances in bioinformatics.[16] DNA synthesis can readily be used to generate stretches of up to 5,000 base pairs; longer ones can be created by splicing together shorter sections using gene splicing techniques.

Technologies have been developed to expand the genetic code and to allow the incorporation of unnatural amino acids into proteins. Genetic code expansion offers the possibility to directly encode these modifications and to produce a modified protein.[33] One strategy called 'codon assignment' involves using genetic engineering to reallocate one or more of the specific redundant natural codons

[8] Decision XIII/17, paragraph 4 issued pursuant to COP-13 acknowledged the operational definition of "synthetic biology" by the Ad Hoc Technical Expert Group 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", and considers it useful as a starting point for the purpose of facilitating scientific and technical deliberations under the Convention and its Protocols".

(Table 1) to encode an unnatural amino acid. The resultants are called genomically recoded organisms (GRO).[34] Another strategy involves using engineering to allow the ribosome to incorporate unnatural amino acids into protein in response to a four base 'quadruplet' codon.[35] It should be noted that the term GRO could be considered under the much broader term LMO.

2.8 Techniques and databases used to study RNA, proteins and metabolites

An RNA transcript is an indication of which genes are active and which are dormant at any given time or under any given set of conditions in an organism and is studied using 'transcriptomics'. Proteins and metabolites are downstream products of translation and biosynthesis respectively (Figure 1). They fulfill important roles in an organism's metabolism and can be studied using a huge variety of techniques, too many to cover in detail in this report. These techniques are grouped under 'proteomics' and 'metabolomics' below, each of which could fill an entire textbook. Below, we have attempted to give a brief overview of aspects of these three 'omics' technologies relevant to this report.

2.8.1 Transcriptomics

The transcriptome of an organism is a measure of which genes are expressed under any given set of conditions. The transcriptome is highly context-specific and influenced by many variables such as sample processing and analytical techniques. Similar to DNA sequencing, high throughput RNA sequencing can be used to determine the total population of RNAs in a sample. Alternatively, microarrays (DNA/RNA chips, biochips), which are microscope slides printed with thousands of tiny spots in defined positions, with each spot containing a unique, known DNA sequence, can be used. These oligonucleotides act as probes to detect thousands of different transcripts simultaneously, relying on quantitative fluorescence from labelled cDNA synthesized from the sample RNA. The output from these techniques is therefore RNA sequences or an indication of which genes are transcriptionally activated. Microarray probing of samples will only give results for what is probed, providing an inference (indirect measure) of which genes are transcriptionally active.

2.8.2 Proteomics

The proteome of an organism is the totality of all proteins that are produced or modified by an organism. In principle, protein sequences produced by an organism can be predicted from the genome, but the proteome of an organism changes under different growth conditions and stresses, amongst other factors. It is therefore important to measure the actual proteome in an organism using experimental techniques to enable a full understanding of the organism's metabolism and expression levels. While there are many techniques available to achieve this, including antibody-based techniques and protein microarrays, most relevant here is the use of mass spectrometric techniques. Mass spectrometry enables the rapid sequencing of single proteins and peptides by determining the mass of the intact protein/peptide and its fragments. These masses can be searched in online databases (e.g. https://www.uniprot.org/) to determine the sequence of amino acids in a protein, and to compare its similarity to other proteins in the database. Protein masses in the database can be obtained by experiment, or by calculation using the masses of each individual amino acid present in a protein/fragment sequence. The large number of proteins present in a proteome means that some sort of separation technique is necessary prior to mass spectrometry.

Some definitions of proteomics, sometimes termed 'structural proteomics' also include the 3-dimensional structures of the individual proteins. The structure of a protein is determined by its amino acid sequence alone,[36] without the need for additional genetic information. A protein may adopt a

'native' active correctly folded structure and alternative inactive 'misfolded' structures. In principle, therefore, it is possible to predict the structure of a protein given only its amino acid sequence, although it is currently still very difficult to do this reliably.[37] For this reason most protein molecular structures are determined using x-ray crystallography, acquisition of spectroscopic data or the use of cryo-electron microscopy, all of which rely on complex and expensive infrastructure and require considerable computational power. Structural data on DNA, RNA and proteins can be found in freely accessible databases such as the protein databank (https://www.rcsb.org/) giving atom coordinates for these structures and associated metadata and linking out to papers describing their function.

2.8.3 Metabolomics

This is the study of the full complement of small molecule metabolites produced by an organism's metabolism under a certain set of conditions. The metabolome contains a large range of different types of molecule with varying characteristics produced at widely different concentrations. Profiling the metabolome is therefore very different from DNA/RNA sequencing (section 3.4) and proteomics and relies on using high resolution separation techniques to separate each metabolite before measuring its mass and fragments using mass spectrometry. The measured data can be analyzed using a database such as METLIN (https://metlin.scripps.edu) which uses this mass spectrometric data to identify a proportion of the metabolites present (many will be unknown), giving degree of certainty. Statistical techniques are important to study how the metabolome changes when conditions change. For instance, it can be used to measure the influence of a toxin or the effect of a gene modification on a metabolome. If metabolites not present in the database are encountered in metabolomic analysis, it will be more complicated to identify these. If this occurs, the metabolite will need to be obtained in its pure form and its chemical structure defined using spectroscopic or other techniques.[38]

2.8.4 Databases

Once generated, there are many databases in which data derived from genetic resources are deposited. These data include: selected metadata of the genetic resource (e.g. the taxonomy of the organism and its geographical origin); the nucleotide sequence data in DNA and RNA which is (genomic/transcriptomic data); amino acid sequences of proteins (proteomic data) and is complemented by structural data of different proteins as identified by x-ray crystallography, cryo-electron microscopy or nuclear magnetic resonance; the data on metabolites isolated and identified from any organism using different spectroscopic techniques and mass spectrometry (metabolomic data); and the epigenomic data which includes for example the pattern of the DNA methylation or histone acetylation.

3. SECTORS THAT UTILIZE DSI AND TECHNOLOGIES/TECHNIQUES ENABLED BY DSI

3.1 Introduction

The 2018 Laird and Wynberg study[1] together with the synthesis of views[9] and accompanying case studies prepared by the Secretariat to the CBD addressed the potential implications of DSI on the objectives of

[9] Synthesis of views and information on the potential implications of the use of digital sequence information on genetic resources for the three objectives of the Convention and the objective of the Nagoya Protocol (CBD/"DSI"/AHTEG/2018/1/2); Available at https://www.cbd.int/doc/c/49c9/06a7/0127fe7bc6f3bc5a8073a286/dsi-ahteg-2018-01-02-en.pdf

the CBD[10]. These provided a wide range of e examples of different contexts and purposes for which DSI can be used. Whereas Section 3 of this study complements these efforts by providing greater technical context concerning the generation and use of DSI, this section provides illustrative insight regarding how DSI is used in specific sectors in the life-sciences which have been (and continue to be) transformed or enabled by this relatively recent revolution.

As considered in Section 3.8, technologies enabled by DSI are becoming ubiquitous in life-science related research and industry, particularly 'omic' technologies which are primarily aimed at the detection of genes (genomics), mRNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics) in biological and environmental samples. These technologies have a broad range of applications across scientific disciplines and we have chosen the following sectors to highlight: taxonomy and conservation (as indicative of basic research); agriculture and food security; industrial and synthetic biology; healthcare applications and discovery of pharmaceuticals. A comprehensive analysis of each sector is beyond the mandate for this Study, so for each sector we provide a brief overview of the sector accompanied by coverage of key trends and examples highlighting the use of DSI and technologies which are enabled by DSI in that sector. For convenience, Table 3 facilitates a sectorial comparison of the application of the different techniques as outlined in Figure 1 and discussed in Section 3.8 (genomics, transcriptomics, proteomics, metabolomics and epigenomics) in each sector.

3.2 Taxonomy and conservation

3.2.1 Overview of sector

This sector covers one of the main objectives of the CBD: "Conservation of biological diversity". DNA barcodes and longer DNA sequences together with DNA sequence databases allow rapid identification of species or higher order taxa such as genus, family or order. This process can now take days whereas species identification using morphological methodology can take many months and relies on type specimens held in national collection as well as taxonomic expertise, which is becoming rarer. In addition to assisting in the discovery of new species, DNA barcodes have a broad range of applications including biodiversity conservation, observing seasonal effects and effects of climate change on species distributions, as well as correcting mistaken identification and labelling on foods and plant-based medicines. A global effort is underway to catalogue all life on earth,[25] which aims to improve our understanding of ecosystems, evolution, ecosystem services and biological assets. The project is complemented by the CBD's Global Taxonomy Initiative (https://www.cbd.int/gti/) – a cross-cutting effort coordinated by the CBD to ensure that taxonomic information and expertise is available to CBD parties.

3.2.2 Key trends and examples

Monitoring of biodiversity. The greater availability of DNA barcodes for many species could assist biodiversity surveys, enable effective conservation measures to be implemented[39] and new species to be identified.[40]

Evaluating biodiversity response to climatic events.[41] Species richness and assemblage changes in response to climatic events can be measured using metabarcoding. This requires sequencing of

[10] Case studies and examples of the use of digital sequence information in relation to the objectives of the Convention and the Nagoya Protocol (CBD/"DSI"/AHTEG/2018/1/2/ADD1); available at https://www.cbd.int/doc/c/7a1d/3057/f5fa0ecb0734a54aadd82c01/dsi-ahteg-2018-01-02-add1-en.pdf

environmental DNA present in the sample, followed by a comparison to available DNA barcodes for a range of species.

Species identity and labeling. Fish sold can be mislabeled, either through accidental misidentification or willful mislabeling of species. Misidentification can occur when two species are phenotypically similar and can be rectified using DNA barcodes and the construction of a phylogeny to show relatedness.[42] DNA barcodes are used on fish sold to ensure lower value or endangered species are not substituted.[43]

3.3 Agriculture and food security

3.3.1 Overview of sector

The Agrifood sector can be split into two parts. 'Agritech' refers to technologies that target farmers whereas 'Foodtech', targets manufacturers, retailers, restaurants and consumers. Jointly, the two have enough reach to impact every part of the production line, from farm to fork. Agriculture uses a broad range of 'omic' techniques, principally to modify crops, create new varieties, and manage agricultural practices. Genetic modification of crops and livestock can give rise to unique traits such as insect resistance/drought tolerance, for example. Other methods of optimizing productivity can be developed using techniques such as marker-assisted selection.

3.3.2 Key trends and examples

Selective breeding. Most breeding products that are currently on the market have not been developed by the use of gene editing, but by classical breeding and require access to physical genetic resources such as plants. Marker-assisted selection can be used to select traits such as pathogen resistance in crops or parasite resistance in livestock. A high-density map of molecular markers for the tomato contains 40 resistance markers which allowed rapid selection of resistant breeds.[44] Complete genetic maps identifying parasite resistance traits in dairy cattle are a first step to breeding cattle resistant to parasites.[45]

Development and characterization of LMOs. In 2014, half of all LMO crops planted were soybeans modified for herbicide tolerance. A bacterial gene incorporated into the soybean plant confers tolerance to the herbicide glyphosate. Thus, producers can chemically control weed species during the growing season. Near-future LMO varieties are being developed with data from transcriptomics, proteomic, epigenomic and metabolomic experiments.[46-50] In addition, gene editing techniques rely on genomic sequences to create minute changes, conferring traits similar to 'traditional' LMOs or for use in rapid or *de novo* domestication.[51-53]

Soil metagenomics. Understanding the soil microbial communities that carry out key ecosystem services may be achieved by metagenomic analysis which identifies the composition and diversity of these communities. A new frontier, 'metaphenomics' looks into the actual functions carried out by viable and active cells under given environmental conditions.[54]

3.4 Industrial biotechnology and synthetic biology

3.4.1 Overview of sector

Industrial biotechnology provides alternative methods to generate industrial products via processes that can be carried out in water, at ambient temperatures, without producing large volumes of waste. Almost 40% of the global market is attributed to bioenergy and a large proportion of the remainder to renewable chemicals, such as solvents and biodegradable plastics. Biotechnology is heavily reliant on genetic resources for the discovery of new products and processes.[55]

Synthetic biology is a novel area of research that is the amalgamation of multiple disciplines such as molecular biology, biotechnology, biophysics and genetic engineering amongst others (see Section 3.7). The global synthetic biology market can be segmented as indicated below with applications across many industries, including pharmaceutical, diagnostics, energy, agriculture, bioplastics and environment[56]:

- by products: Synthetic DNA/genes; Software tools; Chassis/host organisms; Synthetic clones; Synthetic cells
- by technology: Nucleotide synthesis and sequencing; Bioinformatics; Microfluidics; Genetic engineering
- by application: Pharmaceuticals and diagnostics; Chemicals; Biofuels; Bioplastics; Others (Environment, agriculture & aquaculture)

3.4.2 Key trends and examples

Laundry Detergents. Low temperature laundry detergent enzymes (proteins) are developed by analyzing and modifying genes from low temperature adapted microorganisms.[57,58] The three-dimensional structure of the enzyme is used to identify 'hotspots' where amino acid modifications may have the greatest effect. The gene encoding this enzyme can then be modified, resulting in the desired change.

Production of Bioethanol. Related genes from different organisms can be 'shuffled' to produce 'chimeric' enzymes. These can be tested to determine if they have increased productivity, in this case the production of bioethanol.[59] These genes can be reshuffled until enzyme activity is optimized. Shuffled genes that express chimeric enzymes are difficult to trace back to an originating DNA sequence as this is a product of the gene families used and the shuffling process.

Production of Therapeutic and High-value Compounds. Bacterial, fungal and plant systems are now modified to produce therapeutic and high-value compounds through the introduction of multistep biosynthetic pathways.[60-63] For example, a precursor to the antimalarial artemisinin can now be produced using a synthetic biology process.[64] Process development relied on detailed knowledge of the DNA sequence directing the production of artemisinin in the plant, related genes in other organisms and whole genome sequences of alternative hosts, which were engineered for this purpose. Codon optimization was used extensively.

3.5 Healthcare applications and discovery of pharmaceuticals

3.5.1 Overview of sector

Genetic resources are commonly used in the discovery of small molecule pharmaceuticals, and several of these can be found in the list of most-prescribed pharmaceuticals[65], some of which are based on natural product chemicals. Estimates indicate that 20-25% of this market is derived from genetic resources[56] with nearly 2 out of 3 antibacterial agents deriving from genetic resources.[66] Of other importance in healthcare is also the prevention of disease, such as food-borne illnesses and the early diagnosis so that appropriate treatment can be provided.

3.5.2 Key trends and examples

The design of diagnostic tests for infectious disease agents. Design involves analysis of many sequences to identify highly conserved target regions within the pathogen genome that have no homology to other DNA or RNA sequences in the test sample.[68] These can then be used as markers for presence of the pathogen. For example, diagnosis for the Ebola virus could take as long as 3-10 days but detecting the pathogen at 1 day reduces viral infection to almost 0%. A recent study[69] employed a CRISPR-associated

RNA-guided RNA editing enzyme to detect the RNA genome of the Ebola virus in blood samples in under 5 minutes.

Detection of pathogens in contaminated food for disease prevention. Rapid detection of food-borne pathogens ensures food safety. The National Center for biotechnology Information (NCBI) hosts a 'Pathogen Detection' website that shares data on gene sequences for these pathogens.[70] It quickly clusters and identifies related sequences to uncover potential food contamination sources. Genetic information is used for the initial identification of the pathogen causing the disease, but also for the identification of clusters or outbreaks. Identification at an early stage facilitates the implementation of preventive measures, thereby reducing the public health impact. Sequencing information can also be used to predict the resistance to antibiotics, further guiding the treatment that should be followed, and allowing it to be more effective in the presence of resistance genes.

Discovery of new drugs. Bacteria produce a range of important pharmaceuticals.[66] Comparative genomic analysis of microbes can uncover new pharmaceutical compounds. For example, whole genome analysis of the bacterium *Staphylococcus lugdunensis* indicated the bacterium contained a biosynthetic pathway for the previously unknown metabolite, lugdunin, which is effective against antibiotic resistant infections in a mouse model.[71]

3.6 Extent of reliance on DSI and technologies/techniques enabled by DSI

It is evident from the illustrative coverage of each sector, and also the selected examples of uses of DSI-related technologies in different sectors in Table 3, that all sectors considered in this study use different types of information that potentially constitutes DSI and technologies/techniques enabled by DSI. In particular, genomic data appears to be highly utilized in all sectors. Similar trends are expected for other sectors in the life-sciences. Thus, these sectors can be considered while discussing the scope and concept of DSI and assessing the implications of including or excluding particular types of information associated with the underlying genetic resource.

Table 3. Selected examples of the use of DSI-related technologies in different sectors. Relevance of each 'omic' technology is shown in the column headed 'use' and indicated as High (H), Medium (M) or Low (L)

	Taxonomy & Conservation		**Agriculture & Food Security**		**Industrial & Synthetic Biology**		**Healthcare & Pharmaceuticals**	
	Use	**Comment**	**Use**	**Comment**	**Use**	**Comment**	**Use**	**Comment**
Genomics	**H**	DNA barcode database incomplete for many branches of life and relies on established taxonomy and systematics with variable coverage for different divisions of life. Taxonomic organization subject to change.	**H**	Reference genomes and identifying natural variation and trait loci. Metagenomic analysis of soil micro-organisms to understand crop health.	**H**	Accurate/reliable annotations to assign functions to genes. Unknown genes require additional lab work. Proteins with same function may have different DNA sequences.	**H**	Analysis of disease targets and pathogen DNA and RNA sequences to develop treatments and diagnostics.
Epigenomics	**L**		**M**	Understanding heredity in livestock.	**M**	Understanding of phenotypic changes in LMO produced using synthetic biology.	**M**	Microarray data are now the main source for identifying new therapeutic targets with the current shift to personalized medicine.
Transcript-omics	**M**	Identification of metabolically active species in environmental samples.	**M**	Understanding function of different micro-organisms soil microbiome in maintaining crop health.	**M**	Determination of genes which are being transcribed allows up or down regulation of pathways to increase production.	**M**	RNA silencing and gene therapy rely on these data.
Proteomics	**L**		**M**	Determine if LMOs are expressing desired proteins.	**H**	Determination of which proteins are being expressed. Used to identify biosynthetic gene clusters directing the production of metabolites.	**M**	Understanding of proteins involved in production of potential natural product pharmaceuticals.
Metabol-omics	**M**	Profiling of plant metabolites to identify correct phenotype (chemotaxonomy).	**M**	Determine if LMOs are producing desired metabolites.	**M**	Identification and quantification of small molecules being produced, used to redirect metabolic flux to increase production of these small molecules.	**M**	Analysis of metabolites produced by organisms studied for potential natural product pharmaceuticals.
Other					**H**	Codon optimization to achieve expression of modified gene constructs in alternative hosts. Gene editing tools. Molecular structures of proteins.	**M**	Develop of pharmaceuticals and disease diagnostics using gene editing tools.

4. DSI: SCOPE AND TERMINOLOGY

4.1 Introduction

During the 2017-2018 inter-sessional period, parties to the CBD and Nagoya Protocol undertook a number of steps to attempt to clarify the concept of DSI.[11] This process did not yield consensus on the appropriateness of the term 'DSI' nor what it refers to, whether it is limited to DNA and RNA sequences or whether it also covers the amino acid sequences of proteins and the metabolites produced by biosynthetic enzymes, among other types of information.[72] These challenges are not unique to CBD and its Nagoya Protocol as evidenced by related discussions underway in various other UN processes such as the International Treaty on Plant Genetic Resources in Food and Agriculture (ITPGRFA), the Pandemic Influenza Preparedness Framework (PIP) and the process[12] concerning the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction (BBNJ). Various definitions for 'DSI' and equivalent terminology have been published or proposed by organizations, trade bodies and learned societies involved in the discussions across these domains.

In 2018 the Ad Hoc Technical Expert Group (AHTEG) on Digital Sequence Information on Genetic Resources established under CBD and its Nagoya Protocol compiled a broad list of subject matter that may potentially comprise DSI.[13] –This list is useful as it is the most comprehensive breakdown that has emerged from CBD's efforts to date relevant to the utilization of genetic resources. Accordingly, we reproduce the AHTEG list here for convenience and use it as the starting point for our observations in this study:

(a) "The nucleic acid sequence reads and the associated data

(b) Information on the sequence assembly, its annotation and genetic mapping. This information may describe whole genomes, individual genes or fragments thereof, barcodes, organelle genomes or single nucleotide polymorphisms.

(c) Information on gene expression

(d) Data on macromolecules and cellular metabolites

(e) Information on ecological relationships, and abiotic factors of the environment

(f) Function, such as behavioral data

(g) Structure, including morphological data and phenotype

(h) Information related to taxonomy

---

[11] Parties and relevant stakeholders were invited to submit their views on potential implications of the use of digital sequence information on genetic resources for the three objectives of the Convention and a fact-finding and scoping study addressing similar issues was commissioned (Laird and Wynberg study). A synthesis of the submissions received, including case studies and examples of the use of "DSI", and the Laird and Wynberg study were considered by an Ad Hoc Technical Expert Group on Digital Sequence Information on Genetic Resources whose report and recommendations was subsequently submitted to COP14 and its Subsidiary Body on Scientific, Technical and Technological Advice.

[12] The Intergovernmental Conference on an international legally binding instrument under the United Nations Convention on the Law of the Sea.

[13] Report of the AHTEG on Digital Sequence Information on Genetic Resources is available at https://www.cbd.int/doc/c/f99e/e90a/71f19b77945c76423f1da805/dsi-ahteg-2018-01-04-en.pdf

(i) Modalities of use"

This list indicates the types of information that may be relevant to the utilization of genetic resources, however, some elements were not clearly defined such as 'associated data' under category (a). Some of the categories, in particular (e)-(i) were not considered in detail by the AHTEG or in the views on DSI submitted to the Secretariat of the CBD in the 2017-2018 inter-sessional period. The broad scope of the AHTEG list reflects differences of opinion which exist regarding DSI subject matter and this is reflected/inherent in the different terminology proposed to describe the concept of DSI. Building on these previous efforts and reflecting on the terminology being considered in the various UN processes described above, this study attempts to further clarify the concept of DSI by introducing new logical groupings ('broad', 'intermediate' and 'narrow') which may be better suited than the AHTEG list to facilitate discussions regarding scope and terminology associated with DSI, and by posing certain priority questions/issues which need to be resolved if a suitable terminology and scope are to be found.

We commence by drawing a conceptual distinction between data and information and evaluate their flow from the utilization of a genetic resource (Section 5.2). We use this as a basis to propose the new logical groupings for DSI subject matter which are mapped against the AHTEG list and also against alternative terminology to replace DSI in order to help clarify the subject matter and boundaries of these groupings (Section 5.3). We identify priority questions/issues that need to be addressed in order to clarify the concept of DSI by considering the meaning of the terms 'digital', 'sequence' and 'information', in turn, (Section 5.4) and by considering the effect of modifications to DNA, RNA and protein subunits (Section 5.5).

4.2 Understanding the flow of data and information

It appears that a common challenge faced at the CBD and other UN processes in clarifying the subject matter and terminology associated with DSI or its equivalent terms, is in deciding what counts as data and the circumstances in which data is embedded with value and transformed into information (knowledge). This distinction can be difficult to apply in practice, however, data is essentially a means of communicating and facilitating exchanges about the material world. Data describes inherent characteristics of material artefacts as distinguished from research outputs or other value-adding steps that generate knowledge such that its dissemination constitutes the sharing of information (knowledge, claims, models, theories, communities, and so on) as distinct from the underlying data itself.

In the context of a genetic resource, the question arises as to whether DSI should be confined to representational data (such as a DNA sequence 'GTACCTGA …', methylation patterns, and so on) and if not, to what extent it should include processing activities performed with that data to generate information in whatever format, medium, shape, and so on, by data producers, curators, users, and so on. Conceived this way, a key challenge faced across the various UN processes is to determine whether DSI, howsoever called, is limited to DNA and RNA sequences or whether it also captures the amino acid sequences of proteins and/or information generated by cognitive processes applied to such data.

Given the difficulty in distinguishing data from information, beyond this point we use both terms 'data' and 'information' using the most appropriate term in each circumstance. As an approach to address this challenge it is useful to consider the flow of data/information from a genetic resource onwards to DNA, RNA, protein sequences and metabolites as depicted in Figure 6, which also integrates terminology and subject matter components that may assist in clarifying the concept of DSI. It is evident that at each step the data/information it yields becomes progressively further removed from the original genetic resource.

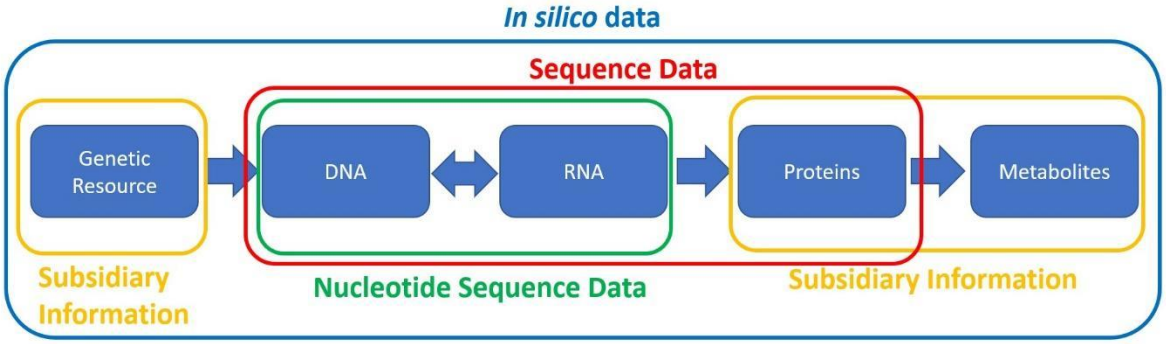


Figure 6. The flow of data/information from genetic resource through DNA, RNA and proteins to metabolites showing the limits/boundaries of some alternative terms used to refer to DSI. Subsidiary information on the genetic resource includes sample metadata, taxonomy, biotic/abiotic environmental factors, traditional knowledge, phenotypic data, ecological relationships and behavioral data amongst others.

4.3 New logical groupings & alternative terminology

To help clarify the concept of DSI we use the flow of data/information from a genetic resource, particularly the degree of biological processing and proximity to the underlying genetic resource, to provide a logical basis to group information that may comprise DSI. This gives rise to four proposed groups, one broad/inclusive group, two intermediate groups and a narrow/defined group, as depicted in Figure 7 and further described below. They are summarized as follows:

- Group 1 - Narrow: concerning DNA and RNA
- Group 2 - Intermediate: concerning (DNA and RNA) + proteins
- Group 3 - Intermediate: concerning (DNA, RNA and proteins) + metabolites
- Group 4 - Broad: concerning (DNA, RNA, protein, metabolites) + traditional knowledge, ecological interactions, etc.

Group 1 has a narrow scope and proximity to the genetic resource and is limited to nucleotide sequence information associated with transcription. Group 2 has an intermediate scope and extends to protein sequences, thus comprising information associated with transcription and translation. Two interpretations for the scope of this group are possible, as discussed below. Group 3 has a wider intermediate scope and extends to metabolites and biochemical pathways, thus comprising information associated with transcription, translation and biosynthesis. Group 4 has the broadest scope and also includes information with the weakest proximity to the underlying genetic resource, extending to behavioral data, information on ecological relationships and traditional knowledge, thus comprising information associated with transcription, translation and biosynthesis, as well as downstream subsidiary information concerning interactions with other organisms and the environment as well as its utilization, among other subsidiary information. Scientifically, groups 1-3 are all based on the molecular structure of macromolecules and small molecules, the information they carry and information associated with their acquisition. Group 4 also includes information that is not related to molecular structure or information associated with their acquisition.

Taking into consideration the proximity of information to the underlying genetic resource and the biological process associated with the generation of the information, is a useful proxy to determine if it is possible to accurately identify or infer the genetic source from which it is derived. For example, in the

case of DNA it may be possible to identify the genetic source, however, certain genes are conserved across wide taxonomic ranges and the sequence may not be traceable to any particular genetic resource but rather to a genus, family or higher taxon. In the case of RNA and protein sequences it is possible to infer the genetic sequence of the source, however, whereas this can be inferred with a high degree of precision/confidence for RNA, the redundancy of the genetic code makes this less precise for proteins (because multiple codons are available to encode an amino acid and so more than one DNA option will be inferred from a protein sequence, see sections 3.2 and 3.3.1). Precision becomes even more challenging with biosynthetic information and inferring the underlying genetic code is not possible from some subsidiary information. Accordingly, proximity has significant implications for traceability to a particular genetic resource and also in identifying the source of information, including whether it has been generated through the utilization of a genetic resource or independently.

Using these proposed groups, we can evaluate the broad list of subject matter potentially comprising DSI as proposed in 2018 by the Ad Hoc Technical Expert Group (AHTEG) on Digital Sequence Information on Genetic Resources, as identified above. We can also use these groups to evaluate a range of terms proposed to replace DSI, including[14]: In silico; Dematerialized Genetic Resources (DGR); Genetic Information (GI); Digital Sequence Data (DSD); Genetic Resource Sequence Data (GRSD); Genetic Sequences (GS); Genetic Sequence Data/Information' (GSD/GSI); Nucleotide Sequence Data (NSD); and Subsidiary Information (SI).[15] These evaluations are shown in Table 4 which is a key reference for the reader to understand the different groups proposed to evaluate the concept of DSI in this study. Please note that in this table additional categories are listed where the original AHTEG report is unclear. In these cases, such as 'associated data' in a.2 of Table 4, which is not defined, we have added a more detailed explanation in the row underneath. Other categories are subdivided to group similar information together. The precise subject matter content, boundaries and definition of these terms are by no means universally agreed, so the evaluation is, of course, subject to a degree of interpretation as some terms may have a narrower scientific or technical meaning and the categories of information corresponding to each other ultimately depends on how the terms are understood or further defined.

Terminology appears to be available to describe DSI with narrow subject matter limited to nucleotide sequences (as proposed in Group 1). These terms could include Genetic Resource Sequence Data (GRSD); Genetic Sequences (GS); Genetic Sequence Data/Information' (GSD/GSI); and Nucleotide Sequence Data (NSD). It is also evident that terminology is available to describe subject matter with broad scope extending beyond transcription, translation and biosynthesis (i.e. as proposed in Group 4). These terms include In silico; Dematerialized Genetic Resources (DGR); Genetic Information (GI).

The terms Digital Sequence Data (DSD), Genetic Resource Sequence Data (GRSD) or Genetic Resource Sequence Data and Information (GRSDI), although previously used in certain contexts to describe Group 1 (narrow), could be used to describe the subject matter of intermediate scope comprising information associated with transcription and translation (as proposed for Group 2) depending on the interpretation

[14] These terms arise from CBD forums, publications, professional bodies and learned societies in the context of the parallel discussions underway in the various UN processes also attempting to clarify the concept of "DSI", howsoever called.

[15] Additionally, during our interviews, further terms were introduced which can be analysed by reference to Table 4 and the discussions above: "Biological sequence information", "Functional sequencing information", "Digital genetic resources and sequence information", "Digital biological code", "Digital sequence information on genetic material" and "Digital biological information". These will not be discussed but could be analysed in the same manner as all the terminology discussed above.

adopted. None of the terms proposed to date appear to adequately capture an intermediate range comprising information associated with transcription, translation and biosynthesis of a genetic resource (i.e. as proposed for Group 3). Overall, the four logical groups proposed in this study provide a nuanced alternative to the 2018 AHTEG list and so may better assist in clarifying the concept and scope of DSI, however, appropriate terminology will need to be evaluated, particularly for the intermediate groups.

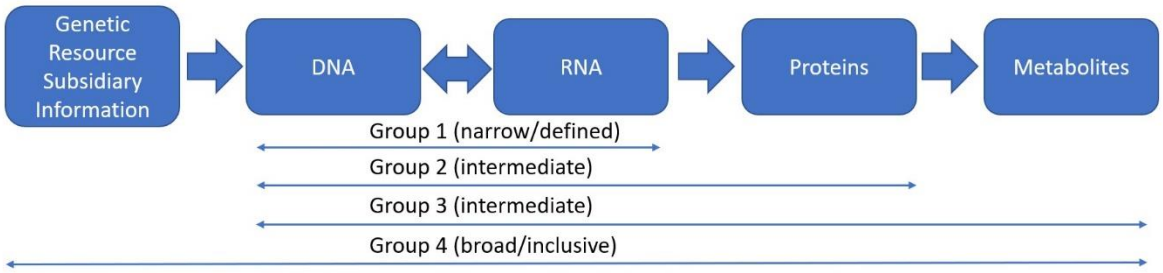


Figure 7. Proposed subject matter groupings for data/information potentially constituting DSI to facilitate discussions concerning DSI scope and terminology. Group 1 only includes data on DNA and RNA sequences, whereas Group 2 also incorporates data/information concerning protein sequences. Group 3 extends to data/information concerning metabolites and Group 4 is the broadest category which extends further downstream beyond metabolites to also include all subsidiary information.

Table 4. Scope of the different current terminologies showing the subject matter groupings as in Figure 7. Some of the AHTEG categories have been subdivided or supplemented with additional subcategories for clarity. + signs in categories were assigned based the definitions available.

		Narrow/Defined (Group 1)					Intermediate (Groups 2 & 3)			Broad/Inclusive (Group 4)			
AHTEG Category	**Component**	**DSD**	**GRSD**	**GS**	**GSD GSI**	**NSD**	**2a**	**2b**	**3**	**In silico**	**DGR**	**GI**	**SI**
a1	Nucleic acid sequence reads	+	+	+	+	+	+	+	+	+	+	+	
a2	Associated data to nucleic acid reads (technical aspects of sequencing experiments: the sequencing libraries, preparation techniques and data files).		+	+	+	+	+	+	+	+	+	+	
b1	Information on the sequence assembly, including structural annotation and genetic mapping. (This information may describe whole genomes, individual genes or fragments thereof, barcodes, organelle genomes or single nucleotide polymorphisms).			+	+	+	+	+	+	+	+	+	
b2	Non-coding nucleic acid sequences		+	?	?	+	+	+	+	+	+	+	
b3	Functional annotation of genes					?		+	+	+	?	+	
c1	Information on gene expression							+	+	+	?	+	+
c2	Epigenetic heritable elements (e.g. methylation patterns).							+	+	+	?	+	+
d1	Amino-acid sequence of proteins produced by gene expression.						+	+	+	+	?	+	+
d2	Molecular structures of proteins.							+	+	+	?	+	+
d3	Data on other macromolecules (not DNA, RNA or proteins) and cellular metabolites. (Molecular structures).								+	+	?	+	+
e	Information on ecological relationships, and abiotic factors of the environment.									+	?	+	+
f	Function, such as behavioral data (this would include environmental influences).									+	?	+	+
g	Structure, including morphological data and phenotype (this would include environmental influences).									+	?	+	+
h	Information related to taxonomy.									+	?	+	+
i	Modalities of use.									+	?	+	+
	Additional undefined elements.									+	?	+	+

Where: DGR = dematerialised genetic resources; GI = genetic information; DSD = digital sequence data; GRSD = genetic resource sequence data; GS = genetic sequence; GSD = genetic sequence data; GSI = genetic sequence information; NSD = nucleotide sequence data; and SI = subsidiary information

4.3.1 Broad scope of subject matter: information associated with biological processing and subsidiary information

Scope

As proposed above, Group 4 is an open-ended category which has the broadest scope and includes subject matter of the weakest or non-existent proximity to the underlying genetic resource. Examples could include behavioral data, information on ecological relationships and traditional knowledge, thus comprising information associated with transcription, translation and biosynthesis, as well as downstream subsidiary information concerning interactions with other genetic resources and the environment as well as its utilization, among other subsidiary information.

Evaluation of Existing Terms

In silico The BBNJ process is considering the term '*in silico*' storage and utilization of data or information.[16] The term is also in use by certain CBD parties. In biology and chemistry this term is used to mean 'performed on computer or via computer simulation', with the reference to silicon, the material from which computer chips are manufactured. It may refer to any data or information held or processed on a computer, all of which fall within AHTEG categories a.-i.

Dematerialised Genetic Resources (DGR)[73] This terminology refers to the informational aspects of genetic resources. It includes the acquisition, digitalization, storage and dissemination of DNA sequences from genetic resources. The separation between the provider of the genetic resource and the eventual user as well as the digital nature of the data prompts the use of the word 'dematerialized'. This information can then be 're-materialized' through gene synthesis and incorporation in living modified organisms (genetic modified organisms). This may only cover the DNA and RNA sequences of these genetic resources, but the word 'dematerialized' may include all types of information relating to the genetic resources from categories a.-i. in the AHTEG study.

Genetic Information (GI)[74] Collective term used to refer to information derived from genetic resources, plant materials and viruses. It was a catch-all term used in discussions around the information under the CBD, the Nagoya Protocol, the ITPGRFA, and the PIP Framework, and encompasses AHTEG categories a.-i.

4.3.2 Intermediate scope: information associated with biological processing involving transcription, translation and biosynthesis

Scope

As proposed above Group 3 has an intermediate scope and extends to protein sequences and metabolites thus comprising information associated with transcription, translation and biosynthesis. Figure 8 shows how the proposed intermediate groupings relate to the scope of the existing terminology (top panel) and also how the scope of proposed group 3 could be interpreted (bottom panel).

Evaluation of Existing Terms

[16] Other terms being considered under this process include 'marine genetic resources in silico', 'digital sequence information', 'genetic sequence data'.

None of the terms proposed to date appear to adequately capture an intermediate range comprising information associated with transcription, translation and biosynthesis of a genetic resource, as proposed by Group 3.

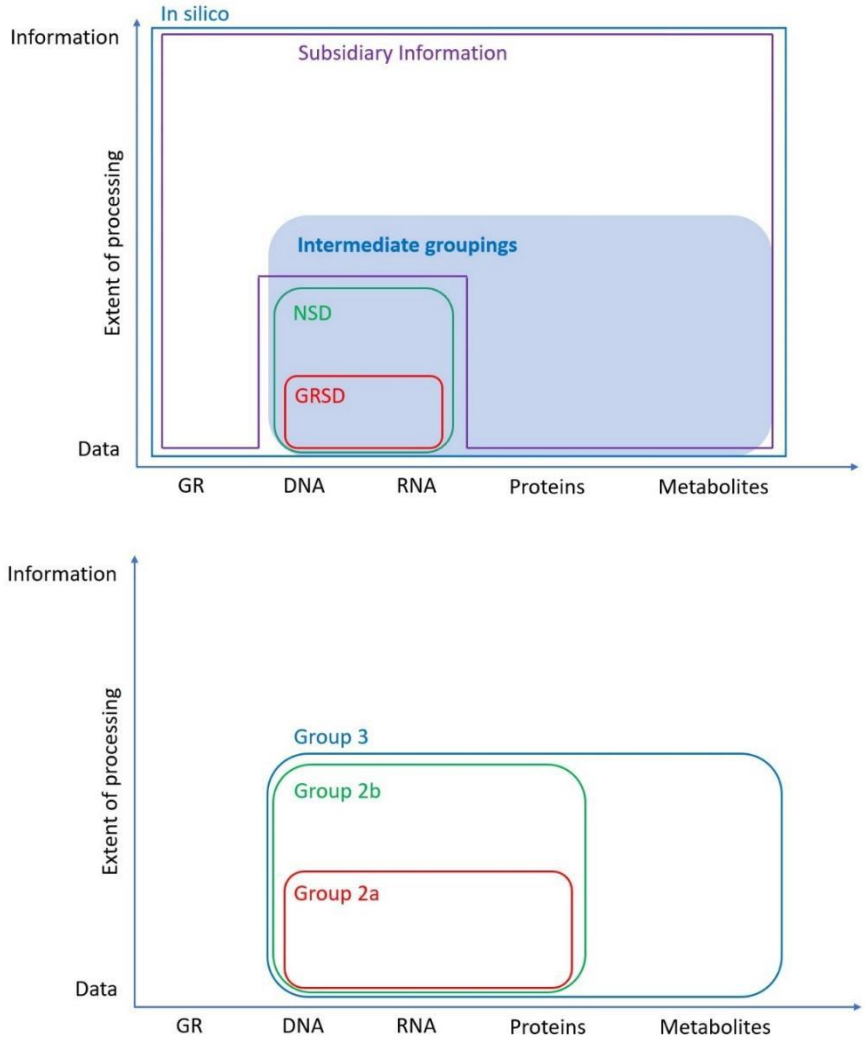


Figure 8. Evaluation of existing terms and proposed groupings to describe DSI. Top Panel: A graphical representation of the main terminologies proposed to replace 'DSI' showing the extent of biological processing carried out on data to convert it to information plotted against flow from genetic resource onwards to DNA, RNA, proteins and metabolites. The potential coverage of the two proposed intermediate groupings is indicated showing that it includes DNA, RNA, protein sequences, metabolites and a defined range of associated data and information selected from AHTEG categories a.-d. (for instance functional annotations of genes, gene expression information, epigenetic data, and molecular structures of proteins). **Bottom Panel:** The different ways that the intermediate subject matter grouping could be interpreted. **Group 2a** includes DNA/RNA sequence data including non-coding sequences, and information on the sequence assembly, including structural annotation and genetic mapping, as well as protein sequence data. **Group 2b** is the same as group 2a in addition to which it includes functional annotation of genes, gene expression information, epigenetic data, and molecular structures of proteins. **Group 3** is the same as group 2b, but adds data on other macromolecules and metabolites, including their molecular structures.

4.3.3 Intermediate scope: data/information associated with biological processes involving transcription and translation

Scope

As proposed above, Group 2 has an intermediate scope and extends to protein sequences, thus comprising information associated with transcription and translation. Two interpretations for the scope of this group are possible, either subject matter is strictly limited to nucleotide and protein sequence data (Group 2a), or it includes information associated with transcription and translation more broadly, for instance, functional annotations of genes, gene expression information, epigenetic data, and molecular structures of proteins (Group 2b).

Evaluation of Existing Terms

Digital Sequence Data (DSD) This includes DNA, RNA and protein sequences. However, since the word 'data' is used, this only refers to raw sequence data derived directly from genome and protein sequencing. Data that has been processed, such as automatic DNA annotation by comparison to other DNA sequences in the database or converting raw DNA data into protein sequences in an automated way, will be out of scope of DSI as it could now be considered 'information' and no longer data (AHTEG category a. only). This is the only existing term, that appears readily available to describe subject matter of intermediate scope comprising information associated with transcription and translation (as proposed for Group 2). However, the term is understood to be limited to raw protein sequence data and so would only be suitable for the narrow interpretation considered for scope in this group (group 2a).

Another term, **Genetic Resource Sequence Data (GRSD)**, which was intended by its proponent, the International Chamber of Commerce, to be limited to nucleotide sequences (see section 5.3.4), could be re-interpreted more broadly to describe DSI subject matter which includes protein sequences.[17] This is because the 'Genetic Resource' pre-fix gives the impression that the term covers sequence data related to a genetic resource more broadly. Although this term would also be suitable only for the narrow interpretation considered for scope in this group (group 2a), the broader interpretation (group 2b) could be accommodated through a minor modification to this term, **Genetic Resource Sequence Data and Information (GRSDI)**, which of course comprises both data and information related to proteins.[18]

4.3.4 Narrow scope: limited to nucleic acid sequence data associated with transcription

Scope

As proposed above, Group 1 has a narrow scope or proximity to the genetic resource and is limited to nucleotide sequence data associated with transcription.

Terms

Genetic Resource Sequence Data (GRSD)[75] The International Chamber of Commerce defines this as: "the description of the order of nucleotides (DNA or RNA), as found in nature, in the genome or encoded by the genome of a given genetic resource. The 'genome' includes nuclear and extra-nuclear DNA, and

[17] Note, in its peer review submission the International Chamber of Commerce objected to the suggested reinterpretation of this term to cover broader subject matter, with a specific reference to protein sequences. In doing so they highlight that their intention is to the contrary, to explicitly exclude protein sequences from DSI subject matter.

[18] The term "genetic sequence data and information" is one of the terms being considered within the BBNJ process, alongside the term "in silico" and other terms.

coding (gene) and non-coding DNA sequences. It does not include other molecules resulting from natural metabolic processes associated with or requiring the genetic resource. GRSD cannot, and does not, include information connected with or resulting from the analysis or further application of GRSD, e.g. sequence assembly, sequence annotation, genetic maps, metabolic maps, three-dimensional structure information or physiological properties related to it. Including information resulting from human interaction on GRSD would result in yielding man-made genetic sequences, which would no longer be considered GRSD." This definition is therefore narrower than the definition in AHTEG category a. as it explicitly excludes metabolites and by omission excludes protein sequences and metadata associated with the genetic resource. However, data on protein sequences can usually be predicted by automated analyses of the DNA sequences although there are exceptions (see section 3.2). Arguments around the use of the word 'data' were given above for DSD.

This definition is very clear in that it expressly includes sequences of all possible forms of DNA discovered to date, in particular non-coding DNA.[76] Non-coding DNA might be excluded by CBD Art 2 which defines 'genetic material' as meaning "any material of plant, animal, microbial or other origin containing functional units of heredity" as no function has yet been ascribed to some types of non-coding DNA.

The use of the word 'genetic' may be important here as it ascribes a function to the data but does not specify the molecular mechanism by which heredity should occur. This therefore potentially allows for the inclusion of modified DNA and RNA (see Sections 3.3 & 5.5), as long as these can transfer genetic information in a hereditary manner.

Genetic Sequences (GS) From PIP framework "the order of nucleotides found in a molecule of DNA or RNA. They contain the genetic information that determines the biological characteristics of an organism or a virus" (Art 4.2)[77] This could refer to the actual DNA or RNA from the genetic resource or the sequence data/information. This definition makes clear the extent of what is included, only DNA/RNA (AHTEG categories a. and b.), and excludes proteins, metabolites and metadata associated with the genetic resource.

Genetic Sequence Data/Information (GSD/GSI). Like GRSD, this refers only to genetic data/information, but additional clarity will be needed to indicate that this is derived from a genetic resource. This includes only DNA and RNA sequences and not protein sequences or information on metabolites, thus encompassing only AHTEG categories a. and b.

Nucleotide Sequence Data (NSD) and Subsidiary Information (SI). NSD is more specific than GSD/GRSD and includes only DNA and RNA sequence data, and expressly refers to the chemical structure of the component nucleotides. It refers only to the AHTEG categories a.-b., and an accessory term, 'subsidiary information' (SI) is introduced to cover metadata associated with the genetic resource, data on proteins and metabolites, thus encompassing AHTEG categories c.-i. These are the terms used by the International Nucleotide Sequence Database Collaboration (INSDC, see DSI study 2/3 for additional discussion). The institutes that run the INSDC also run additional databases that contain information on protein sequences derived from gene predictions and translations of DNA sequences.

A description of the relationship between NSD and SI is given in a recent submission[78]: "NSD include non-coding & coding sequences, regulatory sequences, conserved sequences, genes that encode specific traits, DNA without known function and 'junk DNA'. Larger data elements would include the entire genome of an organism [or, indeed, of a clade (pangenome) or environmental sample (metagenome)]. NSD are aggregated from naturally occurring genetic resources generated as a part of research or

downloaded from INSDC and other databases. Analyses of NSD are interpreted in research to develop understanding of biological diversity at genetic, species and ecosystem levels."

By the specificity of the terminology NSD, it excludes DNA and RNA in which the nucleotides have been modified so that they can no longer be regarded as nucleotides, despite their potential to carry the genetic code and be duplicated (see Sections 3.3 & 5.5). A second comment is the lack of function ascribed to the nucleotides, which is apparent in the use of the term 'genetic' in the previous three definitions. By not referring to function, non-coding DNA is also brought within scope of DSI as is clear from the definition of NSD above.

4.4 Digital Sequence Information

The 2018 Laird and Wynberg[1] study summarizes objections to this terminology and explains why it is not appropriate to describe the elements of 'genetic information' that might be included under the CBD. We build on this by considering each of the constituent elements of 'DSI', in turn, and in the process identify important issues which need to be considered in order to clarify the concept of DSI. We use the Oxford English Dictionary (OED) definitions for 'Digital', 'Sequence' and 'Information' and provide an analysis concerning the suitability/desirability of each term in any in terminology proposed to replace DSI. Where relevant, we also assess implications regarding the type of information that may be associated with the concept of DSI, arising from the use of the constitutive term.

4.4.1 Digital (OED)

"Of signals, information, or data: represented by a series of discrete values (commonly the numbers 0 and 1), typically for electronic storage or processing."

The word 'digital' only refers to the way in which data is held, implying it is in computer memory or data storage, and to counter this it is stated that this data can also be held in other forms such as on paper. However, DNA sequences printed on paper are machine readable, but are not 'digital' in this sense and would therefore be out of scope of DSI, despite conveying the same information.

4.4.2 Sequence (OED)

"The fact of following after or succeeding; the following of one thing after another in succession; an instance of this." A subsidiary definition is given for biochemistry: "The order of the constituent nucleotides in a nucleic acid molecule or of the amino-acids in a polypeptide or protein molecule."

Anything stored on a computer is in the form of a sequence such as '001100100 …' and would be captured by using the term 'digital', 'sequence' and 'digital sequence'. The term 'sequence' is applied to DNA, RNA and proteins, whose subunit nucleotides, for DNA/RNA, and amino acids, for proteins, can be described by sequences of letters or groups of letters (e.g. the amino acid chain MARWAELCEL can also be given as Met-Ala-Arg-Trp-Ala-Glu-Leu-Cys-Glu-Leu). Whilst this information is useful, it gives no indication of the gene function or expression level of these sequences. For DNA, the sequence alone does not indicate gene expression, its effect on phenotype, and many other important characteristics ('Broad' subject matter grouping, Group 4, AHTEG categories c, e, f, g, h).

Length or function of Sequence. The length of the sequence and its function may govern whether a DNA sequence is unique to a particular genetic resource or origin. Table 4 in Study 2/3 concerning databases and traceability associated with DSI, shows that statistically sequences below 30 nucleotides may not be unique, meaning a search of a sequence of less than 30 nucleotides may yield multiple results from

different organisms found in different countries. In addition, it must be considered that not all sequence variation is governed only by random factors, but it is governed by selection that could lead to convergence for some DNA sequences, meaning that the same sequence, longer than 30 residues, could occur in multiple species. Parties need to consider a minimum sequence length taking into consideration the data presented in Table 4 in Study 2/3. In addition, whether non-coding elements such as promotors, which are functional but do not encode proteins, should be regarded as being within the scope of 'DSI' needs to be considered, as should elements such as BioBricks which serve a variety of functions and may be natural, modified or synthetic DNA sequences (Section 3.7).

Environmental and metagenomic DNA. Acquisition of environmental and metagenomic DNA is now common in many research areas (Section 3.5). In the context of the CBD, the genetic resource underpinning such is the combined DNA in the sample and not the unique organisms from which they arise. Whereas it is possible therefore to connect the DNA sequences to the genetic resource, it will be very difficult to connect them to the originating organism, thus raising problems of traceability for such materials. Most of the environmental and metagenomic DNA sequences will be partial or incomplete, but they are still vital to the understanding of community structures and many other applications.

Microarray data. It is not clear whether microarray data (Section 3.8.1) would be regarded as 'sequence' data. The microarray readout is a quantity of light (fluorescence) and is conceivably not directly sequence data. If gene expression (AHTEG category c.) is included in any term used to replace DSI then this brings microarray data within scope of DSI. Microarray data would therefore be included in Intermediate Groups 2b and 3.

Three-dimensional structural information. The word 'sequence' approached in this way would also exclude three-dimensional structural information on DNA, RNA and proteins, which is essential to understand their biological function and interaction with DNA, RNA, proteins and metabolites. This distinction was used in the case between D'Arcy and Myriad Genetics Inc., heard at the High Court of Australia, where the chemical composition of a DNA sequence was regarded as different from the information that this genetic sequence contained. Structural information on proteins (atom coordinates) is contained within standardised text files known as 'pdb' files (Section 3.8.2). Three-dimensional structural information is included in Intermediate groups 2b and 3.

Macromolecules. The use of the word 'macromolecule' in AHTEG category d. could also cause confusion as this includes all DNA, RNA, proteins, polysaccharides amongst others. Polysaccharides are chains of sugar molecules that are frequently encountered in biology and can be regarded as macromolecules that can be represented as sequences. Examples include starch, cellulose and glycogen which act as different types of energy stores. They can form very long linear or branched chains of the same sugar molecule, such as starch, which can contain more than a thousand molecules of glucose joined in a uniform linear way. A more complex example is the recognition by the immune system of complex sequences of sugars in potential pathogens. All of these complex sequences of sugars are the outcome of an organism's metabolism, the interaction of many proteins working together to generate polysaccharides and these could therefore be defined as 'derivatives'.[19] Using the word 'sequence', without clearly defining

[19] Nagoya Protocol Article 2c: " 'Derivative' means a naturally occurring biochemical compound resulting from the genetic expression or metabolism of biological or genetic resources, even if it does not contain functional units of heredity"

sequences of which type of subunit, might therefore bring polysaccharides within scope of DSI. Data on macromolecules including their molecular structures is included in Intermediate group 3.

Alternative representations of metabolites. The use of the word 'sequence' would appear to exclude small molecule metabolites which fall under 'derivatives' under the Nagoya Protocol (AHTEG category d). However, molecular structures can be represented and stored as 'sequences' as SMILES (Simplified Molecular Input Line Entry Specification, Figure 9) which includes information on molecular connectivity without specifying two or three-dimensional coordinates of the atoms in the molecule. In mathematical terminology it is a 'molecular graph' expressed as a unique 'sequence'. For each SMILES there is only one possible molecular graph (molecular structure) and *vice versa*. If it is accepted that metabolites can be described as 'sequences' in this way, this could bring small molecule metabolites within scope if the word 'sequence' is used in the eventual definition. All molecules can be described in this way, including atom-level descriptions of DNA, RNA, proteins, polysaccharides and metabolites, meaning that any definition including the word 'sequence' would include these. In the current proposal for intermediate groups, data on cellular metabolites, including molecular structures is included in Group 3.

A key issue in clarifying the concept of DSI is to consider which types of 'sequence' should be included in any replacement terminology for DSI. If the definition of 'sequence' only includes DNA, RNA and proteins and not sequential representations of small molecules (e.g. as SMILES strings) then this brings with it the possibility of describing DNA, RNA and proteins as SMILES strings which under this interpretation would not be regarded as sequences. Using this approach, the same information is conveyed without using the normal sequence representation of DNA, RNA (using the 4-letter nucleotide code) or proteins (using the 20-letter amino acid code).

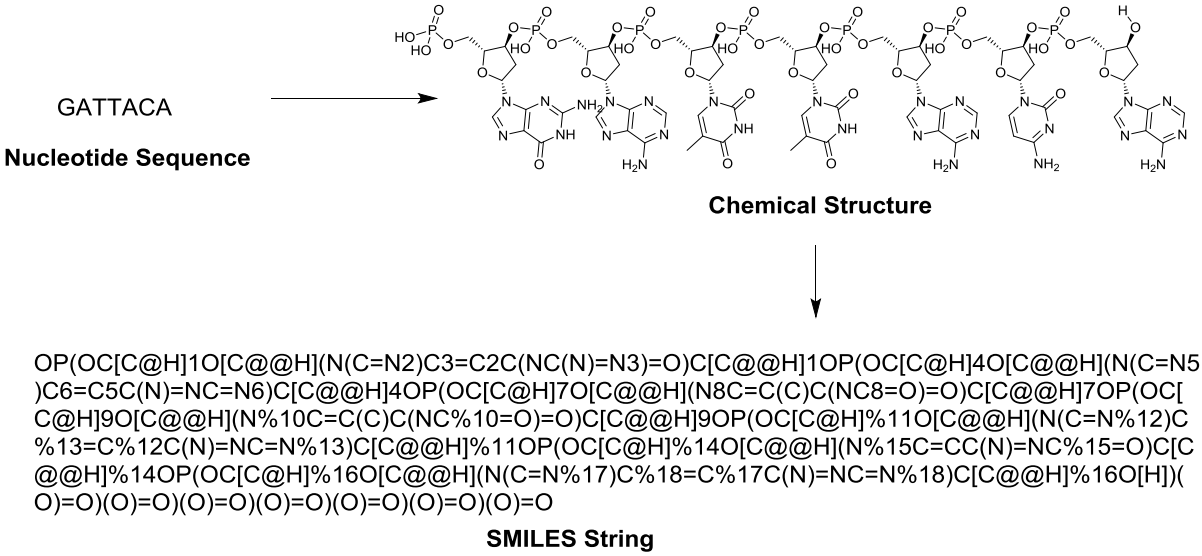


Figure 9. The relationship between nucleotide sequence, chemical structure and SMILES string of the same DNA sequence.

4.4.3 Information (OED)

"That which is obtained by the processing of data."

Data (OED): "Related items of (chiefly numerical) information considered collectively, typically obtained by scientific work and used for reference, analysis, or calculation."

Broader definitions of 'information' include the imparting of information in general, as alluded to in Section 5.2. Genes carry information which is interpreted through the action of transcription and translation and is under the influence of environmental factors to give rise to the phenotype of an organism (See Figure 3). Mutations and selection pressures lead to the evolution of these genes over time, thus altering this information in a continuous process.

The use of the word information in keeping with the OED definition above implies that the raw data (e.g. the raw sequences of nucleotides in DNA/RNA or amino acids in proteins) has been processed in some way or has had some value added by processing. The question is whether automated tools such as automated annotation or converting DNA sequences by translating codons into protein sequences are enough to convert data into information, or whether human intervention or curation is essential (see Section 3.4).

The error rate for the different DNA sequencing methodologies discussed in section 3.4 and Table 2 must be considered here as different levels of data processing are necessary to convert raw 'reads' into an accurate DNA sequence, which may be construed as converting data into information. The distinction between the terms 'data' and 'information' has been heavily discussed in submissions to the CBD DSI process and there appears to be a consensus that the difference between data and information is the level of processing that has been executed. In the context of DSI it will be helpful to develop clarity on what automatic and semi-automatic processing might be included and what might be excluded from the scope of subject matter comprising DSI. For example, sequence alignment could be included as it is a necessary and semi-automated element of developing sequence data for further analysis. This is included in most of the proposed terminology introduced in Section 5.3 except for ICC's original definition of genetic resource sequence data which includes only raw sequence data but excludes aligned sequences.

4.5 Modifications DNA, RNA and protein sequences and their subunits

The modification of DNA, RNA and protein sequences and their subunits (nucleotides, amino acids) were discussed in section 3.3. These must be considered in clarifying the concept of DSI as such modifications influence the possible scope of subject matter constituting DSI, as per the terminology proposed to replace DSI discussed above. The key issue is whether only unmodified DNA, RNA and proteins are within scope of any terminology proposed to replace DSI, and if not, what is the extent of modification that is permitted if modifications are to fall within the new DSI terminology.

DNA Modifications. Designer synthetic DNA sequences can generate wholly new proteins, and as these do not trace back to any genetic resource these do not fall within scope of any definition relating to DSI. Changes to nucleotides, such as modifying the base, sugar or phosphate may no longer be considered nucleotides. Therefore, any terminology to replace DSI that includes 'nucleotide' would leave these novel, synthetic analogues out of scope, even though in principle they may have the same function as DNA derived from a genetic resource. The nucleotide structure can be retained, but the number of bases that can be used can be increased, but as these are not natural and do not derive from a genetic resource these will fall outside of the scope of any terminology proposed to replace DSI. Accordingly, in order to clarify the concept of DSI it needs to be considered whether these modifications should all be regarded as 'unnatural' and whether those that do not arise from a genetic resource should be considered as DSI for policy discussions or not, even if they have the same function as DNA derived from a genetic resource.

If only DNA and RNA sequences are included in the eventual definition of DSI (as per Group 1), then epigenetic methylation of DNA may be excluded. It is important to consider that methylation does not affect the nucleotide sequence and the pattern is not easily predicted, but it does influence gene expression.

Protein modifications. Modifications to proteins including phosphorylation, and compounds containing amino acid subunits such as RiPP and NRPS are one step further along the flow from genetic resource to DNA, RNA onwards and do not fit neatly under the terms 'proteins' or 'metabolites'. To clarify the concept of DSI it first needs to be considered whether proteins are included under any terminology replacing DSI, and if so what level of modification is considered allowable.

5. CONCLUSIONS AND IMPLICATIONS FOR FUTURE DISCUSSIONS CONCERNING DSI

5.1 Subject matter groupings

Considering the flow of data/information from, and its proximity to, an underlying genetic resource provides a basis to group information that may comprise DSI and this gives rise to four logical groupings. To recap, Group 1 has a narrow scope or proximity to the genetic resource and is limited to nucleotide sequence data associated with transcription. Group 2 has an intermediate scope and extends to protein sequences, thus comprising information associated with transcription and translation. Two interpretations for the scope of this group are possible, either subject matter is strictly limited to nucleotide and protein sequence data or it includes information associated with transcription and translation more broadly, for instance, functional annotations of genes, gene expression information, epigenetic data, and molecular structures of proteins. Group 3 has a wider intermediate scope and extends to metabolites and biochemical pathways, thus comprising information associated with transcription, translation and biosynthesis. Group 4 has the broadest scope and includes data/information with the weakest proximity to the underlying genetic resource, thus extending to behavioral data, information on ecological relationships and traditional knowledge, thus comprising information associated with transcription, translation and biosynthesis, as well as downstream subsidiary information. These groupings could be used in future discussions when discussing DSI subject matter and terminology (e.g. instead of the AHTEG list).

The proximity of information to the underlying genetic resource is a useful proxy to determine whether it is possible to accurately identify or infer the source from which it is derived. This has implications for the traceability of information to a particular genetic resource and also in identifying the source of information, including whether it has been generated through the utilization of a genetic resource or independently. If traceability of DSI is important, a narrow scope of DSI subject matter appears more desirable given the technical difficulties in identifying or inferring origin, whereas if traceability is not important a broader scope of subject matter may be able to be accommodated.

5.2 Priority issues to clarify the concept of DSI

Throughout this study we have identified a number of priority issues that should be addressed in order to clarify the concept of DSI. Irrespective of whether the logical groups proposed in this study are adopted, these issues should be used to help guide deliberations concerning the scope and concept of DSI. These issues can be summarized as follows and for each we propose a logical approach that may assist in resolving the issue:

1.) How far along the flow from genetic resource onwards to DNA, RNA, protein sequences and metabolites DSI can be considered to extend. Specifically: whether macromolecules (e.g. proteins, polysaccharides) are included under 'DSI' and whether small molecules (metabolites) are included under DSI – *this can be resolved by utilizing the four groups proposed to clarify the scope of DSI subject matter, in which case all macromolecules (non DNA/RNA) and metabolites would be excluded under Group 1 or 2, whereas they would be included under Groups 3 or 4.*

2.) The distinction between data and information and how this is stored and processed, including the extent to which data has been processed before it can be considered information – *this can be resolved by utilizing the four groups proposed to clarify the scope of DSI subject matter as these have clear subject matter boundaries and so an approach, criteria or definition for distinguishing between data and information is not necessary*.

3.) Types of sequences that are included under any terminology proposed to replace DSI. Specific questions are:

   a. What length of sequence can still be considered as a 'sequence' - *sequences below 30 nucleotides may not be unique and so this may provide a logical threshold below which information should be excluded from DSI subject matter.*

   b. Whether non-coding DNA should be included under 'DSI'- *genetic elements which do not encode proteins (such as promotors) may have a natural functional role in transcription, translation or biosynthesis and on this basis it may be considered an inherent part of the underlying genetic resource, such that it would be illogical to distinguish between coding and non-coding sequences.*

   c. Whether epigenetic heritable factors should be included under DSI - u*sing the same rationale as in b. above, epigenetic heritable factors may have a natural functional role in transcription, translation or biosynthesis and therefore it may be logical to exclude it from DSI subject matter (assuming the rationale for non-coding DNA is also accepted).*

   d. Whether modified DNA, RNA (and proteins) should be included under DSI- u*sing the same rationale as in b. and c. above, naturally modified DNA, RNA or proteins may nevertheless have a natural functional role in transcription, translation or biosynthesis and on this basis these may be considered an inherent part of the underlying genetic resource such that it would be illogical to exclude from DSI subject matter, at least to the same extent that DSI subject matter includes DNA, RNA and/or proteins. Conversely synthetically modified DNA, RNA or proteins cannot be said to have a natural functional role and so on this basis could be considered not to be an inherent part of the underlying genetic resource.*

5.3 Subject matter groupings and life-science sectors

Section 4 provided illustrative insights regarding how different life-sciences sectors utilize DSI, including by comparing and contrasting their reliance on 'omics' technologies (Table 3). Applying the subject matter groupings proposed in Section 5 we are able to consider the implications for each sector if DSI subject matter is construed in a narrow, intermediate or broad manner, as depicted in Table 5 and

described in greater detail below. Unsurprisingly given our earlier observations showing a heavy reliance on 'omics' technologies and trends and technologies enabled by DSI, Table 5 confirms that all sectors rely heavily on DNA and RNA sequence data (narrow group) and on functional annotations of genes and protein sequence data obtained via proteomic techniques (intermediate group 2a/b), while molecular structures of proteins and metabolomic data (intermediate group 3) are particularly important in taxonomy & conservation (as in other fields of basic research, of course), industrial biotechnology, synthetic biology, healthcare and drug discovery. Irrespective of whether a narrow, intermediate or broad approach is used in defining the scope of DSI, all sectors would be within scope as all use information and applications/technologies which rely on such information, within each grouping. Of course, the broader the scope of DSI subject matter adopted, the greater the technologies, techniques and overall activities in each sector that would rely on information that falls within the scope of DSI.

From a technical perspective, this study builds on the 2018 Laird and Wynberg study by providing greater technical context concerning the generation and use of DSI (Section 3) and by providing illustrative insights concerning certain life-sciences sectors that rely on DSI and technologies/techniques, enabled by DSI (Section 4). More comprehensive technical coverage regarding the use of DSI and technologies enabled by DSI in commercially orientated research and development, including insights regarding the extent to which such uses are the subject of patent claims (for example, in the life sciences sectors covered in this study), would help further clarify the concept of DSI by facilitating more nuanced discussions concerning the possible implications of including or excluding particular types of information associated with an underlying genetic resource from the scope of DSI subject matter, within the context of the CBD and the Nagoya Protocol.

Table 5. Selected examples of applying the proposed DSI subject matter groupings to the different life-sciences sectors.

DSI **Subject Matter Grouping**	**Taxonomy & Conservation**	**Agriculture & Food Security**	**Industrial & Synthetic Biology**	**Healthcare & Pharmaceuticals**
Narrow (Group 1) (DNA/RNA)	Most critical to this field of use is DNA/RNA sequence data, and ability to compare these the sequence databases	Reference genomes are needed to carry out any marker assisted breeding or genetic modification. Metagenomes critical to understand health of soil microbiome.	Availability of multiple related gene sequences for comparison is essential for this field.	DNA and RNA sequences essential to this field.
Intermediate (Group 2) (DNA/RNA/Proteins)		Gene annotations needed for marker assisted breeding. Gene expression information relevant. Epigenetic heritable elements are important. Protein sequences less important than DNA/RNA sequences. Proteomic data used to assess effect of breeding/genetic modification.	Gene annotations needed to discover related enzymes. Information on protein sequences is essential for the engineering of enzymes. Molecular structure of proteins is important needed for targeting modifications.	Gene annotations needed to discover related proteins. Information on gene expression is needed to understand essential metabolic processes in pathogens. Protein sequences are important to understand how small molecule metabolites are biosynthesized. Molecular structures of proteins are necessary to understand and engineer metabolite biosynthesis.
Intermediate (Group 3) (DNA/RNA/Proteins/Metabolites)	Metabolome of organism can be used to assist taxonomy.	Metabolomic data important to understand nutritional value of crops.	Information on molecular structure of metabolites needed for many products.	Information on molecular structure of metabolites needed for many products.
Broad (Group 4) (DNA/RNA/Proteins/Metabolites and subsidiary information)	For taxonomic description this includes morphological data. Species richness and assemblage may change in response to abiotic and environmental influences.	Phenotype, ecological relationships, environmental and abiotic factors are relevant.	Understanding external biotic and abiotic stresses that elicit protein and metabolite production.	Ecological relationship and abiotic environmental factors are important to understand pathogen evolution and spread.

6. ACKNOWLEDGMENTS

The following are thanked for their helpful discussions and review of drafts: Abbe Brown, Lydia Slobodian, Sarah Laird, Chris Lyal, Charles Lawson and Amber Scholz. The staff at the CBD Secretariat also deserve thanks: Valerie Normand, Beatriz Gomez, Austein McLoughlin, Kristina Taboulchanas, Worku Yifru and David Cooper.

7. CONFLICT OF INTEREST STATEMENT

Marcel Jaspars is a co-founder of, has shares in, and consultant to GyreOx Ltd, a company that uses marine genetic resources from areas within national jurisdiction to develop potential drug molecules.

8. REFERENCES

1. S. A. Laird, R. P. Wynberg (2018) Fact-Finding and Scoping Study on Digital Sequence Information on Genetic Resources in the Context of the Convention on Biological Diversity and the Nagoya Protocol (CBD/DSI/AHTEG/2018/1/3); available at https://www.cbd.int/doc/c/079f/2dc5/2d20217d1cdacac787524d8e/dsi-ahteg-2018-01-03-en.pdf.
2. E. Schrödinger (1944) What is life? The physical aspect of the living cell. Cambridge University Press, UK.
3. O. T. Avery, C. M. Macleod, M. Mccarty (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.* 79: 137-158.
4. E. Vischer, E. Chargaff (1948) The separation and quantitative estimation of purines and pyrimidines in minute amounts. *J. Biol. Chem.* 176: 703-714.
5. A. J. F. Griffiths, S. R. Wessler, S. B. Carroll, J. Doebley (2012) Introduction to genetic analysis (third edition) W. H. Freeman and Company, New York.
6. L. Pray (2008) Discovery of DNA structure and function: Watson and Crick. *Nature Education* 1: 100.
7. A. Rich, D. Davies (1956) A new two-stranded helical structure: polyadenylic acid and polyuridylic acid. *J. Am. Chem. Soc.* 78: 3548-3549.
8. F.H. Crick (1958) On protein synthesis. In F. K. Sanders (ed.). *Symposia of the Society of Experimental Biology*. Number XII: The biological replication of macromolecules. Cambridge University Press pp. 138-163.
9. R.S. Gardner, A. J. Wahba, C. Basilio, R. S. Miller, P. Lengyel, J. F. Speyer (1962) Synthetic polynucleotides and the amino acid code, VII. *Proc. Natl. Acad. Sci. USA* 48: 2087-2094.
10. A. J. Wahba, R.S. Gardner, C. Basilio, R. S. Miller, J. F. Speyer, P. Lengyel (1963) Synthetic polynucleotides and the amino acid code, VIII. *Proc. Natl. Acad. Sci. USA* 49: 116-122.
11. S. Horowitz, M. A. Gorovsky (1985) An unusual genetic code in nuclear genes of *Tetrahymena*. *Proc. Natl. Acad. Sci. USA* 82: 2452-2455.
12. S. T. Sharfstein (2018) Non-protein biologic therapeutics. *Curr. Opin. Biotechnol.* 53: 65-75.
13. C. Böhler, P. E. Nielsen, L. E. Orgel (1995) Template switching between PNA and RNA oligonucleotides. *Nature* 376, 578-581.

14. S. Hoshika, N. A. Leal, M.-J. Kim, M.-S. Kim, N. B. Karalkar, H. Kim, et al. (2019) Hachimoji DNA and RNA: A genetic system with eight building blocks. *Science* 363, 884-887.

15. F. Sanger, S. Nicklen, A. R. Coulson (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74: 5463-5467.

16. L. J. Kahl, D. Endy (2013) A survey of enabling technologies in synthetic biology. *J. Biol. Eng.* 7: 13.

17. F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen (1982) Nucleotide sequence of bacteriophage λ DNA. *J. Mol. Biol.* 162: 729-773.

18. D. Sim, I. Sudbury, N. E. Ilott, A. Heger, C. P. Ponting (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15: 121-132.

19. J. Besser, H. A. Carleton, P. Gerner-Smidt, R. L. Lindsey, E. Trees (2018) Next-Generation Sequencing Technologies and their Application to the Study and Control of Bacterial Infections. *Clin Microbiol Infect.* 24: 335-341

20. R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd. *Science* 269: 496-498.

21. M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, *et al*. (2000) The Genome Sequence of *Drosophila melanogaster*. *Science* 287: 2185-2195.

22. R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520-562.

23. J. Yu, S. Hu, J. Wang, G. K. Wong, S. Li, *et al*. (2002) A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79-92.

24. S. A. Goff, D. Ricke, T. Lan, G. Presting, R. Wang, M. Dunn, *et al*. (2002) A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92-100.

25. H. A. Lewin, G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington *et al*. (2018) Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. USA* 115: 4325-4333.

26. T. Woyke, G. Xie, A. Copeland, J. M. González, C. Han, *et al*. (2009) Assembling the marine metagenome, one cell at a time. *PLoS One* 4: e5299.

27. S. Mariani, C. Baillie, G. Colosimo, A. Riesgo (2019) Sponges as natural environmental DNA samplers. *Curr. Biol.* 29: R401-R402.

28. J. Eberwine, J. Y. Sul, T. Bartfai, J. Kim (2014) The promise of single-cell sequencing. *Nat. Methods* 11: 25-27.

29. K. Chen, L. Pachter (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput. Biol.* 1: 106-112.

30. P. Horvath, R. Barrangou (2010) CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science* 327: 167-170.

31. S. Mukherjee (2016) The Gene- An intimate history. Vintage, London.

32. A. Cheng, T. K. Lu (2012) Synthetic biology: an emerging engineering discipline. *Annu. Rev. Biomed Eng.* 14: 155-178.

33. H. Neumann (2012) Rewriting translation – genetic code expansion and its applications. *FEBS Lett.* 586: 2057-2064.

34. M. J. Lajoie, A. J. Rovner, D. B. Goodman, H. Aerni, A. D. Haimovich, *et al.* (2013) Genomically recoded organisms expand biological functions. *Science* 342: 357-360.

35. J. C. Anderson, N. Wu, S. W. Santoro, V. Lakshman, D. S. King, *et al.* (2004) An expanded genetic code with a functional quadruplet codon. *Proc. Natl. Acad. Sci. USA* 101: 7566-7571.

36. C. B. Anfinsen (1972) Studies on the principles that govern the folding of protein chains. Nobel Lecture available at:

https://www.nobelprize.org/uploads/2018/06/anfinsen-lecture.pdf

37. 13th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction available at:

http://predictioncenter.org/casp13/

38. P. Crews, J. Rodriguez, M. Jaspars (2010) Organic Structure Analysis. Oxford University Press, New York

39. A. Noreña-P, A. G. Muñoz, J. Mosquera-Rendón, K. Botero, M. A. Cristancho (2018) Colombia, an unknown genetic diversity in the era of Big Data. *BMC Genomics* 19: 859.

40. G. M. Connette, P. Oswald, M. K. Thura, K. J. LaJeunesse Connette, M. E. Grindley, *et al*. (2017) Rapid forest clearing in a Myanmar proposed national park threatens two newly discovered species of geckos (Gekkonidae: *Cyrtodactylus*). *PLoS ONE* 12: e0174432.

41. T. E. Berry, B. J. Saunders, M. L. Coghlan, M. Stat, S. Jarman, *et al.* (2019) Marine environmental DNA biomonitoring reveals seasonal patterns in biodiversity and identifies ecosystem responses to anomalous climatic events. *PLoS Genet.* 15: e1007943.

42. S. P. Iglésias, L. Toulhoat, D. Y. Sellos (2010) Taxonomic confusion and market mislabelling of threatened skates: important consequences for their conservation status. *Aquatic Conserv: Mar. Freshw. Ecosyst.* 20: 319–333.

43. S. J. Helyar, A. D. Lioyd, M. de Bruyn, J. Leake, N. Bennett, G. R. Carvalho (2014) Fish Product Mislabelling: Failings of Traceability in the Production Chain and Implications for Illegal, Unreported and Unregulated (IUU) Fishing. *PLoS ONE* 9: e98691.

44. A. Barone, L. Frusciante (2007) Molecular marker-assisted selection for resistance to pathogens in tomato. In: “Marker-assisted selection: current status and future perspectives in crops, livestock, forestry and fish” [E. P. Guimarães, J. Ruane, B. D. Scherf, A. Sonnino, J. D. Dargie (eds.)], Food and Agriculture organization of the United Nations (Rome).

45. J. I. Weller (2007) Marker-assisted selection in dairy cattle. In: “Marker-assisted selection: current status and future perspectives in crops, livestock, forestry and fish” [E. P. Guimarães, J. Ruane, B. D. Scherf, A. Sonnino, J. D. Dargie (eds.)], Food and Agriculture organization of the United Nations (Rome).

46. P. Gallusci, Z. Dai, M. Génard, A. Gauffretau, N. Leblanc-Fournier, C. Richard-Molard, D. Vile, S. Brunel-Muguet (2017) Epigenetics for Plant Improvement: Current Knowledge and Modeling Avenues. *Trends Plant Sci* 22, 610-623.

47. N. M. Springer, R. J. Schmitz (2017) Exploiting induced and natural epigenetic variation for crop improvement. *Nat. Rev. Genetics* 18, 563-575.

48. M. Eldakak, S. M. Milad, A. I. Nawar, J. S. Rohila (2013) Proteomics: a biotechnology tool for crop improvement. *Front. Plant Sci.* 4, 35.

49. N. Amiour, S. Imbaud, G. Clément, N. Agier, M. Zivy, B. Valot, *et al.* (2012) The use of metabolomics integrated with transcriptomic and proteomic studies for identifying key steps involved in the control of nitrogen metabolism in crops such as maize. *J. Exp. Botany* 63, 5017-5033.

50. A. Kamthan, A. Chaudhuri, M. Kamthan, A. Datta (2015) Small RNAs in plants: recent development and application for crop improvement. *Front. Plant Sci.* 6, 208.

51. T. Wang, H. Zhang, H. Zhu (2019) CRISPR technology is revolutionizing the improvement of tomato and other fruit crops. *Horticult. Res.* 6, 77.

52. Z. H. Lemmon, N. T. Reem, J. Dalrymple, S. Soyk, K. E. Swartwood, D. Rodriguez-Leal, J. Van Eck, Z. B. Lippman (2018) Rapid improvement of domestication traits in an orphan crop by genome editing. *Nat. Plants* 4, 766-770.

53. A. Zsögön, T. Čermák, E. R. Naves, M. M. Notini, K. H. Edel, S. Weinl, *et al.* (2018) De novo domestication of wild tomato using genome editing. *Nat. Biotechnol.* 36, 1211-1216.

54. J. K. Jansson, K. S. Hofmockel (2018) The soil microbiome—from metagenomics to metaphenomics. *Curr. Opin. Microbiol.* 43: 162–168.

55. An explanatory guide to the Nagoya Protocol on access and benefit-sharing; available at https://www.iucn.org/content/explanatory-guide-nagoya-protocol-access-and-benefit-sharing

56. https://www.alliedmarketresearch.com/synthetic-biology-market

57. P.H. Nielsen (2005) Life cycle assessment supports cold-wash enzymes. *Int. J. Appl. Sci.* 10: 131.

58. J. Bjerre, O. Simonsen, J. Vind (2013) Household and personal care today. Vol. 8, p. 37.

59. T. H. Richardson, X. Tan, G. Frey, W. Callen, M. Cabell, *et al.* (2002) A novel, high performance enzyme for starch liquefaction - Discovery and optimization of a low pH, thermostable α-amylase. *J. Biol. Chem.* 277: 26501-26507.

60. R. Sadre, P. Kuo, J. Chen, Y. Yang, A. Banerjee, C. Benning, B. Hamberger (2019) Cytosolic lipid droplets as engineered organelles for production and accumulation of terpenoid biomaterials in leaves. *Nat. Commun.* 10: 853.

61. X. Luo, M. A. Reiter, d'Espaux, J. Wong, C. M. Denby, A. Lechner, *et al.* (2019) Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. *Nature* 567: 123-126.

62. V. Chubukov, A. Mukhopadhyay, C. J. Petzold, J. D. Keasling, H. G. Martin (2016) Synthetic and systems biology for microbial production of commodity chemicals. *NPJ Syst. Biol. Appl.* 2: 16009.

63. A. Cravens, J. Payne, C.D. Smolke (2019) Synthetic biology strategies for microbial biosynthesis of plant natural products. *Nat. Commun.* 10: 2142.

64. C. J. Paddon, J. D. Keasling (2014) Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat. Rev. Microbiol.* 12: 355-367.

65. https://www.webmd.com/drug-medication/news/20150508/most-prescribed-top-selling-drugs

66. D. J. Newman, G. M. Cragg (2016) Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* 79: 629-661.

67. G. Poste (2001) Molecular diagnostics: a powerful new component of the healthcare value chain. *Expert Rev. Mol. Diagn.* 1: 1-5.

68. P. Qin, M. Park, K. J. Alfson, M. Tamhankar, R. Carrion, *et al.* (2019) Rapid and Fully Microfluidic Ebola Virus Detection with CRISPRCas13a. *ACS Sens.* 4: 1048-1054.

69. https://www.ncbi.nlm.nih.gov/pathogens/

70. A. Zipperer, M. C. Konnerth, C. Laux, A. Berscheid, D. Janek, et al. (2016) Human commensals producing a novel antibiotic impair pathogen colonization. *Nature* 535: 511-516.

71. Recommendation adopted by the subsidiary body on scientific, technical and technological advice. Twenty-second meeting Montreal, Canada, 2-7 July 2018 available at:

https://www.cbd.int/doc/recommendations/sbstta-22/sbstta-22-rec-01-en.pdf

72. A. Traoré (The dematerialization of plant genetic resources: A peasant's perspective)- available at: https://www.righttofoodandnutrition.org/files/2._eng_the_dematerialization_of_plant_genetic_resources.pdf

73. C. Lawson, F. Humphries, M. Rourke (2019) The future of information under the CBD, Nagoya Protocol, Plant Treaty, and PIP Framework. *J. World Intellect. Prop.* 22: 1-17.

74. International Chamber of Commerce submission to the CBD (2016) Digital Sequence Information (https://iccwbo.org/publication/digital-sequence-information/)

75. M. A. Bagley, A. K. Rai (2014) The Nagoya protocol and synthetic biology research: A look at the potential impacts. Washington, DC.

76. The World Health Organization (2011) Pandemic influenza preparedness Framework for the sharing of influenza viruses and access to vaccines and other benefits available at: https://www.who.int/influenza/resources/pip_framework/en/

77. Submission of views and information on benefit sharing arrangements from commercial and non-commercial use of digital sequence information on genetic resources- available at:

https://www.cbd.int/abs/DSI-views/2019/NHM-RBGK-RBGE-DSI.pdf

_____