July 2018 CGRFA/EG-MIGR-1/18/3 **E**

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

COMMISSION ON GENETIC RESOURCES FOR FOOD AND AGRICULTURE

Item 4 of the Provisional Agenda

EXPERT GROUP ON MICRO-ORGANISM AND INVERTEBRATE GENETIC RESOURCES FOR FOOD AND AGRICULTURE

First Meeting

Rome, 3–5 October 2018

DRAFT EXPLORATORY FACT-FINDING SCOPING STUDY ON "DIGITAL SEQUENCE INFORMATION" ON GENETIC RESOURCES FOR FOOD AND AGRICULTURE

Note by the Secretariat

1. The Commission, at its last session, established a new work stream on "digital sequence information."[1] It requested the Secretariat to prepare, subject to the availability of the necessary resources, an exploratory fact-finding scoping study on "digital sequence information" on genetic resources for food and agriculture (GRFA) to provide information on, *inter alia*, terminology used in this area, actors involved with "digital sequence information" on GRFA, the types and extent of uses of "digital sequence information" on GRFA, such as:

- characterization,
- breeding and genetic improvement,
- conservation, and
- identification of GRFA

as well as on relevance of "digital sequence information" on GRFA for food security and nutrition, in order to facilitate consideration by the Commission, at its next session, of the implications of the use of "digital sequence information" on GRFA for the conservation

[1] The term is taken from decision CBD COP XIII/16 and is subject to further discussion. There is a recognition that there are a multiplicity of terms that have been used in this area (including, inter alia, "genetic sequence data", "genetic sequence information", "genetic information", "dematerialized genetic resources", "in silico utilization", etc.) and that further consideration is needed regarding the appropriate term or terms to be used.

and sustainable use of GRFA, including exchange, access and the fair and equitable sharing of the benefits arising from their use.[2]

2. The Commission requested its intergovernmental technical working groups and the Expert Group on Micro-organism and Invertebrate Genetic Resources for Food and Agriculture (Expert Group) to review and provide inputs to the draft exploratory fact-finding scoping study prior to its submission to the Commission, for consideration at its next session.[3]

3. The draft *Exploratory fact-finding scoping study on "digital sequence information" on genetic resources for food and agriculture* is contained in this document, for review and inputs by the Expert Group. The draft exploratory fact-finding scoping study been prepared by Jack A. Heinemann and Dorien S. Coray (School of Biological Sciences, University of Canterbury, Christchurch, New Zealand) and by David S. Thaler, Biozentrum, University of Basel, Basel, Switzerland. The content of the draft study is entirely the responsibility of the authors, and does not necessarily represent the views of FAO or its Members.

GUIDANCE SOUGHT

4. Experts are invited to review and provide inputs in writing to the draft exploratory fact-finding scoping study by **15 September 2018** to the Secretariat (cgrfa@fao.org).

[2] CGRFA-16/17/Report, paragraph 86.
[3] CGRFA-16/17/Report, paragraph 90.

Draft Exploratory Fact-Finding Scoping Study on "Digital Sequence Information" on Genetic Resources for Food and Agriculture

Jack A. Heinemann and Dorien S. Coray
School of Biological Sciences, University of Canterbury, Christchurch, New Zealand

David S. Thaler
Biozentrum, University of Basel, Basel, Switzerland

The content of this draft study is entirely the responsibility of the authors, and does not necessarily represent the views of FAO or its Members.

TABLE OF CONTENTS

LIST OF VIGNETTES

ABBREVIATIONS

CBD	Convention on Biological Diversity
CRISPR	clustered regularly interspaced short palindromic repeats
DSI	digital sequence information
DNA	deoxyribonucleic acid
EBI	European Bioinformatics Institute
EU	European Union
GIS/GLIS	Global Information System
GRFA	genetic resources for food and agriculture
GSD	genetic sequence data
GWAS	Genome-wide association study
ITPGRFA (Treaty)	International Treaty on Plant Genetic Resources for Food and Agriculture
mQTL	metabolomic quantitative trait locus/loci
MAS/B	marker assisted selection/breeding
MTA	material transfer agreement
ODM	oligonucleotide-directed mutagenesis
omics	Refers to terms that end in omic, such as genomics, epigenomics, transcriptomics, proteomics, metabolomics
QTL	quantitative trait locus/loci
RNA	ribonucleic acid
SNP	single nucleotide polymorphism

EXECUTIVE SUMMARY

> "The CBD and Nagoya Protocol contemplate regulation of the physical transfer of tangible genetic or biological material from a provider country to a user, pursuant to an ABS agreement. New technologies emerging from synthetic biology fundamentally change that paradigm, however. The genome of a particular species may now be sequenced within a provider country and that information may be transferred digitally to a company or research entity for downloading to a DNA synthesizer. As a result, synthetic biology technologies beg the question of whether ABS requirements should apply to the use of digital sequence information from genetic resources" (Manheim, 2016).

Study questions and key findings

This exploratory fact-finding scoping study examines how digital sequence information (DSI) on genetic resources for food and agriculture is currently being used, how it might be used in the future and what the implications of its use are and might be in the future for the food and agriculture sector.

All uses of DSI on genetic resources for food and agriculture (GRFA) that do or could affect GRFA or the value of GRFA were considered within scope of the study. This included use as food or in agriculture, but also using DSI to discover or add value to materials derived from GRFA (e.g. amyloid forming proteins, enzymes) or to add value to GRFA (e.g. identification of new traits, preservation of endangered populations, diagnosis of pathogens, food preservation) (Figure 1).

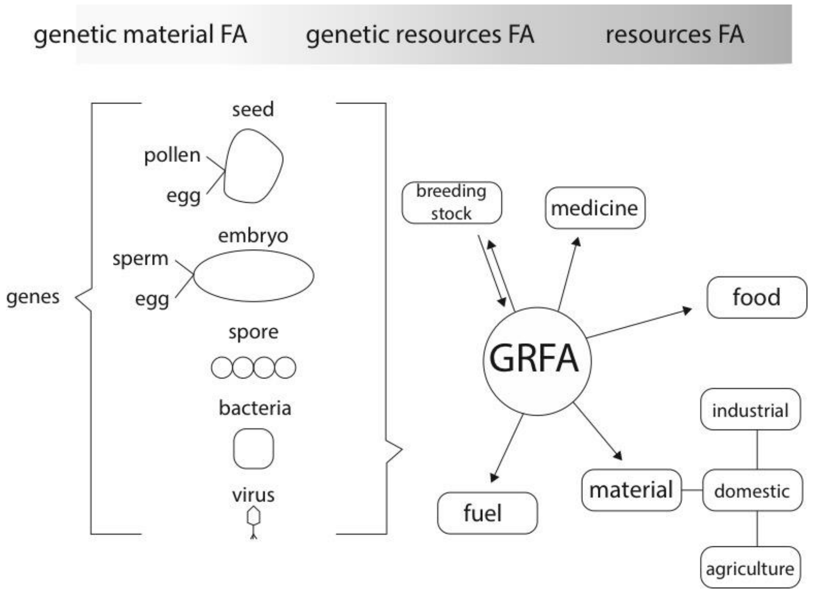


Figure 1. Uses of GRFA and associated genetic material (arrows indicate direction of resource use)

What is meant by DSI? (terminology in use)

- DSI on GRFA currently in use includes multiple kinds of information about various biological materials found in GRFA, used to manage GRFA, or to derive value from GRFA. Some
    - o but not all DSI on GRFA is DNA (or RNA) sequence information;
    - o is sufficient to synthesize a trait without needing to transfer biological genetic material;
    - o DSI on GRFA that is not DNA or RNA may;
        - ▪ be essential to identify or synthesize some traits;
        - ▪ not require DSI on DNA or RNA to identify or synthesize some traits.
- DSI on GRFA is not limited to DSI on organisms that are GRFA.

What are the characteristics of DSI and genetic material?

- DSI on GRFA contributes to food security and nutrition as a fundamental tool for characterization of GRFA and environments, selection and breeding, creation of new products, food safety and management of GRFA.
- DSI is an essential component of technologies used for the characterization, conservation and sustainable use of GRFA (e.g. DNA barcodes in biodiversity surveys). The roles of DSI are increasing with new technologies.
- DSI underpins technology for the synthesis of DNA and some kinds of genetic material. Such technologies could one day allow all kinds of genetic or biological material to be synthesized using DSI.[4]

What are the characteristics of technologies that use DSI, and what implications do these have for GRFA management?

- DSI makes it easier to get value from a genetic resource without possessing it or even its DNA. Actual or potential differences in the characteristics of technologies that use DSI did not vary significantly by subsector (animal, plant, microbe, forest, fishery).
- The study found that DSI was used extensively in all subsectors. Possibly differences in value will emerge where DSI is also useful for non-agricultural applications, such as drug and vaccine development, rather than by subsector application in general.

How is DSI stored, exchanged and shared and what implications does this have for GRFA management?

- DSI on GRFA is stored in electronic digital media. The amount of private DSI on GRFA is unknown. Publicly accessible DSI includes the content and functionality of approximately 1 700 online databases with infrastructure mainly in developed countries. Continuing funding in an open access model is not assured.

What role does DSI have in research, product development, regulation and intellectual property (IP) claims?

- DSI on GRFA is central to product development and intellectual property, and expected to increase in importance especially as DSI on more kinds of organisms becomes relevant to GRFA. DSI is used in regulations on food safety, product labeling, and correct identification of food components, which can be important for the conservation of threatened species. It is used to both diagnose diseases in plants and livestock and to design therapeutics to treat them.

What role does DSI have for farmers and the broader community? What implications do these have for GRFA management?

- Through DSI different kinds of industries and actors become involved with the characterization, conservation and sustainable use of GRFA.
    - Value chains are developing for everything from bionanotechnologies through synthetic biology and biological computing to on-farm hand-held sequencers and customized management advice.
    - The decreasing cost of sequencing and synthesis provide greater access to tools that can be used directly by the public and farmers, but also by researchers. This can be expected to increase the frequency and quantity of DSI transferred across national borders.

[4] The Study uncovered no scientific foundation for the dematerialization of information *per se*. That is, information of the kind described in this report may be transferred between different materials but does not exist separately from at least one physical form during storage, transmition or use. Information from a molecule (e.g. DNA) can be copied to other material such as computer-readable media, sound waves, electrical current or light (e.g. transmitted on a fibre cable). Whereas in these forms sequence information is literarally intangible, this quality also is not able to distinguish between information, genetic material or some kinds of GRFA because a molecule of DNA, a cell (e.g. the GRFA yeast *Saccharomyces cerevisiae*), and even seeds and animals below a certain size, cannot be perceived by human touch.

Introduction

Can DSI about genetic resources for food and agriculture (GRFA) increase the value of biological genetic material and in some cases substitute for possession of the biological genetic material of plants, animals and microbes for some uses in ways that it could not before? Can it create value separately from biological genetic material?

This report has been prepared to introduce a variety of audiences to the outcome of an exploratory scoping and fact-finding study on the uses of DSI on GRFA. It is important to emphasize that the study included GRFA as affected, modified or managed by DSI, but was not limited to DSI sourced from GRFA. The difference in approach places DSI rather than GRFA at the center of the study. Recognizing this emphasis on DSI is critical to understand that the value that might come from DSI *on* GRFA is sometimes different than the value of DSI *from* GRFA.

Importantly, the underlying study did not focus on biosafety or other kinds of risk because it was recognized that other reports were being prepared that included such perspectives, and the objective of this report was to scope what is and could be done with DSI rather than how DSI should or should not be used, or should or should not be regulated. Nevertheless, the study did find that DSI underpins powerful biotechnologies as well as resource inventories and conservation. Socio-economic implications and the potential to cause harm to human health or the environment are noted but not considered in depth.

The report covers existing and emerging uses of, mainly electronic, bio-information technology. These uses are already being applied to the characterization, conservation and sustainable use of GRFA, including equitable access or fair opportunity to benefit. The importance of DSI for these technologies is almost certain to increase. This report is intended to assist policy-makers in adapting to the scientific changes brought about by both quantitative and qualitative changes in information collection, transmission, applications and inherent value even when it may be uncoupled from the original biological genetic material.

GRFA include plants, animals, fungi and microorganisms that we eat (e.g. cassava, poultry, beer, yoghurt), or provide material and fuel (e.g. forest trees), industrial and consumer products (e.g. proteins) or provide genetic diversity (e.g. wild relatives), services (e.g. soil microbes) and resources (e.g. landraces) upon which agriculture depends (Figure 2). In the future, they even may provide computing power (The Economist, 2012). These resources are continuously being improved through activities such as breeding and management, or threatened by pests, pollution, disease or loss of genetic diversity.

Until recently, to use a genetic resource required possessing it. To move the benefits of a genetic resource to other places, the genetic resource itself had be transported (Figure 2), such as tomato seeds from plants that originated in South America to Europe and to North America. With the introduction of tissue culturing technologies, just the cells of some organisms were enough to reconstitute a genetic resource. As the basis for traits became associated with genes, and the material basis of those was linked to DNA, it became possible in some instances to transport just DNA, such as in the form of a plasmid with a gene from a bacterium intended for insertion into the genome of a plant (Zimkus and Ford, 2014).[5]

[5] It is acknowledged that in this passage a broader description of genetic resources is being used than as defined in the Treaty. However, it is included to illustrate that those working with GRFA, that is, in gene banks and government agricultural agencies, often refer to and treat DNA as a genetic resource. Tissue, cells and DNA as genetic resources are discussed here: http://www.nies.go.jp/biology/en/aboutus/facility/capsule.html. DNA as a genetic resource is discussed here: "There is a fifth type of genetic resource that differs from the other four types because it cannot be used to regenerate an organism and usually is not held in genebanks: **Cloned DNA sequences,** or genetic material from other organisms incorporated into crops by molecular techniques (for example, a gene from the bacteria *Bacillus thuringiensis* used for resistance to insects" (Heisey and Rubenstein, 2015).

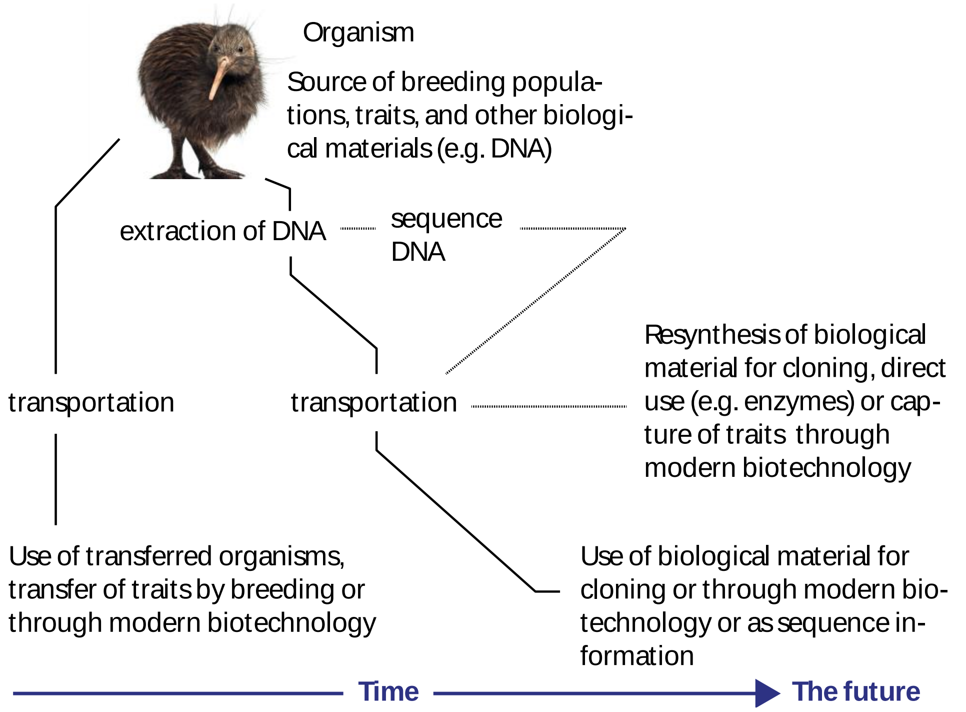

Organism
Source of breeding populations, traits, and other biological materials (e.g. DNA)
extraction of DNA
sequence DNA
Resynthesis of biological material for cloning, direct use (e.g. enzymes) or capture of traits through modern biotechnology
transportation
transportation
Use of transferred organisms, transfer of traits by breeding or through modern biotechnology
Use of biological material for cloning or through modern biotechnology or as sequence information
Time
The future

Figure 2. Mobility of GRFA

In even more specialized instances, the biological genetic material is no longer necessary to transport some traits, even organisms, of value. Although at this time examples are few, the proof-of-principle is established by them. Moreover, there is growing scientific interest and capacity to not only increase the number of examples, but also the kinds of traits and organisms that may be "transported" using only genetic information uncoupled from biological genetic material.

There are still other ways that genetic resources provide value separate from their biological genetic material. For example, in Figure 2 the genetic resource pictured is a kiwi bird, indigenous to New Zealand. At first, such an example seems out of place because by law the kiwi bird is not eaten. However, its unique genome is important to GRFA (Le Duc *et al.*, 2015) both for the genes for traits it carries, and for the genes it does not have. The kiwi bird genome can be used by comparative genomics to provide clues about what unknown genes in organisms that we do eat–both bird and mammal–may be responsible for traits that may be improved or eliminated. Value is extracted through comparative genomics using DSI and does not require possessing a kiwi bird.

To construct for the first time a version of the kiwi bird genome in the form of physical electronic media, that is, in the form used by a DNA database such as GenBank (GenBank), required access to the bird or tissues from the bird. But from as little as a single original sample of biological genetic material, the information could be stored, replicated and transmitted separately as electronic media, and the value of it was obtained without using the kiwi bird biological material. In fact, only in its electronic form could the genome be used for a comprehensive comparative genomics analysis.

DSI on GRFA is not limited to DSI on organisms that are GRFA. However, neither are genes. For example, prior to the development of genetic engineering, the bacterium *Bacillus thuringiensis* would not have been food any more than a kiwi bird. However, it is clear that the genes for its insecticidal proteins have been an important source of traits for some staple crops such as cotton, maize and soybeans (Heisey and Rubenstein, 2015). More than just a microbe that provides ecosystem services to these crops, it has become part of the genetic constitution of GRFA.

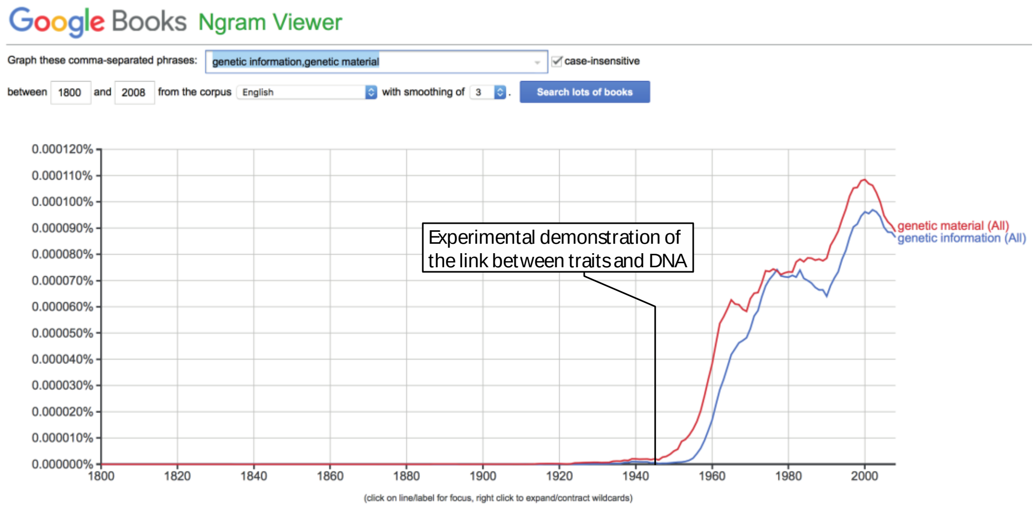
The ability of DSI to expand the range of organisms that become resources for food or agriculture is not special to this technology. However, as DSI makes it possible to recruit traits from non-GRFA organisms, or to use genomes of non-GRFA organisms to identify genes in organisms that are used in agriculture, it extends the boundaries of what may be used as GRFA.

Terminology

A number of different terms are discussed in this report, collectively referred to as “digital sequence information” (DSI) in accordance with decision XIII/16 of the Convention on Biological Diversity (CBD) and the decision of the Commission on Genetic Resources for Food and Agriculture (CGRFA) to establish a new work stream on DSI. They include, *inter alia*, genetic sequence data, genetic sequence information, genetic information, dematerialized genetic resources, and *in silico* utilization (paragraph 86 CGRFA, 2017).

It is not assumed that everyone using the various terms associated with DSI means the same things by them.[6] Therefore, the study examined how the terms are used by different actors. This exercise is useful for providing context. It also raises awareness of when different users may be talking past one another without realizing that they are talking about similar things, or indeed are using similar words but with a very different understanding of them.

This technical scoping study drew primarily from the scientific and related literature but also from other sources such as interviews with scientists and members of civil society. As a consequence, DSI is used in general to refer to that which the scientists and technicians creating and using DSI generally think it applies. This “at the coal face” definition has obvious but unavoidable limitations. First, the science is constantly evolving, making more precise definitions presently impossible. Second, the scientific community notably does not use the term DSI



Google Books Ngram Viewer

Graph these comma-separated phrases: genetic information,genetic material ☑ case-insensitive

between 1800 and 2008 from the corpus English with smoothing of 3 .

Figure 3. Use of the terms “genetic material” and “genetic information”.
Note: The terms “digital sequence information” and “genetic sequence data” returned no results. Note increase around the time of the first experiments showing that genes were made of DNA.
Source: (Avery *et al.*, 1944).

[6] “[T]he primary problem in any legislative document that uses technical terminology – determining how that term is understood, and by whom. Where the document’s target audience is technical (technical experts and regulators) then the term will be understood and applied technically. Regulations will be developed citing particular scientific texts or statistics that will give a precise definition of the element and a standard for its measurement. Where the target audience is not highly technical, however, use of a technical term can be problematic, since it may mean different things to different audiences” (Tvedt and Young, 2007).

(Figure 3), so some judgement must be exercised in linking scientific usage and that of the CBD and CGRFA.

Notwithstanding those limitations, the use of operational examples provides a strong sense of the scope of real things referred to by the terminology associated with DSI. Over 1700 database repositories of information are in routine use as sources of DSI (Various, 2017). A few are specialized in storing information about nucleic acids, such as DNA and RNA. Even these store information about, or convert DNA information into, amino acid sequences of proteins. Other types of databases specialize in storing information about differences in sequences. For example, single nucleotide polymorphism (SNP) databases reveal differences in versions of the same gene or genomes of the same species. This variation can be the basis for desirable traits.

An illustrative definition of DSI is used for the purposes of this report. DSI refers to the kind of information in or that might be added to databases of the kind currently in use and collated by the scientific journal *Nucleic Acids Research* (NAR).

GRFA is at least *both* material of biological origin and functional units of heredity, as it is defined in the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) to mean "any material of plant origin, including reproductive and vegetative propagating material, containing functional units of heredity"(ITPGRFA). The material in a genetic resource includes *inter alia* RNA, proteins, metabolites and the describable interactions between molecules that may be genetically determined, environmentally induced, or a combination of the two (Figure 4).

Some of these other molecules may also be genes. While it has long been known that the material form of a gene is a nucleic acid (usually DNA), there are other kinds of molecules that also are essential for the propagation of some traits (Heinemann, 2004; Strohman, 1997). This ever-growing number of materials that appear indispensable for the inheritance of some traits are called epigenes.

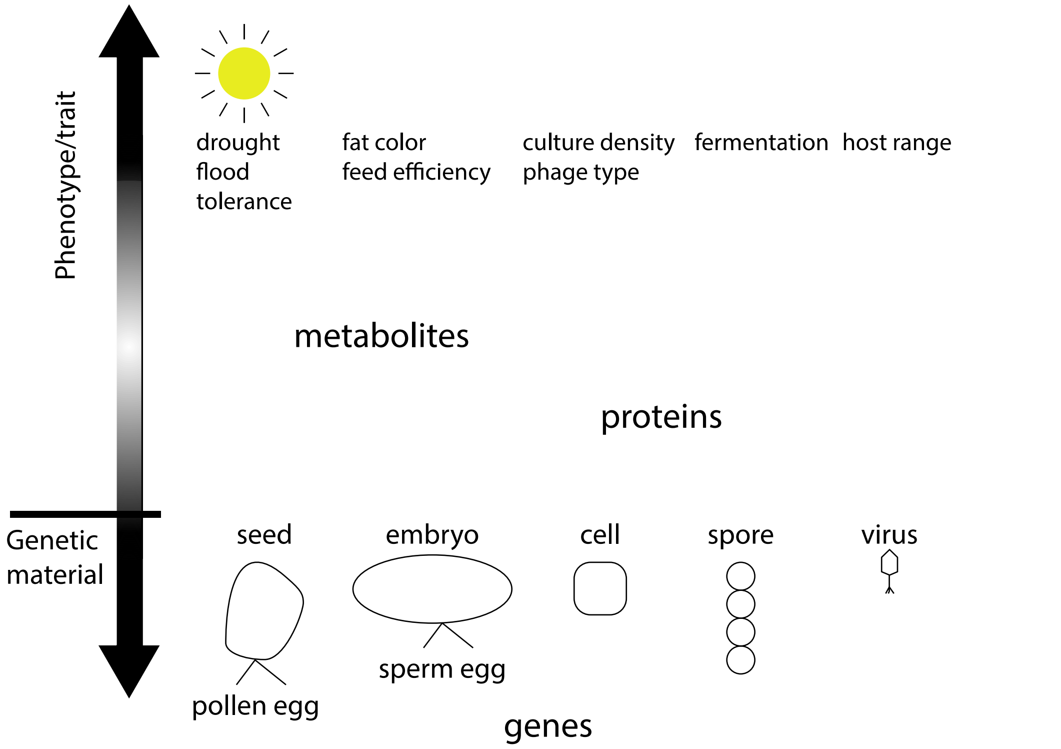


Figure 4. GRFA is both material of biological origin and functional units of heredity

The context of genetic information can be critical for trait propagation, separate from the DNA of a genome. In a significant number of cases DNA sequence is necessary but not sufficient to establish a hereditary characteristic. Examples include epigenetic information as elaborated in the technical report.

In analogy with the "genetic code" as carried by DNA, traits for which at least one other kind of non-nucleic acid molecule is required for propagation are likened to possessing codes, such as the metabolic, histone, sugar, splicing, tubulin, signaling, ubiquitin and glycomic codes (Barbieri, 2018; Buckeridge, 2018; Gabius, 2018; Marijuán *et al.*, 2018; Prakash and Fournier, 2018). They are studied in an emerging area called Code Biology (Barbieri, 2018).

How DSI on GRFA is used by scientists now portents a time when DNA itself may not be center stage, or at least must share the stage with many more forms of DSI. Knowing this helps to explain why there are so many kinds of databases of DSI on GRFA, and how each contributes essential information necessary to reconstruct a range of different traits.

In short, DNA and RNA sequences alone are sometimes sufficient and sometimes not to identify a trait; some traits would be impossible to transfer using only DNA. A trait can sometimes even be identified using sequence information from molecules other than DNA. Therefore, to restrict consideration of DSI on GRFA to DNA databases would neither be true to how scientists are already using DSI on GRFA nor fully inform future-ready policy.

Beyond these practicalities of modern genetics is the technology to synthesize biological and biological genetic material using neither the original source organism nor any biological material from it. Complete genome synthesis has been achieved for types of bacteria (Gibson, 2014). Complete synthesis of a few viruses has already been demonstrated (Wimmer, 2006). A new machine in prototype stage may one day achieve this for cells, possibly plant and animal zygotes (Boles *et al.*, 2017). The synthesis of engineered and novel organisms uses DSI of DNA, proteins and possibly other molecules along with subunits of these molecules that may be produced in a laboratory as inputs. Some genetic materials can be constructed using DSI and chemistry, separate from biological genetic material (Figure 5).

In summary, it is possible to identify common elements in the various terms that have been collected under the heading DSI. However, this does not mean that those using different terms are necessarily talking about the same thing, or that the meaning of terms is static. Different actors can be drawing upon discipline conventions that do not perfectly overlap with other disciplines, such as illustrated by the different history of use of the term genetic material by classical applied animal and plant breeders compared to molecular biologists in the middle of the last century.

Operational adoption of DSI appears to be mostly unaffected by terminological disputes. From the viewpoint of how DSI is being used on GRFA and the management of GRFA, all information that can be gathered into databases is being used. No database is superior to all others in every application of DSI to GRFA. DNA databases are linked to genomic, transcriptomic and proteomic databases (and sometimes these are features of a single database). Metabolomic databases sometimes can be more useful for understanding a GRFA than a corresponding DNA database. The power of the databases is increased because they link phenotypic and metabolic traits to DNA genes, or epigenes.

It is beyond the scope of this study to resolve discontinuities in terminology adopted by different users. Like the definition of the gene itself, sometimes an example (i.e. DNA) is more practical for those working in the discipline than is a more flexible heuristic (i.e. the material basis of heritable traits). For the purpose of communicating the findings of the report on DSI, the term is tied to the operational descriptions of actors that map with reasonable consistency to what is in and what could be in the future stored in databases of the type described by *Nucleic Acids Research* (NAR), and used by the science of bioinformatics.

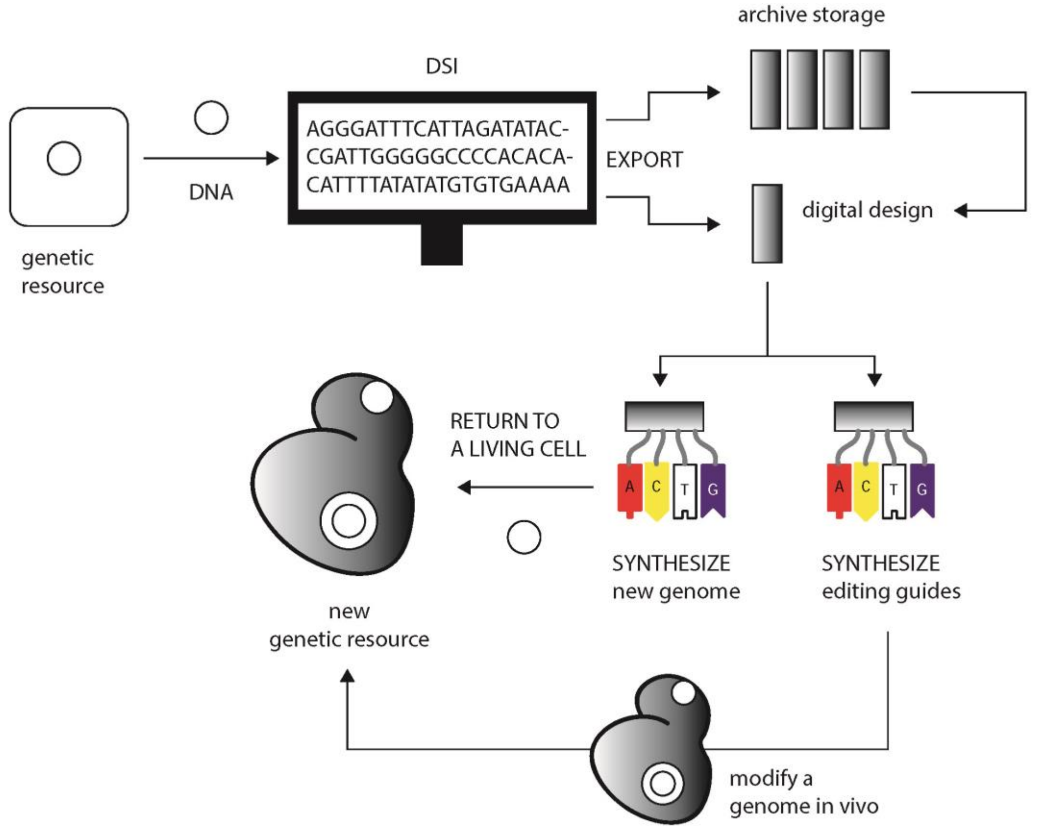


Figure 5. Moving information between digital and living worlds
Note: From top left, a cell or organism is a genetic resource from which material such as genomic DNA (circle) is extracted. The order of bases in the genome is determined by sequencing and the information is stored as a digital sequence. The information may or may not be further modified and a new genome written, or a genome edited, using the digital information. The synthesized DNA is transferred to a living organism (e.g. yeast) for final assembly of large pieces and may be further transferred to other cells to create a new genetic resource.
Source: (Gibson, 2014).

DSI carries information, as do genes

> "Any biology textbook published within the last sixty years states as a fact that genes encode and transmit information. This statement derives from the mid-1940s, when biologists started to equate the activity of genes with a code that transmits instructions and governs all processes and features of living organisms, from respiration to eye color" (García-Sancho, 2015).

DSI has qualities that make it as different to text as is e-mail to a handwritten letter. While at the interface a handwritten letter consisting of a string of letters appears the same as those letters displayed on a computer screen, the latter is supported by a large variety of unseen but essential information. Without the accompanying algorithms, schemata and metadata, the string of letters would be meaningless even to the computer.

It is tempting to imagine that a DNA sequencing machine "sees" the nucleotides in a molecule of DNA and lists them off in order such as TTATGCCAT. However, in reality the machine generates a host of data that is transformed by sophisticated processes governed by underlying software that automates decisions about the chemical structures in the molecule. That is why each section of a genome is sequenced many times (called "depth of read") to increase the confidence in having identified the correct nucleotide at each location.

To have meaning in relation to a trait, the string of letters of a DNA sequence must be read in a particular direction and that information is part of the explicit or implicit schemata of a database. The trait may be dependent on events that cannot be easily or fully predicted just from the DNA sequence, such as the number of post-transcriptional and post-translation modifications that occur between gene and trait.

Beyond the actual sequences of DNA and RNA, additional DSI can take the form of counting the number of times a sequence occurs in a sample or a genome (called copy number variation or CNV). These are large gains or losses of genomic DNA seen in comparisons of genomes of organisms in the same species. Along with other kinds of markers (e.g. SNPs), CNVs contribute to estimates of genetic variability and can explain significant differences in phenotypes.[7]

In transcriptomics the copy number for each sequence forms the basis for gene expression analysis and linking the contribution of RNAs to a quantitative trait. The number of copies of each amino acid sequence yields analogous data for proteins (Figure 6).

It is these and other elements of information that enable DSI to faithfully encode GRFA information. Transferability of information between media makes the different kinds of media interchangeable for storing the information, allowing either loss-less transfer or transfer with known amounts of loss between media.

These principles underlie many common and familiar technologies, such as telephones. The human ear receives sound as waves propagated in the medium of air. Telephones do not transmit sound waves directly; telephone handsets transduce the sound waves as electronic signals and encode them for transmission. These signals are then transmitted through the telephone network in various forms such as electrical or optical (light) signals plus data about the call.
The recipient's telephone then decodes and transduces these signals to create sound waves perceivable by the human ear. All of this is enabled by the use of structured formats and conventions.

DSI is highly structured using formats and conventions with attributes beyond simple and minimally structured text files. As with the technology supporting the layers of information necessary for a conversation over a telephone line, a database has content, schemata and meta-data, making it very different from a sequence of symbols as may be written on a piece of paper (Figure 6).[8] The database has organization that elevates the content from data to information. Schemata make it possible to navigate and act on the content of the database. Various forms of meta-data, descriptive, structural and adminstrative, provide context.

[7] "In farm animals, several traits are caused by CNV affecting genes or gene regions. For example, the *Dominant white* locus in pigs includes alleles determined by duplications of the KIT gene. CNV also affects the *Agouti* locus in sheep and goats and contributes to the variability of coat colour in these two species. CNV in intron 1 of the *SOX5* gene causes the pea-comb phenotype in chicken and the late feathering locus in this avian species includes a partial duplication of the *PRLR* and *SPEF2* genes" (Fontanesi *et al.*, 2010).

[8] This is not to say that paper would be insufficient as a medium to achieve the sophistication of some kind of 'electronic' database. The comparison being drawn is to a series of letters on a piece of paper and what is required for the same utility being routinely provided now by electronic databases.

Information field	Entries	Information field	Entries
Enzyme nomenclature		Functional parameters	
EC number	3 869	K_m value	28 134
Recommended name	3 509	Turnover number	3 986
Systematic name	3 182	Specific activity	11 787
Synonyms	17 707	pH optimum	14 037
CAS registry number	3 552	pH range	3 929
Reaction	3 518	Temperature optimum	6 147
Reaction type	4 123	Temperature range	908
Enzyme structure		Molecular properties	
Molecular weight	12 329	pH stability	2 931
Subunits	7 416	Temperature stability	6 825
Sequence links	33 099	General stability	5 398
Post-translational modification	1 112	Organic solvent stability	311
Crystallization	1 003	Oxidation stability	349
3D-structure, specific PDB links	6 142	Storage stability	6 505
Enzyme–ligand interactions		Purification	11 176
Substrates/products	47 630	Cloned	2 015
Natural substrate	7 668	Engineering	797
Cofactor	6 217	Renatured	199
Activating compound	6 217	Application	338
Metals/ions	13 173	Organism-related information	
Inhibitors	56 336	Organism	40 027
Bibliographical data		Source tissue, organ	19 347
References	46 305	Localization	7 935

Figure 6. A peek at the meta-data of BRENDA
Note: “The database covers organism-specific information on functional and molecular properties, in detail on the nomenclature, reaction and specificity, enzyme structure, stability, application and engineering, organism, ligands, literature references and links to other databases” (Schomburg *et al.*, 2002). It can be used to calculate or simulate metabolic pathways using “the information of substrate/product chains and the corresponding kinetic data of the preceding and following enzymes” (Schomburg *et al.*, 2002).

In Table 1A, the Google Translate[9] tool was given a complete sentence and asked to translate that sentence from English into another language. The translation was substituted for the original English text for translation back to English. Fidelity of translation from English to either Chinese, Russian or Arabic was high as apparent from the high fidelity of the back translation.

Compare for example, the difference in translation of the sentence (with content, schemata and meta-data when sent to Google Translate) and the same words translated out of context as just a series of words, even when in the same order as the original sentence. If each word of the same sentence is translated independently and then assembled into a phrase, the accuracy of the translation declines, as shown in the last row of Table 1B. This use of the tool Google Translate

[9] https://translate.google.com

strips it of important contextual information and rules, the context of language that is encoded in schemata and meta-data.

Table 1A. Characteristics of information interchangeability illustrated by translation into different UN working languages.

from*	to	back
Fidelity of translation from English to either Chinese, Russian or Arabic was high as apparent from the high fidelity of the back translation.	翻譯從英文翻譯成中文，俄文或阿拉伯文的忠誠度高，這從後面翻譯的高保真度中可以看出。	The loyalty of translations from English to Chinese, Russian or Arabic is high, as can be seen from the high fidelity of later translations.
Fidelity of translation from English to either Chinese, Russian or Arabic was high as apparent from the high fidelity of the back translation.	Верность перевода с английского на китайский, русский или арабский была высокой, как видно из высокой верности перевода назад.	The faithfulness of the translation from English into Chinese, Russian or Arabic was high, as can be seen from the high fidelity of the translation back.
Fidelity of translation from English to either Chinese, Russian or Arabic was high as apparent from the high fidelity of the back translation.	كانت دقة الترجمة من الإنجليزية إلى الصينية أو الروسية أو العربية عالية كما يتضح من الدقة العالية للترجمة الخلفية.	The accuracy of translation from English to Chinese, Russian or Arabic was high, as demonstrated by the high resolution of background translation.

Table 1B. The difference between a sequence of words and a sentence.

from**	to (word by word translation)	back
Interchangeability of information allows it to be transferred between media without or with minimal loss.	互換性 的 信息 允許 它 至 是 轉入 之間 媒體 無 要么 同 最小 失利	The interchangeability of information allows it to be transferred to the media without having the same minimum loss
fidelity of translation from english to either chinese russian or arabic was high as apparent from the high fidelity of the back translation	保真度 的 翻譯 從 英語 至 或 中文 俄語 要么 阿拉伯 是 高 如 明顯的 從 該 高 保真度 的 該 背部 翻譯	Fidelity of translation From English to or Chinese Russian Either Arab Yes high Such as obviously From This high Fidelity of This Back translation

* Translations provided in series left to right by Google Translate. The order was top to bottom English to Chinese, English to Russian, English to Arabic.

** Translation of each word from English to Chinese provided by Google Translate, then the entire phrase was translated into English.

DNA sequence information can be transferred to DSI and then again transferred back through synthesis of a DNA molecule (Figure 5). This characteristic is shared by the kinds of databases chosen to illustrate DSI, all of which contain DSI on GRFA. A protein’s amino acid sequence can be transferred to DSI and then again transferred back through synthesis of a protein molecule, or the DSI of the amino acid sequence itself can be converted into a DNA sequence for the synthesis of a DNA molecule that when placed back into a cell produces the protein molecule.

DSI in the form of genome sequences can be used to predict transcriptome sequences, which in turn can be used to predict proteome composition and this can be used to predict the metabolome. As importantly, a description of the metabolome would allow the design of a proteome that could produce that metabolome and then a genome that could produce that proteome. In other words, the knowledge of many databases may soon be sufficient to recreate traits or material of value without necessarily producing copies of the source biological genetic material.

While all the above can be done, it is neither intended nor implied that doing so is trivial. In some cases it is routine while in others it would be at the cutting edge of, or just beyond, current technology. The degree of difficulty determines the pace by which such things are done, how well the outcome matches the prediction, but not the possibility of them being done from a technical point of view.

In summary, DSI is stimulating discussions of what genetic material can be, similarly to how artificial intelligence is challenging concepts of the brain. Is the biological material of neural tissue the only medium in which complex thinking can occur or be stored? Would a computer have to be a copy of a human being in order to have “thoughts” of some value to humanity? Does genetic material need to be biological material? In a biological system the answer appears obvious. However, as a storage medium for genetic information, a sequence can be held by either a biological material or the kind of material used by a computer. As DSI makes it possible to store and transmit genetic information back to biological material, it challenges the boundaries of what may be used as GRFA.

Use of DSI in characterization, conservation and sustainable use of GRFA

DSI on GRFA is applied to describe and characterize genetic resources and contribute to their conservation and sustainable use. Characterization provides information about organisms and ecosystems that may be used for basic research, identification and monitoring and to draw comparisons with other organisms and ecosystems. Molecular markers, phenomics and bioinformatics are used in the characterization of GRFA (Lidder and Sonnino, 2011).

Molecular markers are used to identify and track alleles for quantitative traits. When different combinations of alleles of different genes cause different phenotypes, the trait is said to be quantitative. DSI is used to create the tools for following different alleles through breeding.

Barcoding is a technology that improves identification, particularly of species that are difficult to distinguish between by morphology. Barcoding uses DNA sequences from a gene or genomic location for species identification and discovery (Lidder and Sonnino, 2011). A variation of barcoding is metabarcoding, where an estimate of biological diversity is derived from direct isolation of DNA from a sample, such as a soil sample (Orwin *et al.*, 2018).

DSI has become a critical component of breeding by contributing to the rate and accuracy of trait mapping to develop targets for marker-assisted selection (MAS). Moreover, it is important for calculating genomic estimated breeding value (Varshney *et al.*, 2014).

DSI also contributes to efforts to conserve GRFA. The means to identify individuals with defined genetic variants associated with particular traits informs estimates of the population’s effective size (Ne) which is needed to ensure that there are sufficient numbers for breeding (Lidder and Sonnino, 2011). The information may be used to manage both domesticated and wild GRFA. Likewise, for some ecosystem functions, it is not enough to have sufficient numbers of one genetic type, and instead it depends on having both the right mix of variants or species (qualitative) and in the right ratios (quantitative). For example, the rumen microbiome varies with feedstock. The variance is not in what microorganisms are in the rumen, but the ratio of the different species to one another (Henderson *et al.*, 2015).

Likewise, sustainable use of GRFA depends on maintaining sufficient genetic diversity to ensure that species or subspecies do not go extinct, that diseases and pests can be managed, and that the impact of wild and farmed populations on the environment can be mitigated. DSI is developed to track the movement of genes between domesticated and wild populations of GRFA. For example, movement from domesticated to wild populations might endanger small wild populations if it caused outbreeding depression (Lidder and Sonnino, 2011). Monitoring gene flow is thus also important for conservation efforts.

Sustainable use of GRFA also depends on environmental factors. For example, most crop plants require pollinators such as bees. Because of decreasing bee populations, use of commercially supplied species has been increasing in some places (Suni *et al.*, 2017). Are these commercial releases effective? How do they impact wild bee populations and pollination? For example, "Managed species are commonly reared or transported under conditions that facilitate disease transmission, and both managed bumblebees and honey bees have a list of serious pests and pathogens that can 'spillover' into natural populations" (Lozier and Zayed, 2017).

Bees from 22 sites in Massachusetts that varied in their intensity of commercial bee use were genotyped using DSI on nine microsatellite loci to determine if commercial releases of certain bee species might impact populations of wild bees, or pollination of crops. The research found that despite the commercial releases, wild bees were found on cranberry crops more often. The research suggests that management of bees for services to pollination should include a determination of whether commercial bees or bumble bee habitat surrounding agricultural fields, intended to promote wild bee populations, would be the more effective option at lowest risk to bee health (Suni *et al.*, 2017).

Furthermore, sustainable use has economic dimensions. Farmers need adequate returns from their products. Certification and traceability add value to agriculture by meeting consumer demands. DSI is used to meet certification and traceability standards, as well as to ensure compliance with some food safety standards.

DSI is used to identify and diagnose pests and disease (both genetic and infectious), but also to discover and design new pesticides and therapeutics, such as antibiotics and vaccines.

Sequencing

DSI is generated from organisms and ecosystems by extracting particular molecules and where appropriate "sequencing" them. For example in the case of DNA, that is to determine the order of subunits in the molecule (Oldham, 2009). DNA is a polymer of nucleotides abbreviated as A, T, G and C. The actual sequence of these four subunits defines the individual molecule of DNA. Similarly, a protein is a polymer of amino acids (of which there are 20 different kinds in use in most organisms). The actual sequence of the amino acids in a protein defines the individual protein molecule.

Sequencing capacity has grown exponentially. As many as 2.5 million plant and animal species may have complete genome sequences, and between 100 million and 2 billion people may have their genomes sequenced, by 2025 (Stephens *et al.*, 2015). The rate of growth in this capacity, already possibly doubling once every seven months, foreshadows a future where all genomes may one day be described and variations in them recorded in near real time, globally.

Sequencing capacity is also becoming more distributed. This is due to both a decrease in costs and miniaturization. Devices that fit in the palm of a person's hand and which plug into the USB port of a laptop are now available (Jain *et al.*, 2018; Yong, 2016)). Devices the size of a credit card, which can be used to sequence the DNA in saliva or blood samples, are being developed to diagnose the most common microbial pathogens in animal GRFA (Chen *et al.*, 2017). Coupling the card with the camera on a smartphone provides real time diagnostics for veterinarians and farmers. In principle, these tools could one day be used on plant and soil samples.

Reconstruction

Molecules can be synthesized directly using DSI. Synthesizing DNA or RNA molecules based on a particular desired sequence has been done since the early 1970s. However, those were very expensive molecules and also very small. It is now possible with great accuracy to synthesize gene-sized DNA molecules (1 000+ nucleotides) at a cost of about USD0.15 per base pair (Petrone, 2016).

Both sequencing and reconstruction contribute to food security and nutrition through use in modern breeding programmes. An example is the use of DSI in MAS applied through SNP (pronounced "snip") chips (Box 1).[10] For MAS, DSI in the form of SNP databases can be used to screen germplasm to concentrate genetic material with a high frequency of desired genotypes ("phenotype to genotype").[11] Screened and selected stocks are preferentially used in breeding (Weller *et al.,* 2017).

Box 1. How is DSI used in a SNP chip?

A SNP chip is usually made from glass coupons upon which are attached many fragments (oligonucleotides) of DNA. High-density chips can have DNA fragments for hundreds of thousands of different oligonucleotides, each associated with a different genetic marker. All of these oligonucleotide sequences and then molecules were derived from DSI.

Genomic DNA from an organism is compared to the DNA on the chip to identify SNPs. Oligonucleotides corresponding to DNA sequences in genes of interest are synthesized. These oligonucleotides differ at one nucleotide, representing different SNPs. The DNA derived from the genome of an organism to be genotyped is applied to the chip. DNA sequences that are able to bind (called hybridize) are perfectly complementary to an oligonucleotide on the chip, as if they were two strands of a double-stranded DNA molecule.

The hybridization pattern on the chip is visible to sensitive electronic detectors, creating a digital profile of the SNPs in the genome being examined. Each marker is assigned a particular sequence in the genome being tested, based on the particular variable nucleotide that was represented at that place on the chip. That profile is converted into a DSI database of SNPs.

SNP information is among the kinds of genotype and phenotype information that is held by databases. SNPs are a type of molecular marker. Among others that are in use to varying degrees include RFLP (restriction fragment length polymorphisms), RAPD (random amplification of polymorphic DNA), AFLP (amplified fragment length polymorphisms) (for a review, see Yang *et al.*, 2013)). SNP-based tools are becoming the most common for genotyping because SNPs are large in number and well distributed in genomes, have proven stable and accurate, and are adaptable to high-throughput analyses. Although SNP chips are a significant improvement in genotyping, MAS can still take many generations in some kinds of organisms to achieve the desired outcome (He *et al.,* 2014; Weller *et al.,* 2017)).

As the information value (and not just the scale of content) of DSI grows and the ease and rate of collection increases, "genotype to phenotype" approaches to breeding are being developed. Marker identification can be accelerated using genotype-by-sequencing (GBS) techniques that

[10] For a user-friendly and brief introduction, see https://www.mun.ca/biology/scarr/DNA_Chips.html.

[11] The phenotype to genotype approach associates phenotypes with genetic markers that are converted to DNA sequence information and used to screen offspring with desired genotypes for further breeding. In the emerging genotype to phenotype approach, genomes are sequenced and analysed for predicted phenotypes, or they are synthesized based on predicted phenotypes, selecting genomes that match.

use whole genome sequencing for genome wide association studies (GWAS) in GRFA (He *et al.*, 2014).

GWAS mapping, also called linkage disequilibrium (LD) mapping, identifies statistically significant phenotype-genotype quantitative trait loci associations (Varshney *et al.*, 2014). Originally developed for high resolution GWAS in maize, GBS has been applied to multiple plant species with complex genomes, such as wheat, barley, rice, potato, cassava, apples and pine (He *et al.*, 2014; Norelli *et al.*, 2017; Pan *et al.*, 2015). Use in livestock animals such as cattle and chickens is also advancing (Brouard *et al.*, 2017; Wang *et al.*, 2017b). The key distinction between historical applications of MAS and GBS is that the latter allows "breeders to implement GWAS, genomic diversity study, genetic linkage analysis, molecular marker discovery, and genomic selection (GS) under a large scale of plant breeding programs. *There is no requirement for a priori knowledge of the species genomes* as the GBS method has been shown to be robust across a range of species and SNP discovery and genotyping are completed together" (emphasis added to He *et al.*, 2014).

GBS techniques can in certain cases increase the efficiency of breeding. The technique can be customized to local breeds. It has been easiest to apply to plants and more challenging for livestock, because of the different levels of homozygosity (Brouard *et al.*, 2017). The much higher density of identified SNPs potentially decreases the size of adult populations necessary to make associations between SNPs and traits (Brouard *et al.*, 2017; Ibeagha-Awemu *et al.*, 2016). That makes it applicable both for breeding GRFA and for conservation of wild species, where it can be used to more accurately calculate effective population size (e.g. in salmon as described by Larson *et al.*, 2014).

The capacity to reconstruct molecules other than nucleic acids is also increasing. Machines such as the recently described "digital-to-biological" converter build both DNA and protein molecules and they have already built a virus from scratch. The converter "receives digitally transmitted DNA sequence information and converts it into biopolymers, such as DNA, RNA and proteins, as well as complex entities such as viral particles, without any human intervention" (Boles *et al.*, 2017). Such machines are in prototype stage.

Editing

Gene/genome editing is more likely for the foreseeable future to be the way to modify traits than synthesizing organisms *de novo*. Editing alters DNA sequences *in vivo*. Hence, DNA does not have to be removed from the organism, manipulated and then returned to it. For some applications, a few changes to the genome may be sufficient. Where more than this is required, editing can be applied multiple times to the same genome, or be used to replace large sections of DNA.

Several advancing technologies including editing are dependent on DSI. These include the suite of technologies referred to by various names, including the new breeding techniques, gene drive systems, oligonucleotide-directed mutagenesis (ODM) and RNA/DNA interference (also known as gene silencing).

Each of these techniques relies upon sequence knowledge of a "target" DNA or RNA molecule. In most cases, the techniques also require a co-factor composed of an oligonucleotide molecule. The sequence of that nucleic acid molecule is also determined by reference to sequences in databases.

Gene editing techniques can be used to alter a genome at multiple locations, introducing gene variations that combine to create complex traits. There is a growing number of applications in GRFA (Lamas-Toranzo *et al.*, 2017). Gene editing techniques can be used to generate genetic variation later screened for desirable phenotypes or, once a desirable variant of gene (called an allele) has been identified, that variant can be created in another germplasm (Box 2). Doing so has the potential to avoid the so-called drag that comes from breeding two different parents that each have a different mix of desirable and also undesirable alleles at many different locations. In addition, the desired allele can be identified in any species or stock that shares the equivalent gene.

The value may be realized by the owners of the germplasm or those that use the germplasm (e.g. farmers) of the engineered organism. However, the mobility of the DSI allows others to derive value without being connected to the biological genetic material. These include, for example, the bioinformaticians who identify the valuable DSI or those who sell the materials needed for gene editing.

Box 2. Recent applications of gene editing in GRFA.

Tomatoes

The *cis*-regulatory elements of genes are important for gene expression but are not sequences that vary the protein product of a gene. Variations in expression level have been found in both quantitative trail loci (QTL) and genome-wide association (GWAS) studies (both of which are dependent on DSI) to be a significant source of phenotypic diversity.

Researchers developed a CRISPR/Cas9 gene drive system in tomatoes "to rapidly and efficiently generate dozens of novel cis-regulatory alleles for three genes that regulate fruit size, inflorescence architecture, and plant growth habit" (Rodríguez-Leal *et al.*, 2017). The components of the gene editing CRISPR/Cas9 system were designed from DSI and introduced as transgenes into the tomato genome. A novel genetic system that ensured stable inheritance of the transgenes was devised to increase the number of changes that could accumulate. Offspring with desired phenotypes were isolated over multiple generations.

Sheep

The *BCO2* gene of Norwegian sheep is important for determining the trait "yellow fat". Mutations in this gene prevent production of an enzyme that converts β-carotene into retinaldehyd. The accumulation of β-carotene is often desirable because it can be converted by people into Vitamin A. Fat with accumulations of β-carotene is characteristically yellow in color.

The gene editing tool CRISPR/Cas9 guided by DSI of the *BCO2* gene was used on the one cell stage of Tan sheep zygotes creating either biallelic or monoallelic *BCO2* disrupted modified animals (Niu, 2017). Biallelic animals (those with both *BCO2* genes mutated) had significantly more yellow color in fat.

Dematerialized, or *in silico*, use

The volume of DSI being generated by whole genome sequences is already on the scale of "big data". Some DSI is highly structured, such as that describing genomes. Other DSI is unstructured, such as graphics. The science of computational biology using the tools of bioinformatics provides a methodological approach to unifying these variant kinds of information, further demonstrating the interchangeability of information in various forms.

Some uses of DSI may again be applied to genetic resources directly, and some not. For example, DSI adds value to GRFA by helping breeders and farmers to more efficiently select breeding stock (Figure 7). Historically genes and variants of genes have been retrospectively associated with particular traits (phenotypes) and then the DSI of those associations was used to screen out of the breeding programme the germplasm with the fewest DNA markers for the mix of desired traits or with undesirable genes and alleles of genes (Woolliams and Oldenbroek, 2017)).

The DSI can have value beyond this use too. For example, the same databases can be applied to management of wild GRFA or to assess the diversity in commercial livestock breeds or crop cultivars (Yang *et al.*, 2013).

DSI on GRFA does not have to only add value to GRFA. It can add value in other ways. It can

- increase the value of products when used to screen foods for pathogens or microbes associated with safety (Higgins *et al.*, 2018; Rantsiou *et al.*, in press);
- increase the value of products through label certification, verify label ingredients or detect substitutions of endangered species in food products (Di Pinto *et al.*, 2015);
- be used to detect or diagnose both disease and infectious diseases in plants and animals, or as an antibiotic resistance surveillance tool (European Food Safety *et al.*, 2017);

- be applied to characterizing ecosystem function or conservation of species or sub-species through techniques such as barcoding or meta-barcoding (Orwin *et al.*, 2018);
- be used to design DNA or RNA-based pesticides or therapeutics (Heinemann *et al.*, 2013; Mogren and Lundgren, 2017).

With the miniaturization of collection tools, such as hand-held DNA sequencers and credit card-sized, real-time diagnostic devices, it is likely that more people will be contributing to the creation of DSI. At such scales, it may be possible to create value chains from personalized advertising and direct-to-the-consumer sales as is happening with human genome sequencing (Contreras and Deshmukh, 2017).

As the cost of synthesis declines, and should tools for synthesis, such as the prototype digital-to-biological converter become available, then DSI may add value through production of biological material other than genetic material *per se*, and may also be able to create genetic material from what was previously considered to be only biological material. Such products potentially include the manufacture of vaccines and peptide antibiotics. In addition, the larger the DSI databases become the more useful they will be for the design and then production of novel biological (e.g. enzymes or synthetic biological circuits) and/or biological genetic materials (e.g. genomes).

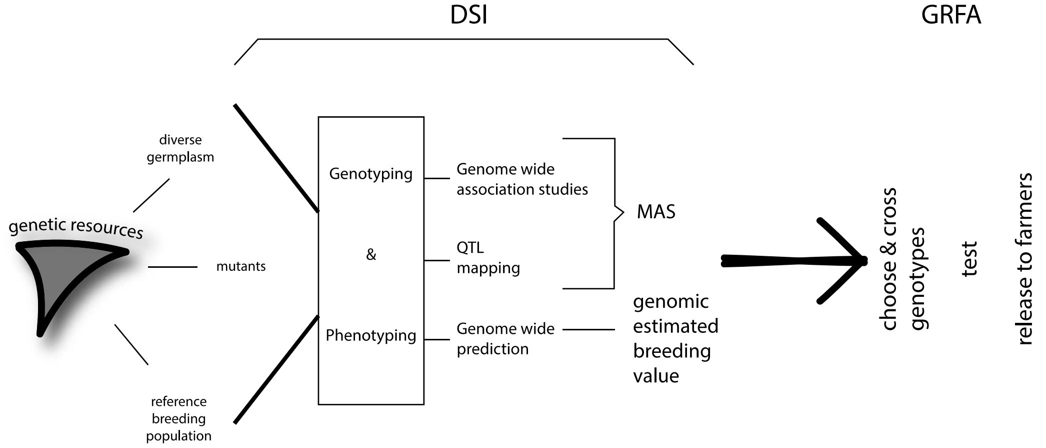


Figure 7. Stylized depiction of plant and animal breeding assisted by DSI
Source: (Varshney *et al.*, 2014).

Implications for access and benefit-sharing

An estimated annual contribution of ~USD300 000 000 is provided by governments to maintain >1 700 public databases. This cost is mainly borne by wealthier countries, which also largely host the computers within which the databases reside. Long-term public funding for the databases is not guaranteed (Editor, 2016).

Big data has big demands on infrastructure. The scale of genome sequence information alone can be so large that even in developed countries it is sometimes more efficient to transport and exchange hard drives than it is to electronically transmit the data (Marx, 2013). What might take a week or a month to transmit electronically in a developed country may be effectively impossible to transmit in poorer countries. Thus, while access may in theory be free or nearly so, in some cases it is largely only hypothetical.

Turning access into derived benefit also has challenges (Helmy *et al.*, 2016). Processing power and human resources are significant limiting factors to use.

Whether or not access and benefit-sharing (ABS) legislation includes the use of DSI on GRFA for the conservation and sustainable use of GRFA, including exchange, access and the fair and equitable sharing of the benefits arising from their use, the research and commercial sectors may react in unexpected ways. Some fear that any additional compliance or subscription costs may slow the distribution of assets for upstream research (Manzella, 2016) as the information is either not generated or it is kept secret. However, as recorded in interviews for this study, failure of legislation to provide a framework for ABS might have similar effects. Researchers and businesses may prioritize the value of DSI to country of origin, or to indigenous peoples, and either keep the data secret or abandon its collection altogether.

Conclusion

DSI as captured by the physical media of databases and how it has been used for many decades in research, industry and public resources management was examined in this study. This included DSI on GRFA that has added value to these genetic resources. DSI, for example, has been used since the 1990s by companies that specialize in monitoring transgenes in international trade or certify seed stocks as “GMO free”, or test for potential patent infringements (Heinemann, 2007; Heinemann *et al.*, 2004).

The study also looked at new and envisioned uses relevant to GRFA. The actual and especially potential value of DSI on GRFA is rapidly increasing because highly anticipated new technologies would be impossible without it. Some of these technologies even convert DSI on physical electronic media into biological genetic material using only machines and chemical precursors.

This study was informed by the latest scientific literature, with many examples taken from papers in the last two years. Some of these references inform the future looking nature of the study. Those references in general come from premier, mainstream scientific journals or interviews with scientists that are at the forefront of developments.

While DSI has long had value separate from the biological genetic material that it describes, some of the projections of future value may fail to be realized, or even later viewed as hype. Nevertheless, there is wide-scale enthusiasm among both basic researchers and scientific entrepreneurs for emerging technologies that can reify DSI as GRFA, or add value to GRFA. There has been no abatement of effort to further develop these and indeed even newer technologies. Such high-quality projections cannot be ignored for the purposes of informing a policy discussion, especially on *potential* value of DSI on GRFA.

I. INTRODUCTION

This exploratory fact-finding scoping study examines how DSI on genetic resources for food and agriculture is being used currently, how it might be used in the future and what the implications of its use are and might be in the future for the food and agriculture sector. The study covers:

- terminology used in this area;
- current status of biotechnologies (including identification, characterization, breeding and genetic improvement and conservation of GRFA) in the management of GRFA and agroecosystems and future developments;
- the types and extent of current and future uses of DSI on GRFA in biotechnologies; and
- actors involved with DSI on GRFA, and the relevance of DSI now and in the future on GRFA for food security and nutrition.

1965 Nobel laurette Jacques Monod famously said that "What was true for *E. coli* is true for the elephant". There would prove to be many exceptions to this prediction about the biology of a bacterium and an elephant. Nevertheless, Monod's remark was insightful for guiding the new science of molecular biology which was to eventually discover remarkable similarities in how genes and genomes work across all of life. The quality of the prediction he made is especially evident when comparing organisms that do not describe the extreme ends of the biological spectrum (Lidder and Sonnino, 2011).

The modern equivalent of Monod's assertion is that what can be done with DSI in one organism at least may be done in any other organism of similar type. When it comes to the use of DSI, what is true for the elephant is true for the goat. What is true for *Candida albicans* (a pathogen) is true for the *Saccharomyces cerevisiae* (the yeast used to make fermented beverages and bread). For some uses of DSI, what can be done in *E. coli* can also be done in the elephant.[12] This simple prediction has been true enough to make the use of DSI routine in the characterization, conservation and sustainable use of both organisms that are and are not considered to be genetic resources for food and agriculture (GRFA) (CBD, 2018).

The Monod anecdote informs us that sometimes when discussing the use of DSI on GRFA, it is both prudent and necessary to use an example involving an organism that some might not consider to be GRFA (at least not at the moment). At the cutting edge of science, the example proves the principle. Although it remains to be seen how many proofs of principle will become routine uses of DSI on GRFA, some uses of DSI have long been part of biological research on GRFA and its management.

Monod's prediction has also increased in value and prescience with changes in technology over time. When the insecticidal qualities of the soil bacterium *Bacillus thuringenesis* became apparent, this species of bacteria became a GRFA or at the very least a resource for management of GRFA. Subsequently, individual fragments of DNA from *B. thuringenesis* were found to confer upon crop plants, such as cotton, soybeans and maize, the trait of toxicity to insects (Heisey and Rubenstein, 2015). Genetic engineering technology was used to create crop GRFA that had one or more *B. thuringenesis* genes, making these bacteria unequivocally a genetic resource.

In one way, the use of DSI is an incremental advance in that progression of *B. thuringenesis* from soil microorganism to GRFA. Based on the DNA sequence of the genes from *B. thuringenesis* for the insecticidal traits, entire genes can be synthesized from chemical percursors and then inserted into crop plant. Isolating the bacteria from soil is not necessary.

[12] For example, the gene editing technique CRISPR/Cas9, which uses DNA molecules synthesized from DSI, can be used in both *E. coli* and elephants. "The plan sounds wild – but efforts to make mammals more mammoth-like have been going on for a while....Church says that he has edited about 14 such genes in elephant embryos" (Reardon, 2016)

In other ways, DSI is a significant advance on the use of *B. thuringenesis* DNA. For example, the DNA sequence of the gene from the bacterium can be altered in such a way as to create the same protein in plants, but is expressed more efficiently than would be the literal sequence from the bacterium.[13] Bioinformatics was the science of studying the sequences of genes that revealed these differences between plants and *B. thuringenesis*. Additionally, DSI has been used to design and create novel combinatorial toxicities by combining regions of different toxin genes (Honee *et al.*, 1990).

Comparative genomics can use DSI on genomes from different species to identify valuable genes, extending the use of DSI beyond natural breeding populations. For example, the copy number variation (CNV) information of the cattle genome was used to build a high resolution array to map CNVs in the less well study genomes of goats (Fontanesi *et al.*, 2010).

This makes it possible to extend DSI to organisms that are both outside the natural breeding range and traditional descriptions of GRFA. The kiwi bird is an example of an organism that by law is not food or farmed but is edible and historically was eaten. It has a unique genome that is important to GRFA both for the genes for traits it carries, and for the genes it does not have. The kiwi bird genome can be used by comparative genomics to provide clues about what unknown genes in organisms that we do eat – both bird and mammal – may be responsible for traits that may be improved or eliminated (Le Duc *et al.*, 2015).

The variety of uses of DSI in biology in general is growing along with other changes in biotechnology and computation. Those changes are illustrated by improvements in three co-evolving technologies (Figure 5). The first of these is in scale of DNA sequencing (“reading”).[14] The second is in scale of DNA synthesis to manufacture genomes (“writing”) or make DNA molecules used to modify existing genomes (“editing”). “Since 1975, reading and writing platforms have exhibited increases in throughput of three-million-fold and one-billion-fold” and a million-fold reduction in cost in the past 10 years in both reading and writing (Chari and Church, 2017).

The third of these is in computational power for analysis of DNA sequences along with increases in storage and transmission of data. As an indicator of growth in processing power, the number of transistors per chip has been doubling every two years until 2004 (Waldrop, 2016). Genomics data is estimated to use exabytes[15] of storage by 2025 (Stephens *et al.*, 2015).

1.1 Co-evolving technologies

DNA sequencing technology has its origins in the 1970s (Mardis, 2017). For many years sequencing was done on a small scale, was laborious and expensive. Large public and private investments in the genome sequencing projects of the late 1980s fertilized the development of technologies that could affordably deliver the sequence of entire genomes.

With increased capacity has come decreased cost (NIH, 2016). The first human genome sequence delivered in 2000 is estimated to have cost USD500 000 000–1 000 000 000. Higher quality data could be produced by 2006 at an estimated cost of only USD14 000 000–25 000 000, and by the end of 2015 it was as low as USD1 500 per genome.

DNA synthesis technology also has its origins around 1970, with the production of molecules called oligomers composed of small numbers of nucleotides (Amarnarth and Broom, 1977). This technology too has improved to where now it is possible to synthesize DNA molecules that correspond to genes (e.g. on the order of over 1 000 nucleotides), which are then connected to make even larger, up to chromosome sized, molecules.

[13] This is due to different codon preferences for the same amino acid in the different organisms.

[14] Reading is defined as the use of massively parallel sequencing technologies to decipher nucleotide order and writing as the use of either genome-editing tools or DNA synthesis to make changes in DNA molecules (Chari and Church, 2017).

[15] A unit of information equal to a quintillion (10^{18}).

The scale at which information about genetic resources can now be gathered, stored and transmitted is vast compared to even the end of the last Century, and capacity is rapidly increasing. As many as 2.5 million plant and animal species are expected to have complete genome sequences, and between 100 million and 2 billion people may have their genomes sequenced, by 2025 (Stephens *et al.*, 2015). The rate of growth in this capacity, already possibly doubling once every seven months, portends a future where all genomes may one day be described and variations in them recorded in near real time, globally.

Costs of synthesis are currently about USD0.15 per base pair (Petrone, 2016). The cost to synthesize the human genome, as a collection of gene sized molecules, would be USD450 000 000. This would not include the cost of assembling the fragments into a whole genome. Not only is the synthesis cost prohibitive, but presently the assembly would be unmanageable (Perkel, 2017).

Nevertheless, there is growing evidence that in time technological advancements will overcome barriers to the construction of viruses and organisms using only DSI. Small genomes can be synthesized either entirely or in large parts that are then assembled in yeast cells in the laboratory (Perkel, 2017). The complete synthesis of five chromosomes of the yeast *Saccharomyces cerevisiae* is evidence that even larger genomes may soon be assembled, if not outright synthesized (Richardson *et al.*, 2017).

Not just genomes but also obtaining other material necessary for a virus or cellular organism to reproduce is becoming more independent of possessing the original biological material (Gibson, 2014; Jewett and Forster, 2010). Indeed, the history of work on this extends back many decades. In the 1960s, researchers isolated and purified *in vivo* synthesized T4 and lambda (λ) viral genomes and mixed them together with isolated and purified viral coat proteins to reconstitute viruses *in vitro* (Edgar and Wood, 1966; Kaiser and Masuda, 1973; Weigle, 1966). By 2002, the 7 500 RNA nucleotides long poliovirus genome was synthesized. To recreate the viral genome, the researchers assembled together shorter synthesized DNA molecules (Wimmer, 2006). The RNA genome was synthesized using enzymes that assemble RNA polymers from DNA "templates". The RNA genome was combined together in a test tube with the enzymes that increased the number of RNA genomes and that use RNA to construct proteins. This mixture was able to produce infectious viral particles (Wimmer, 2006).

The critical role of DSI in the reconstruction of the poliovirus was emphasized retrospectively by one of the researchers. "Our experiment has thus overthrown one axiom in biology—namely, that the proliferation of cells or, for that matter, viruses depends on the physical presence of a functional genome to instruct the replication process. It was believed that without parental genomes, no daughter cells or progeny viruses would arise. We have broken this fundamental law of biology by reducing poliovirus to a chemical entity, *which can be synthesized on the basis of information stored in the public domain*—an experimental proof of principle that is applicable to the synthesis of all viruses" (emphasis added to Wimmer, 2006). Since then, several other viruses have been reconstructed in this way (Wimmer, 2006) indicating that this proof of principle could be extended to viruses relevant to GRFA.

In the case of poliovirus reconstruction, the enzymes needed for the production of the virus particle were isolated from human cells. This step may not be necessary in the future. Machines such as the "digital-to-biological converter" that are now being tested can, from a mix of precursor molecules and an electronic DNA sequence, synthesize new DNA genomes and the proteins (Boles *et al.*, 2017). The converter "receives digitally transmitted DNA sequence information and converts it into biopolymers, such as DNA, RNA and proteins, as well as complex entities such as viral particles, without any human intervention" (Boles *et al.*, 2017).

De novo synthesis from DSI is neither always the intention nor the most important new technology. Instead, the technology of gene editing, where an existing genome can be converted incrementally to conform to a new design (Chari and Church, 2017; Perkel, 2017), will more likely be the more common application of DSI at least into the foreseeable future. Its generality was demonstrated by changing multiple occurrences of a particular DNA sequence *in vivo*. The DNA sequence TAG was changed at 321 separate locations in the bacterium *Escherichia coli* genome using short “mutagenic” DNA molecules, making it possible to now assign the sequence code TAG to a new kind of amino acid for incorporation into proteins (Gallagher *et al.*, 2014).

Gene editing can be applied to GRFA. These techniques have been used to create new traits in plants. This can be used to increase the efficiency of insertion of transgenes or to modify existing alleles based on the DNA sequences in other species. For example, using a technique called oligonucleotide-directed mutagenesis (ODM), herbicide resistance traits were introduced into canola and flax (Songstad *et al.*, 2017).

Complementing the changes in DNA reading and writing, and the physical information technology infrastructure, have been changes in bioinformatics, the science of working with biological information (Lidder and Sonnino, 2011; Oldham, 2009). Bioinformatics, including genomics and synthetic biology, has been more than a field in which DNA sequences are compared. Within this sector has emerged a culture of large shared databases, tools and software for collaborative work.

The synergy of the changes in the technologies and culture of working with large datasets discussed in this section have reduced costs and increased capacity in genome sequencing and DNA synthesis. Nevertheless, the direct costs, infrastructure requirements and training will remain significant barriers to those in developing countries (Helmy *et al.*, 2016).

1.2 Interchangeability of material

Bioinformatics has been applied to extracting information from both DNA sequences and the scientific literature (Al-Aamri *et al.*, 2017). These kinds of information can be presented in very different ways, ranging from highly structured to unstructured. They can also be carried by different media, from DNA molecules, to written text, to electronic media. Still, the informational content can be studied using a shared methodology – bioinformatics.

DNA sequence information, in the form of DSI, is transferable between the DNA molecules of organisms, to the media of electronic storage and transmission, and back to molecules of DNA (Boles *et al.*, 2017). These different kinds of material are interchangeable for holding information. Information is thus fungible, allowing value to be created from biological material without access to the biological genetic material in a laboratory, environment or country.

In addition, information may be used in ways that are special to the medium, whether that be electronic or biological. A DNA molecule of a particular nucleotide sequence can be used in ways that an electronic medium of equivalent informational content about the sequence cannot be, and vice versa. The actual or potential value of these two material forms of the same information may vary according to whether it is coupled or uncoupled (“dematerialized”) from biological genetic material.

1.3 What’s next

This report summarizes the findings of the study examining terms associated with DSI to assist policy-makers and others to identify similarities and differences in how different sectors may view what is and is not DSI. Recognizing the existence of divergent views, it was a practical necessity to adopt a working definition of DSI for use in reporting the findings of the study.

This study scoped both current and *future* uses of DSI on GRFA, with an emphasis on actual or potential contributions to food security and nutrition. In the near and long-term future, the tools and techniques applied to working with information in DNA sequences will be extended to more biological materials than DNA, as well as more forms of DSI which will be used on GRFA the way that either DNA molecules or DNA sequences are now.

The potential implications of the use of DSI on GRFA for the characterization, conservation and sustainable use of GRFA, including exchange, access, and the fair and equitable sharing of the benefits arising from the use of DSI, are discussed.

Importantly, the use of DSI makes it possible to derive benefit from information about GRFA that can be used in other kinds of organisms or applications, and for information about organisms that are not GRFA to be used on or for the management of GRFA. This report therefore includes study findings of GRFA becoming a source of potential new pharmaceutical or industrial products, which constitute new potential income pathways for farmers.

A future looking analysis at the cutting edge of research in a scientific area runs the risk of appearing speculative, of describing science fiction rather than science. The report is therefore built upon concrete examples in a series of vignettes. The vignettes are not meant to be a comprehensive mapping of DSI on GRFA, but to demonstrate that which is, or nearly is, possible.

The study upon which this report is based was *exploratory* and *scoping*. It presents a “big picture” view of DSI on GRFA, not limited to GRFA affected by DSI. The examples used throughout this report illustrate the relevance of DSI to food security and nutrition. Although a sector-by-sector analysis was not requested, examples from all sectors are provided. Sub-sector-focused analysis could be useful in the future to answer more detailed questions about DSI on specific kinds of GRFA. The report of the study does not present comprehensive and fully detailed information due to constraints on length. However, its extensive bibliography will guide those interested in more details.

II. TERMINOLOGY AND DEFINITIONS

DSI has no universally agreed definition. This is reflected in decision XIII/16 of the Convention on Biological Diversity (CBD) and the decision of the Commission on Genetic Resources for Food and Agriculture (CGRFA) to establish a new work stream on DSI. Both decisions note that "there is recognition that there are a multiplicity of terms that have been used in this area (including, inter alia, "genetic sequence data, "genetic sequence information", "genetic information", "dematerialized genetic resources", "*in silico* utilization" and that further consideration is needed regarding the appropriate term or terms to be used" (paragraph 86 CGRFA, 2017).

How these and related terms are used in the relevant research literature will be described to explore potential implications of the use of DSI on GRFA for the conservation and sustainable use of GRFA, including exchange, access and the fair and equitable sharing of the benefits arising from their use.

2.1 Terminology

2.1.1 Genetic sequence data

Genetic sequence data (GSD) is a term consistent with the nucleotide order in molecules such as DNA. It is also a term used by the World Health Organization (WHO) in its Pandemic Influenza Preparedness (PIP) framework. In context, it is clear that WHO is referring to nucleotide and/or amino acid sequences associated with viruses because it recommends that GSD be stored in GenBank or the Global Initiative on Sharing All Influenza Data (GISAID), which do store this kind of information. For example, GenBank is the US National Institutes of Health's "genetic sequence database, an annotated collection of all publicly available DNA sequences" (GenBank).

2.1.2 Genetic information, genetic material and genomic sequence data

DSI is a term so far rarely found in the scientific literature (Figure 3). The terms GSD, genetic information, genetic material and genomic sequence data are more common. In the scientific literature, genetic information is generally described in material terms. There is a long scientific tradition of calling DNA sequences genetic information or data, such as in the phrases "genetically transmitted information" (Anonymous) or "genetic information in DNA is conveyed by the sequence of its four nucleotide building blocks" (Watson *et al.*, 2014).

Genetic material is sometimes used in the scientific literature as it is used in international instruments such as the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA, Treaty) to mean "any material of plant origin, including reproductive and vegetative propagating material, containing functional units of heredity" (ITPGRFA). More often, it is used to mean DNA, genes or genetic information. In the famous 1953 paper by James Watson and Francis Crick describing the structure of DNA, they equated DNA with genetic material: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material" (Watson and Crick, 1953). The equation of subcellular macromolecular structures with the genetic material was made in 1935 by Nobel laureate H.J. Muller referring to structures on the scale of chromosomes (Muller and Prokofyeva, 1935). This terminology remains in common usage today (e.g. Griffiths *et al.*, 2000; Ofir and Sorek, 2017; Stegemann and Bock, 2009).

Different concepts of genetic material probably reflect different terminologies in agricultural and molecular biology sciences. Prior to the consolidation of the hypothesis that DNA could be the material form of the gene, geneticists were restricted to working on material of cellular or larger scales and considered these to be the minimum reproducing unit.

Referring to genetic material as "any material of plant origin, including reproductive and vegetative propagating material, containing functional units of heredity" has a long tradition in management of GRFA, and is consistent with classical applied genetics because the predominant tool is breeding. The agricultural genetics literature from the 1940s explicitly equated "genetic material" with that which could recreate the plant (seeds or propagules) (Weiss, 1943). In all the cases described above from the molecular biology literature, genetic *information* had an associated physical, tangible, material. This is consistent with how geneticists view inheritance. In nature, genetic information is routinely exchanged between different media (Annex).[16] The essence of heredity is the transfer of DNA sequence information, not the atoms of the physical molecules of DNA that store the information. A child may contain no atoms from the DNA molecules of his/her grandparents, but he/she inherits their genes. The genetic information in the grandchild has a physical continuity from the grandparents, but not a conservation of molecular material.

The commercial use of information may also take a material view (Mueller, 2003) such as in the following examples of a patent and a Material Transfer Agreement (MTA).

- A patent granted in 2004 claimed any DNA molecule at least 95 percent identical to a specified sequence from the human and GRFA pathogen *E. coli*, and all computer readable media that contain similar DNA sequences (Dillon *et al.*, 2004). Computer readable media with DNA sequences were equated with DNA molecules of a sequence of nucleotides.[17]
- An MTA on genomic information that clearly defined data *per se* as the object of transfer. "A model MTA…defined 'material' as 'the annotated draft assembled rice genome *sequence* described in [name of the publication] and *the medium on which it is provided*' (emphasis added). The MTA also covered modifications (e.g. multiple sequence alignments and/or gene predictions) and unmodified derivatives ('substances or data created by the recipient which contain/incorporate the material and includes without being limiting replicated forms of the material and all cells, tissues, plants and seed containing/incorporating any part of the material')" (recounted in Manzella, 2016). It is worth emphasizing that the MTA formally recognized that the genetic information is in a physical – material – form whether it is DNA or instead a medium used by a computer.[18]

2.1.3 Dematerialized genetic resources

Dematerialization of the use of genetic resources was described by the Secretary of the ITPGRFA (Treaty) as "the increasing trend for the information and knowledge content of genetic material to be extracted, processed and exchanged in its own right, detached from the physical exchange of the plant genetic material: value is increasingly created at the level of the processing and use of such information and knowledge" (emphasis added to FAO, 2013).

In paragraph 86 of the CGRFA-16/17/Report (CGRFA, 2017), "dematerialized genetic resources" is included as a term associated with DSI. Dematerialized genetic resources could be what some refer to as intangible genetic resources, or be included in derivative products (Bagley, 2016; Bagley and Rai, 2013; Tvedt and Young, 2007). Intangible when referring to property can mean "information, reputation, and the like" and "is generally characterized as non-physical property where owner's rights surround control of physical manifestations or tokens of some abstract idea or type" (Moore, 2000).

[16] For example, in the retrovirus lifecycle (see Annex).

[17] "In one application of this embodiment, one or more nucleotide sequences of the present invention can be recorded on computer readable media. As used herein, 'computer readable media' refers to any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. A skilled artisan can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide sequence of the present invention" (Dillon *et al.*, 2004).

[18] Of course this is not to say that only when information can be commoditized does it have actual or potential value.

Intangible genetic resources distinguishes between "invisible" information and "visible" genetic material and as a concept contrasts with materialistic scientific conceptions discussed above, but also from "dematerialized use" where information is a commodity or has a value independent of biological material. As the Secretary said: "Breeding is no longer so focused on plant genetic resources as *raw materials*" (FAO, 2013). If the focus is only this, then "your Treaty which, even in embryonic form, address the non-material values of genetic resources, such as Article 17, and Articles 13.2a and b, you may, in seven to ten years' time, find yourself with a static museum of raw materials, rather than a dynamic framework which facilitates and governs the ongoing innovative uses of genetic resources, at the level of their epistemic value, i.e. their information and knowledge values, including characterization data, genomic data, traits and even environmental data" (FAO, 2013). Likewise, in the First Meeting of the Expert Consultation on the Global Information System on Plant Genetic Resources for Food and Agriculture (GLIS), information associated with germplasm was described alongside "value addition activities" (paragraph 6 GLIS, 2015).

The value of the genetic material of seeds is in the biological material that may be used or exchanged. Dematerialization *of the use* of genetic resources, where value is created at the level of the processing and use of, for example, DNA sequence in databases, is made possible because of the contextual and quantitative information embedded into it. As is discussed in more detail in Chapter III, knowing DNA sequences of the plant genome can be used to identify or design physical and analytic tools (for "markers") that can "be used to follow the inheritance of linked traits during the breeding cycle, reducing phenotyping requirements" (GLIS, 2015). To sell or use these is not to use the seed. As such, both use and value have been uncoupled from the seed or its exchange, but it nonetheless supports, for example, breeding and conservation outcomes.

As is discussed in more detail in Chapters III and IV, information and knowledge content of genetic material is being gathered on a scale that creates even more opportunities for value. Genome sequences have become the scale of "big data" (Rajan, 2015). Information on genetic resources provides the basis for levels of analysis that would be impossible with smaller data sets. Qualitatively new skills and knowledge emerge from this resource, not all of which require, or must be applied to, biological material.

2.2 Common characteristics of terms

The wide range of terms associated with DSI and discussed above have some characteristics in common. With the possible exception of the term "dematerialized genetic resources", they all refer to things with the ability to interchange information stored in physical material, such as DNA, with information *in silico*, that which is used in, for example, electronic media.

The terms differ in broadness of the kinds of information that apply to GRFA. GSD can be used in a way that is more specific than is genetic information which is more specific than *in silico* utilization and dematerialized use of genetic resources. Setting aside metadata, GSD could mean no more than a series of letters, such as A,G,C,T, representing a sequence of nucleotides in DNA.

In practice, however, GSD is always linked to metadata and schemata, which are contextual (Figure 6). They range from such things as sequence orientation to geographic and phenotypic data and are ubiquitous in the databases (Manzella, 2016). For an example of how contextual DSI from biome studies is used in GRFA, see Box II.1. These data also may be used separately from the sequence data.

The terms also have in common that they describe kinds of information that are amenable to study using the same methodologies. The same methodologies can be applied to information in DNA sequences and the scientific literature (Al-Aamri *et al.*, 2017). Bioinformatics is a science of such methods. It is the "organization and analysis of biological and related information, usually involving the use of computers to develop databases, retrieval mechanisms, and data analysis tools, especially in the fields of molecular biology, structural biology, and genetics" (NLM). Synthetic biology is another. It links bioinformatics and biological engineering, from design through to construction.

2.2.1 Beyond DNA

"[I]nstead of asking 'Which resources are genetic resources?' we are tempted to ask 'When is a biological resource not a genetic resource?' Alas, the answers to this second question have also proven illusive" (Tvedt and Young, 2007).

While some of the terms used seem to be limited to information on DNA, other terms embrace more than DNA. In fact DSI could be understood not to be limited to information on DNA sequences. Instead, DSI could embrace all genetic information in living things. DNA sequence information alone is not always sufficient to characterize diversity or select for desired phenotypes of GRFA (Box II.2). The difference between a desirable trait and its alternative therefore may be identified, stored and transmitted as DSI that is not a DNA sequence.

The over 1 700 databases of DSI in routine use (Table II.1) range in coverage from familiar DNA DSI such as in GenBank to LincSNP, which links single nucleotide polymorphisms (SNPs) to disease, LNCediting, a database for functional edits to a specific class of RNA molecules (and which only occur in RNA), RCSB that associates genes and proteins with three-dimensional protein structures, on to BioGRID that lists gene and protein interactions as well as post-translational modifications to proteins and PlaMoM that aids prediction about macromolecules transported in the plant phloem based on stored DSI (Various, 2017).[19] The European Bioinformatics Institute (EBI) in the United Kingdom is one of the largest archives of biology data and only 10 percent is DNA sequences (Marx, 2013). All the categories except "Human Genes and Disease" (with just two databases) have content relevant to GRFA.

In the First Meeting of the Expert Consultation on the Global Information System (GLIS) on Plant Genetic Resources for Food and Agriculture, the experts associated "a wide range of different types of data available about PGRFA" including genomics but also beyond DNA sequences, such as phenomics. Relevant expertise of experts appointed to the Scientific Advisory Committee on the GLIS Article 17 included "bioinformatics and molecular genetics; the 'omics, in particular genomics, phenomics and proteomics; management of environmental and geo-spatial data about plant genetic resources" (Appendix 5 GLIS, 2015).

2.2.2 Codes

Macromolecules, including ones that are not nucleic acids, are increasingly being discussed in the scientific community as possessing forms of genetic information. Those discussions use terms such as epigenetics, epigenome, non-DNA inheritance and prions. A new field being dubbed "code biology" is devoted to the actual or potential characteristics of diverse biological materials to be amenable to bioinformatics-type studies (Barbieri, 2018). How DSI on GRFA is used by scientists now portents a time when DNA itself may not be center stage, or at least must share the stage with many more forms of DSI such as the metabolic, histone, sugar, splicing, tubulin, signaling, ubiquitin and glycomic codes (Barbieri, 2018; Buckeridge, 2018; Gabius, 2018; Marijuán *et al.*, 2018; Prakash and Fournier, 2018).

[19] The journal provides a Web link to the databases (NAR).

Box II.1. DSI of the rumen biome and its use in management of animal genetic resources.

There are about 200 species of ruminant animals and nine species that are domesticated as livestock, including cattle, buffaloes, sheep, goats, reindeer and camels. The estimated population size is 3.6 billion animals worldwide, with the vast majority in developing countries (Hackmann and Spain, 2010; Steinfeld *et al.*, 2006).

The rumen harbors a complex microbial ecosystem that allows these animals to digest plant material that other animals and humans cannot efficiently use (McSweeney and Mackie, 2012). This is due to the particular mix of micro-organisms found in the rumen. These microbial communities break down complex plant polysaccharides that are indigestible or unproductively used by monogastric animals (de Tarso *et al.*, 2016).

The rumen in ruminant livestock has a highly diverse anaerobic microbiome. The rumen microbiome is a source of metabolic diversity. Knowledge about it can be applied to improve animal health and productivity, lignocellulose digestion and to the mitigation of climate change (Campanaro *et al.*, 2017).

Within the last few years, several significant metagenomic surveys of the rumen microbiome have been completed, creating DSI that is applied to the dual challenges of increasing feed efficiency and reducing greenhouse gas emissions (Campanaro *et al.*, 2017; Henderson *et al.*, 2015). For example, an end product of anaerobic digestion is methane. Most of the 37 percent of anthropogenic methane comes from fermentation in the rumen of ruminant livestock. Unfortunately, methane is a greenhouse gas with 23 times the global warming potential of CO_2. Changes to fermentation efficiency could reduce methane emissions and improve retention of feed energy for the animal (Gerber *et al.*, 2013; Henderson *et al.*, 2015).

In one study, rumen microbiomes were surveyed from 742 animals from seven different global regions (Henderson *et al.*, 2015). The diversity of the rumen microbiome was surprisingly consistent. Differences were in the ratio of the sizes of the populations of different species, and in which species had the largest populations. The most important determinant of these difference was diet because the majority of microorganisms in the rumen ferment the most feed (Henderson *et al.*, 2015).

Diet drove the demographics of the microbial community and the greatest variation was among the bacteria in the microbial community of the rumen. Many rumen bacteria are known only through genomic DNA sequences. However, DSI from harvested DNA sequences could still be used to predict the metabolic phenotypes of the microbes, not just infer the potential from a description of the microbial species. Parenthetically, this study is an example of both the value of DSI for metagenomic analyses and international collaborations resulting in the release of DSI on GRFA from both developing and developed countries.

In a complementary “genome-centric metagenomics” approach to surveying different ruminants on different diets, another study extracted rumen fluid (containing the microbiome) and fed cultures of the same microbiome different diets. Metabolic pathways were constructed based on detected genes from the metagenomic sequencing because they differed depending on the polysaccharide/protein composition of the experimentally varied diets (Campanaro *et al.*, 2017).

The two different methodologies used by these studies identified the key microbes contributing to methane production. Such knowledge might lead to manipulations of the metabolome of the rumen or diets of livestock for improved management of animal genetic resources.

Box II.2. DSI contributes to identifying link between genotype and phenotype.

The genotype and the environment determine the phenotype. Over time, the environment, including anthropogenic influences through breeding, may select and amplify genetic differences between populations (Boggess *et al.*, 2013). However, the environment influences phenotype in more ways than just allelic variation. For example, different environments can result variations in the proteins produced from the same gene. The protein diversity of a single genome relative to the number of genes "can be considered to increase 5-fold by considering all [the RNA] splice forms per protein. When the ~300 [post-translational modifications] are considered the number of distinct protein chains rises to around greater than 500,000. On top of this precise proteolytic processing can convert the parent chain to two or more stable daughter chains giving rise to up to potentially 1,000,000 protein chains" (Overall, 2014).

Critically, "*understanding the genetic sequence alone does not provide this information*" (emphasis added to Overall, 2014). Not all traits can be identified from DNA (or more broadly genome) sequences. That is one reason why DSI on GRFA is not just DNA databases.

Functional omics is changing breeding of GRFA from the traditional approach of crossing desired phenotypes to concentrate desired genetic traits, to the use of full phenotype and genotype DSI to predict the "phenotype from a known genotype (genomics) and a full accounting of all environmental effects. This is the approach that will eventually provide an extraordinary improvement in animal and plant production systems, but is predicated on a comprehensive understanding of biological proteomics and metabolomics" (Boggess *et al.*, 2013).

With the help of DSI, phenotypic detail at the omics level is accessed. Using omics DSI informs searches for genes or consortia of genes that provide a particular phenotype. Derived DNA sequences can be used as a tool for breeding, such as when the sequences are used in MAS to enrich for particular genes. Phenotype DSI can be used to synthesize a genome that produces the desired phenotype, or to edit a genome to create the phenotype. Through such uses of DSI, DNA becomes the product, rather than the source, of the information that produces value from a genetic resource. For example:

- Researchers interested in finding new proteins that could be used in place of antibiotics found that they could design those proteins from an analysis of proteins with varying degrees of antibacterial activity (Speck-Planche *et al.*, 2016). Proteins with antibacterial activities are made by all organisms, including GRFA, just as they can be used to treat infections in GRFA. The existing proteins were identified through DSI that is annotated DNA sequences or amino acid sequences, or from the scientific literature. The fundamental information necessary was not DNA sequence information, but either inferred or known amino acid sequences. Using bioinformatics approaches, models were built that could predict with high accuracy what sequence of amino acids in a protein would have antibacterial activity against the largest number of pathogens, or be specific to types of pathogens (Speck-Planche *et al.*, 2016). From this information, new genes could be sought in existing DNA databases to find the organisms producing peptides with desired antibacterial activities, or new genes could be synthesized that would correspond to proteins of the desired attributes. The fundamental information necessary to produce value from an antibacterial peptide came from publicly accessible databases, but the products were not just a recreation of already described genes or extracted from an organism taken from a particular environment.
- Not only is the proteome diversity larger than can be predicted from the number of genes in a genome (Wilhelm *et al.*, 2014), but also metabolite diversity is larger than can be predicted from the known number of enzymes (Moriya *et al.*, 2016). This suggests that either enzymes have considerable 'moonlighting' activity and engage in more diverse reactions than currently known, or that many unassigned genes encode enzymes not assigned to particular substrate (metabolites).

This inspired researchers to search for DSI that described enzyme-substrate structures, rather than DNA sequences of enzyme genes, because the latter approach requires "functional associations to other identified enzyme genes in the context of pathway or genome information" (Moriya *et al.*, 2016). Using databases such as KEGG, the researchers used whole-structure comparison of substrate‒product pairs of metabolites of a pathway to identify the likely features of proteins that would serve as catalysts and then to work from that to identify DNA sequences that are predicted to give rise to such proteins (Moriya *et al.*, 2016).

Where the DNA alone is not explanatory, DSI may reveal epigenetic or epiallelic variations (Fortes and Gallusci, 2017). Epigenes leading to selectable traits can be mapped as epiQTL, in analogy to QTL for genes. EpiQTL have been linked to a large number of important crop traits and stability is important for trueness-to-type in cultivars (see Table 1 in Rodríguez López and Wilkinson, 2015). Where these are due to differences in DNA methylation, specific markers can be developed, in analogy to DNA markers for MAS.

- Genome-wide changes in methylation of DNA nucleotides can occur during micropropagation of clones, resulting in significant phenotypic changes. "For instance, in oil palm, mantled inflorescence syndrome was found to be associated with global changes to C-methylation status during micropropagation, and caused catastrophic reductions in yield among all affected plants and incurred huge costs to the industry" (Rodríguez López and Wilkinson, 2015). Phenotypic variation from changes in epialleles is an important source of diversity even in highly inbred staple crop cultivars, and may contribute more phenotypic plasticity than allelic variation (Rodríguez López and Wilkinson, 2015). Methylation distinguishes epialleles of the gene FIE1 in rice and P1-rr in maize. These genes affect flower phenotypes or pigments in the grain (Rodríguez López and Wilkinson, 2015).
- Tibetan pigs are adapted to the high altitude environment of 4,000+ metres. Adaptation has been linked to various allelic markers (Jin *et al.*, 2018). The Tibetan pig is now also being farmed in low land areas, providing an ideal comparator for identifying epigenetic markers associated with acclimation to relative hyperoxic conditions. The methylation status of genes in the genomes of pigs reared at different elevations was determined using bisulphite DNA sequencing identifying genes related to hypoxia, oxygen transport and energy metabolism associated with significantly different expression levels (Jin *et al.*, 2018).

Table II.1: NAR Database Summary Paper Category List

Nucleotide sequence databases
RNA sequence databases
Protein sequence databases
Structure databases
Genomics databases (non-vertebrate)
Metabolic and signaling pathways
Human and other vertebrate genomes
Human genes and diseases
Microarray data and other gene expression databases
Proteomics resources
Other molecular biology databases
Organelle databases
Plant databases
Immunological databases
Cell biology

2.2.3 Epigenetics

DNA has been used as a synonym of the gene for over a half century. It is proven to be the physical basis of inheritance (Heinemann, 2004). However, there are other materials that also transmit some traits either infectiously or during reproduction, as does DNA. These are referred to as epigenes and they also are the physical basis of inheritance of some traits (Springer and Schmitz, 2017). One kind of epigenetic inheritance is the propagation of differential methylation patterns on DNA.

DNA is composed of four chemically distinct nucleotides, A, G, C and T, and the sequence of these nucleotides is part of what is archived by a genetic sequence database such as GenBank. A common modification of nucleotides is to add a methyl group (CH_3). The addition and removal of methyl groups is associated with functions varying from protecting DNA from the actions of nucleases to effects on gene expression and differentiation of cells in a multicellular organism, all of which might also be the basis of important traits.

Breeding combines and recombines genetic diversity within different genomes represented by changes in the DNA sequence in each gene, but also by changes to the DNA around a gene (sometimes thousands of nucleotides away) and to variations in the chemical modification of nucleotides in a gene, and other ways (Schafer and Nadeau, 2015).

Modification of nucleotides is a dynamic process that may occur multiple times within the life of an individual cell, but may also be stable and in some cases heritable (Tao *et al.*, 2017). Whereas sequence variants of a gene are called alleles, modification of the same underlying DNA sequence yields multiple variants of the gene, called epialleles (Gallusci *et al.*, 2017). Epiallelic variation also can be the basis of desirable traits that provide value from GRFA (Gallusci *et al.*, 2017; Springer and Schmitz, 2017).

2.2.4 Phenomics

Information from biological systems is also being used to augment, even replace, DNA DSI on GRFA. Genomic DNA studies "will eventually provide an extraordinary improvement in animal and plant production systems, but is predicated on a comprehensive understanding of biological proteomics and metabolomics" (Boggess *et al.*, 2013).

Genotype from DNA sequences from many individuals, breeds and species can be directly compared. Similarities provide information on evolutionary relationships and identify variations in genes that are linked to desirable traits. DSI is as valuable as a tool to indicate what different genomes do not have in common as it is to identify what they do (Remington *et al.*, 2005). Comparing the flightless nocturnal kiwi bird genome to other birds can identify genes that are associated with different traits.

Phenotypes are the characteristics, or traits, of an organism. The correspondence of DNA sequence to phenotypes is not the same for all phenotypes. Simple cases in which a single DNA sequence alone completely determines a phenotype are comparatively easy to document. However, many phenotypes are the outcome of many different genes and some genes influence many different phenotypes. Moreover, phenotypes are the product of both genetics as reflected in DNA sequence and also environmental influences including temperature, altitude, diet, water availability and microbiome.

Phenomics is "the use of large scale approaches to study how genetic instructions from a single gene or the whole genome translate into the full set of phenotypic traits of an organism" (NIFA-NSF, 2011). It draws upon information about gene by environment interactions that are measured using techniques collectively called "omics" technologies (Lidder and Sonnino, 2011).

Omics technologies can identify the entirety of specific molecules in a tissue, cell or organism (Heinemann *et al.*, 2011). Omics includes genomics, transcriptomics, proteomics, metabolomics and even larger groupings. For example, liver and skin cells have very different sets of RNAs (transcriptomics), proteins (proteomics) and metabolites (metabolomics), and these differences can vary across time or due to other variables, such as development or stress.

Omics techniques provide useful information but none alone provides a complete picture of an organism. For example, "it is essential to measure protein levels and post-translational protein modifications to reveal information about stress inducible signal perception and transduction, translational activity and induced protein levels" (Ghatak *et al.*, 2017). In addition to the value of phenotypic data, omics DSI can be used to augment breeding of GRFA. "Eventually, these processes will provide more direct insight into stress perception then [sic] genetic markers and might build a complementary basis for future marker-assisted selection of drought resistance" in crops (Ghatak *et al.*, 2017).

The metabolome is the set of metabolic products arising from the sequence of reactions in metabolism. Software platforms using databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG), MapMan, KaPPA-view, MetaCyc and MetGenMap facilitate the identification of the associated genes and metabolites by mapping metabolomic data onto metabolic pathways (Abdelrahman *et al.*, 2017). This provides breeders with metabolomic quantitative trait loci (mQTL) markers.

In one application, metabolites from the leaves of 289 distinct young maize lines were used to associate metabolomics and agronomic traits to genomic markers (Riedelsheimer *et al.*, 2012). Genome-wide association mapping linked 26 metabolites with SNPs. The power of the analysis was sufficient to identify plausible candidate genes consistent with the function of the metabolite in maize.

2.3 Conclusion

This chapter provided a summary of views on the terms associated with DSI. Consideration of the terms identifies meanings that might be shared, or at least points to some common illustrative examples of what they mean. Whatever terminology is used to describe DSI, it is important that it be inclusive for the current and potential future uses in the food and agriculture sector to avoid becoming irrelevant.

With inclusivity in mind, this report uses DSI to mean *inter alia* that which could be held by any existing or future database of the types summarized by *Nucleic Acids Research* (Various, 2017) on its website (NAR) and used within the science of bioinformatics.

III. CURRENT STATUS OF BIOTECHNOLOGIES USING DSI IN THE MANAGEMENT OF GRFA AND AGROECOSYSTEMS AND THE TYPES AND EXTENT OF CURRENT USES OF DSI ON GRFA IN BIOTECHNOLOGIES

DSI has many existing and potential future roles in GRFA management and on GRFA used in biotechnologies. Key applications have relevance to the three objectives of the Commission and the Treaty, conservation and sustainable use of GRFA and the fair and equitable sharing of the benefits arising out of their use.

This chapter briefly discusses the conservation and sustainable use of GRFA, including microorganisms, fungi, plants and animals. Each section includes at least one substantial vignette that maps an existing or emerging role for DSI to characterization, conservation or sustainable use of GRFA.[20]

These vignettes are intended to illustrate scientific and commercial uses, drawing from the credible, peer-reviewed and latest scientific literature. That necessarily restricts examples to the past, present or near future. However, it is evident that all the kinds of information described in the previous chapter are being taken from and used on GRFA.

3.1 Characterization

> “Scientists can now rapidly read the DNA of an organism—even a plant—anywhere. Researchers at the Royal Botanic Gardens, Kew, have recently reported on their use of a handheld real-time DNA sequencing device that allowed them to identify the various species of an entire field of plants far faster than could be done using previous methods. This was the first time genomic sequencing of plants has been performed in the field. They highlight the new opportunities that real-time nanopore sequencing (RTnS) offers for plant research and conservation” (Palminteri, 2017).

DSI on GRFA is applied to describe and characterize genetic resources. Characterization provides information about organisms and ecosystems that may be used for basic research, identification, monitoring and to draw comparisons with other organisms and ecosystems. Molecular markers, phenomics and bioinformatics are used in the characterization of GRFA (Lidder and Sonnino, 2011).

Characterization involves the systematic collection of information about genotypes, traits, phenotypes, genetic diversity and distance, population structure and size and geographical distribution (Lidder and Sonnino, 2011).

Molecular markers are used to identify and track alleles for quantitative traits. When different combinations of alleles of different genes cause different phenotypes, the trait is said to be quantitative. DSI is used to create the tools for following different alleles through breeding.

Barcoding is a technology that improves identification particularly of species that are difficult to distinguish between by morphology. Barcoding uses DNA sequences from a gene or genomic location for species identification and discovery (Lidder and Sonnino, 2011). A variation of barcoding is metabarcoding, where an estimate of biological diversity is derived from direct isolation of DNA from a sample, such as a soil sample (Orwin *et al.*, 2018).

DSI has become an important component of breeding by contributing to the rate and accuracy of trait mapping to develop targets for MAS. Moreover, it is important for calculating genomic estimated breeding value (Varshney *et al.*, 2014). It also assists in setting priorities for the GRFA to be conserved.

Vignette III.1.1: DSI from whole genome sequencing for MAS in yam

The white Guinea yam, *Dioscorea rotundata*, grown in West and Central Africa, is both 96 percent of yam produced globally and an important crop for local food security (Tamiru *et al.*, 2017). A breeding programme for yam could help to reduce their susceptibility to disease and nematodes. Because the yam are usually propagated clonally, as tubers, breeding for these traits

[20] It is possible that a vignette could apply to more than one section of the report.

is slow. Further complicating improvement is the high heterozygosity of yam because it makes them not amenable to linkage analysis using the segregating progeny of an F2 generation and recombinant inbred lines.

DSI plays an important role in other approaches to yam breeding and management DSI from whole genome sequencing was used to identify a genetic marker for sex identification that could be used at the seedling stage. The DSI resources underpin MAS in a genomics-assisted breeding programme. "Molecular markers, such as simple sequence repeats (SSRs), indels, and SNPs, can, for the first time, be developed for various applications in Guinea yam, including linkage mapping, genome-wide association analysis, genomic selection, and MAS" (Tamiru *et al.*, 2017). By providing improved tools for breeding, DSI on yams contributes to food security and nutrition of local communities and increases the potential for income from the crop.

Vignette III.1.2: Metabarcoding DSI to characterize microbial ecosystems

Food security and nutrition is improved by healthy soils, which in turn are dependent on soil micro-organisms. Metabarcoding, a high-throughput variation of barcoding using metagenomic or genomic DNA, can be applied to describe the microbial ecosystem on everything from store-packaged food (Higgins *et al.*, 2018) to large agroecosystems. While barcoding can be used as a tool for the conservation of species, metabarcoding can be used for conservation of ecosystem functions.

Quantitative microbial community structure descriptions were built from DSI of 16S rRNA DNA profiles of ecosystems dominated by five different uses, natural and planted forest, vineyards and improved and unimproved grasslands in New Zealand (Orwin *et al.*, 2018). Metabarcoding profiles based on nucleotide sequences correlated well with the more traditional tool, phospholipid fatty acid profiles, and as reliable predictors of ecosystem function.

Describing the microbial community helps to identify land uses and diagnose soil functions such as nutrient levels and decomposition rates. For example, soils with higher levels of *Acidobacteria* may have less drought tolerance than soils with *Actinobacteria* for those soil functions. Ultimately, DSI built from metabarcoding might help to improve food security and nutrition through either direct supplementation with specific kinds of microbes or amendments that alter the soil microbial community.[21]

Vignette III.1.3: Characterization of epigenetic markers in forestry trees

Forest trees are in general long lived and thus can experience significant environmental variability. Their long generation times make it more difficult to breed them for adaptive traits. This creates special challenges for breeding and management of tree biology and forestry. Fortunately, trees tend to have a high degree of phenotypic plasticity. Understanding the underlying heritable basis for plasticity is a key goal for breeding. "Diverse environmental stresses and hybridization/polyploidization events can create reversible heritable epigenetic marks that can be transmitted to subsequent generations as a form of molecular 'memory'. Epigenetic changes might also contribute to the ability of plants to colonize or persist in variable environments" (Bräutigam *et al.*, 2013).

DSI of epigenetic markers is important for forestry. An example of an epigenetic marker is DNA methylation of DNA sequences. These markers have been used to demonstrate, among other things, a relationship between response to water deficit and the genes associated with phenotypic plasticity in poplar, an important insight for breeding. A heritable epigenetic marker in eucalyptus has also been linked to phenotypic variation in cellulose content, an important trait for use (Bräutigam *et al.*, 2013).

[21] Another example is the Ecobiomics Project (CBD, 2018).

3.2 Conservation

Plant and animal genetic resources are needed for food security and nutrition and, for breeds at risk, genetic resources are needed for traditions and cultural values or to satisfy emergent niche markets (CGRFA, 2012). With "75 percent of crop genetic diversity [having] been lost in the past century; 17 percent of the world's livestock breeds...classified as being at risk of extinction" these resources are under threat (FAO, 2017).

Strategies for conserving genetic diversity include *ex situ*, *ex situ-in vivo*, and *in situ* conservation (CGRFA, 2012, 2014). Effective management of diversity requires that the particular mix of conservation strategies is suited to the objectives, which are specific to the genetic resource (e.g. see Table 1 of CGRFA, 2012). In general, all three strategies must to some degree ensure that relevant organisms are identified and that the population is large enough to hold representative genetic diversity in the collection or environment, the organisms or germplasm are viable, the diversity can be accessed when needed, and information is available and easily disseminated, e.g. by electronic means (CGRFA, 2014).

DSI in the form of species surveys also contributes to species conservation (CBD, 2018). DSI is frequently used for identification of species and for assessing genetic diversity within species. Other tools, such as metabolomics and proteomics, can provide even more detailed assessments of diversity. Diversity measures based on the genome only can miss other heritable differences in diversity, for example by selecting populations that tend to produce a certain kind of proteome. Moreover, those measures exclude epigenetic characterizations. What appear to be similar populations based on DNA sequences may not be equivalent based on epigenetically determined traits (see Box II.2).

Vignette III.2.1: DSI applied to halting loss of plant GRFA

DSI contributes to the preservation of genetic diversity, contributing to improved food security and nutrition. Italy's Cilento region is a UNESCO heritage site protecting the biodiversity at the origin of the Mediterranean diet. Italy is home to several common bean landraces. Some landraces are in danger of extinction. Over 60 percent of the common bean ecotypes have been lost over the last 100 years (Hammer *et al.*, 1996). In order to describe the genetic diversity among landraces, as part of a conservation effort, 12 different Italian landraces were barcoded.

Barcodes refer to DNA sequences that are diagnostic of the origin of DNA in a sample. In conservation work, the objective is to use species-specific barcodes to identify all species represented in a sample. Both the tools (i.e. primers) for amplifying diagnostic barcodes from a sample, and the barcode itself, are stored and transmitted as DSI of DNA sequences in databases.

The amplified DNA from the landraces was sequenced and then compared with DSI in international databases (De Luca *et al.*, 2017). The study found that *in situ* conservation by farmers, geographical isolation, and the biology of the bean all contributed to genetic distinctiveness of the landraces. The tools and methods developed in the study are an ongoing resource to help preserve the genetic diversity of the landraces, which are "part of the social history and cultural identity of the communities living in the area". The study authors concluded that the work "highlighted that conservation of landraces is important not only for the cultural and socio-economic value that they have for local communities, but also because the time and conditions in which they have been selected have led to that genetic distinctiveness that is at the basis of many potential agronomical applications and dietary benefits" (De Luca *et al.*, 2017).

Vignette III.2.2: DSI applied to halting loss of animal GRFA

> "The increasing availability of genomic tools, accelerated through novel applications and decreasing costs of massively parallel high-throughput sequencing (HTS) technologies and computational resources, will create new avenues for research in the field of conservation genetics...Moreover, genomic tools can be directly used in wildlife management by helping to diagnose the causes of population declines, for testing the health of managed wildlife populations, and by directly improving the fitness of managed populations through 'omics assisted breeding" (Lozier and Zayed, 2017).

DSI contributes to the conservation of threatened species, such as pollinators, contributing to improved food security. Bees provide food as honey and they are essential pollinators for >80 percent of flowering plant species including plant GRFA (López-Uribe *et al.*, 2017). Bees pollinate an estimated one-third of the food we consume and boost agricultural productivity of an estimated two-thirds of crops (Suni *et al.*, 2017). In many environments, bee populations are threatened although the causes are not always known and may be multiple, including habitat destruction, disease and agrichemicals (Lozier and Zayed, 2017).

DSI on bee genomes and microbiomes can contribute to preservation and restoration of bee populations. Whole genome sequences of 11 bee species are already available. This important resource is just a start, however, because there are an estimated 16 000 bee species (Lozier and Zayed, 2017). One application of these sequences is to test the possibility that DNA or RNA-based pesticides (see vignette 4.1.2) could cause off-target acute, chronic or behavioral effects on bees (Lundgren and Duan, 2013). Bioinformatics using DSI helps to inform testing of these pesticides for effects on bees (Mogren and Lundgren, 2017). From a literature survey, 101 small double stranded RNAs that were designed as pesticides because they cause gene silencing in a target pest, were found by bioinformatics analysis potentially capable of binding target RNA molecules in honey bees. Such *in silico* analysis is insufficient to either prove or disprove the possibility of off-target effect of any dsRNA, which also requires a plausible exposure pathway, but it helps to identify risk hypotheses that can then be further tested (Heinemann *et al.*, 2013).

3.3 Sustainable use of GRFA

Global food security depends on the ability to produce a nutritious food supply without depleting the ecological resources upon which it relies. Genetic resources in food and agriculture are under increasing demands to feed more people on less land, or in warming oceans. The means of production are challenged by natural disasters and the increasing intensity of weather events due to climate change.

Sustainable agriculture is dependent on sustainable development and biodiversity. This is recognized in the Sustainable Development Goals such as 2 on food security and sustainable agriculture, 14 on life below water, 15 on life on land, as well as 12 on sustainable consumption and production (FAO, 2017). Agriculture is not just a way to produce food, but also a way to practice *in situ* conservation, improve child health and gender equality (Taberlet *et al.*, 2008; UNEP/UNCTAD, 2008). Some 2.6 billion people worldwide are directly engaged in agriculture, and in developing countries 30–60 percent of all livelihoods come from agricultural and allied activities (IAASTD, 2009).

DSI contributes to these broad goals of sustainable use. It is used in reproductive biotechnologies such as sperm and embryo sexing in animals, disease diagnostics and control (e.g. vaccine development), chromosome set manipulation, tissue culture, MAS, and genetic engineering, among others (Lidder and Sonnino, 2011). DSI is developed to track the movement of genes between domesticated and wild populations of GRFA. For example, movement from domesticated to wild species might endanger small wild populations if it caused outbreeding depression (Lidder and Sonnino, 2011).

DSI is used to discover and design new pesticides, biofertilizers and probiotics (Lidder and Sonnino, 2011).

Furthermore, sustainable use has economic dimensions. Farmers need adequate returns from their products. Certification and traceability add value to agriculture by meeting consumer demands. DSI is used to meet certification and traceability standards, as well as to ensure compliance with some food safety standards.

Vignette III.3.1: Landscape genomics DSI used in animal GRFA breeding

Improvements to GRFA through breeding contribute to food security and nutrition. The use of defined markers can help to accelerate breeding programmes by enabling the identification of desired traits even before they are evident by observation (Figure 7). For example, sequences from the chromosomes linked to improved performance of cereal crops under drought conditions (e.g. seedling establishment, root vigor, or transpiration efficiency) have been identified

(Richards *et al.*, 2010) as has a marker for resistance determinants to stem rust in wheat (Periyannan *et al.*, 2014). Using DSI, variants in alleles of genes can be used to select individuals with desired combinations.

SNP analysis is a tool that uses DSI (Box 2). A SNP is the position in a genome where a single nucleotide (A, G, C or T) differs between two genomes, with one usually being a reference genome (Yang *et al.*, 2013). Up to 90 percent of the differences between genomes within the same species are SNPs (Koopaee and Koshkoiyeh, 2014). The different versions of the DNA sequence are called alleles. Each genome can contain thousands of SNPs that can be used as physical markers for breeding and conservation of species. SNPs can be identified using DSI in specialist databases, such as LincSNP (Various, 2017).

SNPs are used in, for example, dairy cattle to predict performance (yield, milk protein content, fertility) in different environments (e.g. high altitude, heat, humidity) to choose individuals that will perform best (predict how offspring will perform in a different environment based on SNPs, e.g. bull from Europe, offspring in Australia) (Haile-Mariam *et al.*). A library of at least 10 000 SNPs is generally required for genomics analysis of livestock (Woolliams and Oldenbroek, 2017).

The distribution of SNPs in genomes is used to infer relatedness between individuals. In conservation of endangered species, whether natural populations or livestock breeds, SNP data can be used to choose the most appropriate (usually least related) individuals for breeding programmes to ensure the highest possible genetic diversity in the population.

SNPs contribute to landscape genomics which examines the order of genes in genomes, which genes belong to the same neighborhoods, how often this is true in a species and between species and finally links such data to phenotypes (Brbić *et al.*, 2016). Landscape genomics can also be used to detect selection signatures. Landscape genomics is a promising approach for optimizing genetic potential by adapting breeding stock to locations. The alternative approach, to optimize traits under a defined set of conditions, fails when those conditions cannot be recreated where the livestock are raised.

Researchers have used SNPs, geographical data and information about production systems to describe indigenous goat populations in South Africa (Mdladla *et al.*, 2017). “Adaptation of animal genotypes to the environment is central to the coping of local livestock populations to changing climatic and production objectives and is a key parameter to use in conservation and improvement programs” (Mdladla *et al.*, 2017) and therefore to food security and nutrition.

Landscape genomics studies on goat populations validate other approaches to identifying breeds of unique genetic character and potential value. The combination of lower costs to generating relevant DSI for such studies and the diversity of traits represented by locally adapted animal populations could produce important gains for marginalized farmers in less developed countries (Mdladla *et al.*, 2017).

The field of landscape genomics illustrates well how local climate and topography data must be combined with DNA sequence data to extract greater value from a genetic resource. The actual or potential value of some forms of DSI cannot be realized if disconnected from other forms of DSI.

In another application of landscape genomics, the genomes of five indigenous African cattle were sequenced (Kim *et al.*, 2017). These cattle breeds are spread across highly different agroecological zones. Because of this, the place in which the cattle live can have a significant effect on their genes over time. If a quantitative trait such as greater heat tolerance or tick resistance is desired, those traits are more likely to be revealed through comparisons of genomes of cattle adapted to environments that significantly differ in temperature and types of pests.

“African cattle populations represent a unique genetic resource for the understanding of the role of natural and artificial selection in the shaping of the functional diversity of a ruminant species. Moreover, unravelling their genome diversity may provide new insights into the genetic mechanisms underlying their adaptation to various agro-ecosystems” (Kim *et al.*, 2017).

The sample size was far from exhaustive as an examination of cattle genetic resources in Africa – with an estimated 150 breeds on the continent – but sufficient to demonstrate Africa-specific adaptations. Geography-genome patterns were identified that likely associate quantitative trait loci (QTL) of populations of cattle more tolerant of heat and other aspects of the climate and local diseases, among other things.

Vignette III.3.2: Epigenome DSI used in plant and animal GRFA breeding

A study of three pig breeds, the western Landrace and Tibetan and Rongchang Chinese breeds, investigated differences in the pattern of adipose tissue deposition (Li, 2012). These traits are important for breeding livestock that produce the most desirable meat at the lowest resource input. The study revealed that the fat distribution pattern was an epigenetic trait arising from the pattern of methylated nucleotides in around 100 genes in each of 44 groups from eight adipose tissues.

Variations in the epigenome may be environmentally induced or part of the developmental pattern of the organism (Springer and Schmitz, 2017). For example, the frequency of methylation on cytosine (C) nucleotides varied more between leaves of the same inflorescence on the shrub of *Lavandula latifolia*, grown commercially for production of essential oil for aromatherapy and fragrance and also for honey production and horticultural value, than they did on average between individual shrubs (Alonso *et al.*, 2017). The size and number of seeds produced by the leaves in an inflorescence was correlated with environmentally influenced sectorial methylation levels.

A methylation pattern may be propagated in two different ways. One is where the pattern persists through gametogenesis, as in imprinting (Li *et al.*, 1993), and, at least in plants, the other is where the organism may be propagated clonally, as through cuttings (Springer and Schmitz, 2017).

In addition to modification of DNA, the proteins that bind to chromosomes and control their structure, called histones, are also chemically modified. This variation in modification of the histones influences how tightly packed a chromosome is in different regions, further modulating the expression of genes in that region (Strzyz, 2017). Histone modifications vary over time and by type of cell, but can also be heritable and are thus part of how different cells in a multicellular organism differentiate to create different tissues and organs. Histone modifications have been called the histone code (Jenuwein and Allis, 2001; Prakash and Fournier, 2018; Rando, 2012).

A recent patent emphasizes the commercial importance of epialleles of epigenetic traits in GRFA. The primary claim of the patent is to provide a method to maintain in crop plants commercially important epigenetic traits through breeding. The patent distinguishes not between plants with different DNA sequences in their genomes, but different epialleles of DNA identical genes. The agronomic performance is dependent on the stable inheritance of the epigenetic marker (Fromm, 2017).

As has been highlighted elsewhere in this report, DSI is stored with a suite of contextual information (i.e. metadata and schemata) that makes the representations useful. Increasingly, contextual information includes epigenetic markers such as DNA and histone modifications because they are the basis of traits with actual or potential value. Databases specializing in methylated nucleotide patterns include MethDB for DNA and MeT-DB for RNA methylation (NAR). Databases providing DSI on histones and modifications include CR Cistrome, HHMD, Histome, Histone Database, and WERAM (NAR).

Vignette III.3.3: Microbial metagenomic and metabolomic DSI used in food safety

Food security is dependent upon access to food that is safe to eat. DSI can and will more frequently be used to ensure food safety. For example, metagenomic and metabolomic DSI contributes to technologies and approaches limiting spoilage.

The DNA in a sample of microbial communities found on various retail foods is converted into DNA sequence to develop a metagenomic profile (Higgins *et al.*, 2018). DSI generated from the DNA sequences can be used to identify specific potential metabolic activities or link species identification with known species-specific traits, or to identify pathogens (King *et al.*, 2017). The type of communities on food varied depending not just on product, but the socio-economic status of the neighborhood. The information has the potential to improve food safety by identifying and controlling risks at the local level. The data was used to predict the range of microbial metabolic activities, finding a predicted 761 on lettuce and 710 on deli meat (Higgins *et al.*, 2018). This information reflected what nutrients were available to the microbes and what they would produce, helping to predict the likelihood of spoilage.

Metagenomics-generated genome DNA sequences can be coupled with transcriptomics (sequencing all RNA molecules) or proteomics (identifying all proteins) on food samples to describe the physiological state of the community, further increasing confidence in the association between a species' identified genome and its predicted metabolism (King *et al.*, 2017).

When combined over time or geography, or by industry, these types of local analyses contribute to "big data" sets. Trends or other diagnostic indicators of causes to either threats to food safety or opportunities for improvements may be identified.

Vignette III.3.4: DSI on microbes applied to the utilization of GRFA

> "In several industrialized countries, public health agencies and regulatory bodies [e.g. Food and Drug Administration (FDA), Centers for Disease Control and Prevention (CDC), Public Health England, European Center for Disease Prevention and Control (ECDC)] are now using WGS [whole genome sequencing] *routinely to characterize* clinical isolates of selected foodborne pathogens and to support epidemiological investigations" (emphasis added to Rantsiou *et al.*, in press).

Food security and sustainable income for farmers are dependent on confidence of access to safe food. A significant challenge comes from post-harvest loss due to contamination by microbes that can be human pathogens. DSI generated from whole genome sequencing of foodborne pathogens is used in surveillance programmes. WGS is replacing older technologies and approaches to monitoring foodborne pathogens (Rantsiou *et al.*, in press).

Dedicated databases such as GenomeTrakr in the United States of America and the ECDC in the European Union help to link outbreaks with the food or environmental source of the pathogens (Rantsiou *et al.*, in press). The databases will also be able to identify genes that contribute to antibiotic resistance or virulence, and thus help to inform medical responses.

Vignette III.3.5: DSI for plant and animal GRFA health

Food security and use of GRFA is dependent upon the health and vigor of GRFA. Infectious diseases that are important in horticulture and animal husbandry can have multiple effects on GRFA. They can threaten entire industries; such was the worry when *Pseudomonas syringae* pv. actinidiae was found on New Zealand's kiwifruit vines in 2010 (McCann *et al.*, 2013). They can threaten trade such as the foot and mouth outbreak in the United Kingdom in 2001 (Knight-Jones and Rushton, 2013).

Finally, they can be the source of pathogens that transfer to people, such as influenza from birds and swine (WHO, 2016). This is not only an important consideration for human health, but carries the risk of reputational damage as illustrated by the concerns surrounding bovine spongiform encephalopathy in cattle from the United Kingdom in the 1990s (Hope, 1995).

Importantly, pathogens are an important resource for the design and manufacture of medicines such as vaccines. For example, instead of sending seed influenza viruses to different countries for large-scale vaccine manufacture, they can now be synthesized on location using transmitted DSI and amplified for vaccine manufacture (Donner, 2013). Vaccines could be made available weeks or months earlier.

Other important veterinary and medical responses also can be expedited using DSI. Epidemiology of antibiotic resistant bacteria through genotyping of strains can provide good predictions of which drugs the pathogen will still respond to and, even if they have some resistance, how much drug should be used (Burnham *et al.*, 2017; Carrico *et al.*, 2013). The database Resfinder is dedicated to identifying genes or gene variants that occur in resistant bacteria (ResFinder). Bacteria can travel quickly through both wild and domesticated animals or infected plant material, and they do not respect political borders. The collection and sharing of DNA sequences from such pathogens for the primary purpose of informing diagnosis and therapy, thus containing outbreaks as early as possible using the minimum amount of antibiotic, is an example of DSI providing shared benefits.

Vignette III.3.6: DSI for product certification, labeling and traceability

Sustainable use of GRFA is dependent upon access to markets and consumer confidence in supply. Many consumers participate in food systems through product labeling and certification schemes (Clarke, 2010). Their choices help shape industry and government priorities.

Product labeling verification using DSI can both be used for traceability (Sultana *et al.*, 2018) and compliance with laws that promote the conservation and sustainable use of genetic resources. Fish fillets traded in one Italian market were found to be mislabeled 32 percent of the time, and sometimes species of high conservation value were substituted for those listed on the label (Filonzi *et al.*, 2010). In another market, products were mislabeled 82 percent of the time and substitutions included threatened species (Di Pinto *et al.*, 2015).

“The study also highlighted that threatened, Vulnerable (VU), Endangered (EN) and Critically Endangered (CR) species considered to be facing a high risk of extinction has been used in the place of commercial species…Additionally, traceability may improve the management of hazards related to fish safety, as well as guaranteeing product authenticity, providing reliable information to customers, enhancing supply-side management and improving product quality and sustainability” (Di Pinto *et al.*, 2015).

Vignette III.3.7: DSI for generating new products from GRFA

New products from GRFA increase both income security and the financial sustainability of farmers. DSI is a critical element in new product development. The world food system is estimated to be worth USD4 trillion (Chaudhry and Castle, 2011). Income-generating activities in this market include reducing food production and post-production waste. Along these lines, nanotechnologies are expected to contribute new products to the food system relevant to production (e.g. agrichemicals and feed additives), processing, safety (e.g. detoxification), preservation (e.g. antimicrobials in packaging), taste, absorption of nutrients, and labels for traceability and authenticity (Chaudhry and Castle, 2011; Chen and Yada, 2011). The value contribution of nanotechnology to food packaging alone is estimated to exceed USD20 billion, up from the 2006 value of only USD4 million (Chaudhry and Castle, 2011).

Bionanotechnologies may also provide new income from waste products. One source of new products is animal blood as a source of proteins and new bioactive compounds (Bah *et al.*, 2013). Presently, blood is collected at abattoirs at a rate of 4–7 percent of carcass lean meat. It is a high volume material but is mainly sold as low-value animal food or fertilizer, or incurs a cost for disposal.

New uses of blood would be an income for farmers. Blood carries high concentrations of the protein hemoglobin. Thus, new applications for this protein could create a new value chain for a by-product. Amyloid fibrils with useful bionanotechnology qualities can be made from hemoglobin (Jayawardena *et al.*, 2017).

DSI can be used to predict whether a protein will form an amyloid fibril. Under the right conditions, all proteins denature and may refold as aggregates (Knowles and Buehler, 2011). One form of aggregate is an amyloid fibril. Fibrils may be used to form nanowires.

Linking a small domain of one protein that can form amyloid fibrils to another protein is often sufficient to cause the second protein to also form amyloid fibrils. Knowledge of the amino acid sequence of that fibril-forming domain was used to make a variant of the metalloprotein rubredoxin which then spontaneously assembled into an electricity conducting nanowire (Altamura *et al.*, 2017).

Some proteins that form amyloid fibrils also are prions, proteins that behave like genetic material performing information storage and transfer functions (Chernoff, 2004; King and Diaz-Avalos, 2004; Knowles and Buehler, 2011; Wickner *et al.*, 2004). The domains of the proteins that confer this ability are predicted using sophisticated models trained on DSI (Toombs *et al.*, 2012), and thus may be identified in DSI or modified from the DSI of existing genetic resources (Antonets and Nizhnikov, 2017). This is useful both because some prions are associated with neurodegenerative diseases in people and animals, including livestock (e.g. 'Mad Cow Disease') and because proteins forming these nanostructures may be grown as potentially valuable nanofibers.

Amyloids have a number of exploitable properties, ranging from a biological matrix for surface adhesion to catalytic scaffolds as well as adhesives and structures for the storage of peptide hormones, and as a genetic material (Knowles and Buehler, 2011). These properties can be used to construct nanotubes through which a nanowire can form, or to construct biosensors and drug delivery media (Hamada *et al.*, 2004; Hauser *et al.*, 2014).

3.4 Conclusion

DSI may be used in various ways to support conservation and sustainable use of GRFA, promoting food security and nutrition. In priority areas of breeding, DSI is being used to associate desired phenotypes with genotypes that then might be amplified or improved through breeding. It is also being used to identify and then select genotypes that are likely to display desired phenotypes. These uses are complementary and may be combined.

DNA information is becoming as important for the discovery of new traits and biological functions and genes as are observable phenotypes. The CRISPR/*cas*9 system of bacteria, which now is the basis of a powerful tool for gene editing, was discovered first as an unexplained feature in genomic DNA sequences (Marraffini and Sontheimer, 2010). Only later were these sequence features linked to traits.

The examples show that DSI other than strictly DNA sequence information (and associated metadata) are also used for the conservation and sustainable use of GRFA. In some cases, DNA sequence information alone would not be enough to achieve the same outcome. For example, having whole genome sequences from indigenous African cattle would not identify putative quantitative trait loci associated with heat tolerance. Only when combined with meteorological and other kinds of data is it possible to identify relevant genes and desirable alleles. Likewise, it might be possible one day to use metabolomic data to characterize ecosystem function, or to reconstruct that function.

IV. WHAT ROLE DOES DSI HAVE IN RESEARCH AND PRODUCT DEVELOPMENT AND GRFA MANAGEMENT?

DSI has predominantly but not exclusively been a descriptive tool used in the biotechnologies of characterization, conservation and sustainable use of GRFA. For example, genomic DNA is sequenced and the sequences are used to help select genomes with alleles of genes that are associated with desired phenotypes (Figure 7). DSI is taking an ever increasingly active role in management and manufacture of GRFA. In this evolving role, DSI will augment capacity in GRFA management but also significantly increase potential for uses that could cause harm.

The shift in the value of DSI is increasingly uncoupled from a source biological material. This is in part due to the degree of independent specialization among those who use DSI or supply analyses based on DSI, and to the way technologies use DSI.

4.1 DSI in the management of GRFA

DSI from both GRFA and organisms that are not GFRA can be used to improve food security and nutrition. As discussed previously, DSI is used in comparative genomics drawing from non-GRFA organisms to associate traits with genes, and alleles with QTL. Organisms that are not GRFA are also important to optimize food production, limit waste and maximize farmer income.

Vignette IV.1.1: Microbiome DSI applied to quantitative traits of plant and animal GRFA

"'We're all of a sudden having tools to study entire communities rather than single microbes,' says Maggie Wagner, a plant biologist at North Carolina State University. Cheap sequencing means it's much easier to go out prospecting and cataloguing microbes. That's exactly what Indigo has spent the past several years doing. With its network of collaborators, the company has collected microbes from plants all around the world" (Zhang, 2016).

Microorganisms comprise the greatest biological and genetic diversity as well as numbers on the planet, and have done so far longer than any other kind of organism (Lau *et al.*, 2017). These organisms are not only a source of nutrients such as from fermented foods, but underpin all agricultural productivity. Soil can harbor thousands of different species of microbes in every gram and soil environments are among the most complex (Fierer, 2017; Lau *et al.*, 2017). Many of the microorganisms are only known through DSI, because they cannot be grown in the laboratory.

DSI from both metagenomics and molecules that are not nucleic acids (Orwin *et al.*, 2018) sourced from soil can inform farmers about soil functionality and its limits, possibly leading to customized management advice or products including supplements and probiotics. Alternatively, microorganisms might be harnessed as bio-reporters to monitor soil conditions or processes that would otherwise be difficult or expensive to monitor (Fierer, 2017). Soil microbes interact with plants and animals, and have demonstrated the potential through this interactions to mitigate effects of climate change (Lau *et al.*, 2017).

A microbiome may be unique to each organism or species and so then, within limits, microbiomes might be important for influencing how each responds to diet or even affects reproduction. Thus, DSI generated from characterization of the microbiome is the basis for optimizing food production and increasing food security and nutrition. Microbiomes may form specifically around different genetic variants (Chu, 2017; Luca *et al.*, 2018). In the fruit fly *Drosophila melanogaster*, a male's microbiome influences fecundity, and has other trans-generational effects, and thus provides a mechanism for how the microbiome and genome could co-evolve (Morimoto *et al.*, 2017).

Gene-by-microbiome interactions could be improved through breeding just as are quantitative trait loci. Metagenomic surveys of plant root biomes are intended to identify gene-by-microbiome interactions to determine how microbial consortia interact with individual plants to increase yields, and reduce the impacts of abiotic stress and disease (Busby *et al.*, 2017; Mauchline and Malone, 2017). Likewise, the microbiome of the rumen in ruminant livestock is a target of manipulation to lower methane gas emissions and assist in efforts to combat climate change (Tapio *et al.*, 2017). Metagenomes of ruminants can be combined with animal genetics to help to reduce methane emissions and contribute to climate change mitigation.

Vignette IV.1.2: DSI for pest management

Weeds, insects and disease have significant impacts on pre-harvest crop losses (Popp *et al.*, 2013). Weeds can lower the nutritional value of the crop or introduce undesirable toxins or allergens and insects can also help spread infections from bacteria and fungi (Heinemann, 2007). DSI from genomic and metagenomic profiles is being used to design DNA or RNA pesticides (Sammons *et al.*, 2015b; Tuttle and Woodfin, 2014). These pesticides are often referred to as types of “Biologicals” or “Agricultural Biologicals” to distinguish them from synthetic chemical pesticidal active ingredients. The *in vitro* produced nucleic acids are introduced into cells of pests using chemicals that lower the barriers to uptake of nucleic acids (Shaner and Beckie, 2013; Tuttle and Woodfin, 2014).

The new pesticides could be much less toxic to humans and damaging to the environment because the active ingredient specifically targets a complementary RNA or DNA sequence in the cells of the intended target, and may not cause such an interaction in organisms of other species (see vignette 3.2.2).[22] In the case of RNA-pesticides, the small RNA molecule may be used to cause silencing of a critical gene, as in RNA interference (Heinemann *et al.*, 2013), or similarly to a gene drive system as a guide to attack the DNA of the target pest (NCB, 2016).

To achieve this specificity, the base order of nucleic acid active ingredient is synthesized according to the expected DNA or RNA sequence derived from the genomes of the pest species. In other words, the active ingredient molecule(s) is made using DSI.

Biologicals show promise as herbicides and insecticides (Sammons *et al.*, 2015b; San Miguel and Scott, 2016; Van *et al.*, 2011; Zhu *et al.*, 2014). With real time monitoring, these pesticides could be further refined to work more effectively with local variants of pests, and to counter the evolution of resistance. Future advances may also see them developed to target mammalian pests or as livestock therapeutics, including prophylactically changing susceptibility to pathogens (NCB, 2016).

The use of RNA or DNA-based pesticides could also be linked to different technologies. Presently, herbicides based on RNA active ingredients are being developed to silence the genes that make weeds resistant to a chemical-based herbicide (Sammons *et al.*, 2015b). The RNA molecule makes the weed susceptible again, and the chemical-based herbicide kills the weed. Together, the two active ingredients restore the effectiveness of the chemical-based ingredient.

It might be possible to also genetically engineer crops with particular alleles of resistance genes (e.g. transgenes) sold together with the herbicidal combinations of RNA active ingredients and chemical herbicide to which they are immune, but would be toxic to both weeds or to crop plants raised from other genetically modified herbicide-resistant germplasm[23] (Ader *et al.*, 2011; Navarro, 2014; Sammons *et al.*, 2015a).

4.2 DSI in synthetic biology applied to GRFA

In general terms, synthetic biology is the field of science where biological function is fabricated and may be applied at a manufacturing or environmental scale (Piaggio *et al.*, 2017). DSI is used to both discover molecules with the desired biological functions or to design and manufacture

[22] This is aspirational and controversial. See: (FIFRA, 2014; Heinemann *et al.*, 2013)

[23] “Transgenic crops with one or more herbicide tolerances may need specialized methods of management to control weeds and volunteer crop plants. The method enables the targeting of a transgene for herbicide tolerance to permit the treated plants to become sensitive to the herbicide. For example, an EPSPS DNA contained in a transgenic crop event can be a target for trigger molecules in order to render the transgenic crop sensitive to application of the corresponding glyphosate containing herbicide” (Ader *et al.*, 2011).

them. The complete *in vitro* synthesis of a genome that was designed to be the minimum necessary to support viability of a cellular organism and its reproduction profoundly demonstrated how synthetic biology can be used (Gibson, 2014). Future applications range from manufacturing biofuels and food (Hayden, 2014) to ecosystem management through to gene drive systems (Piaggio *et al.*, 2017).[24]

Synthetic biology goes beyond creating just a new substrate-enzyme pair and encompasses "building, modeling, designing and fabricating novel biological systems using customized gene components that result in artificially created genetic circuitry" (Tyagi *et al.*, 2016). The fabricated genetic components therefore do not have to be copies of existing biological materials, such as existing alleles of genes, or even produce existing proteins. Indeed, synthetic genomes are being built using an altered genetic code allowing incorporation of novel amino acids into synthetic proteins (Chin, 2017). The fabricated genes, however, are designed based on observation of the natural world including its genetic diversity inspiring human-engineered combinations.

The goal of some synthetic biology work is to be able to predict multi-component interactions that are stable through reproduction. Synthetic biology can use multi-component, heritable interactions to build biological computers (Abels and Khisamutdinov, 2015). DNA, for example, can be the means of computation and environmental sensor (Sainz de Murieta and Rodríguez-Patón, 2012) where the computer could be placed inside genetic resources (The Economist, 2012).

Switches, especially those with switch state memory, are important for the developing field of biological computing. Describing those switches can yield new information about component function (Haynes *et al.*, 2008). Switch state memory can be used as a tool for monitoring an environment to improve productivity and quality of crops, livestock, forestry and fisheries, and to maintain healthy populations of wild species.

Switches can take different forms. A simple switch with memory is achieved with an invertible DNA sequence (Haynes *et al.*, 2008). More complicated switches arise from alternative heritable physiological states (Kotula *et al.*, 2014). Paradigm examples of bi-stable transcription state switches are the lac operon switch state inheritance (Novick and Weiner, 1957) and the bacterial virus λ lifecycle decision switch (Ptashne *et al.*, 1982).

Some examples exist of engineered switches (Khalil and Collins, 2010). The virus λ epigenetic switch has been used in context to amplify a biological signal (Heinemann, 1999). More recently, it was engineered to develop bacteria that served as environmental sensors with relevance to management of GRFA. In *E. coli* the engineered switch was set to detect low levels of antibiotic when passing through a mouse gastrointestinal system, but which could easily be applied to livestock. Bacteria that had sensed the antibiotic remained in a stable epigenetic state distinguishable from the pre-exposure state for up to five days (Kotula *et al.*, 2014).

4.3 DSI can have value separate from biological genetic material

"23andMe has managed to amass a collection of DNA information about 1.2 million people, which last year began to prove its value when the company revealed it had sold access to the data to more than 13 drug companies. One, Genentech, anted up USD10 million for a look at the genes of people with Parkinson's disease. That means 23andMe is monetizing DNA rather the way Facebook makes money from our 'likes'" (Regalado, 2016).

Big data is a term about the scale of data, but also the scale of data that is difficult to use. Structured data are the easiest form of data to analyze. Unstructured data, for example, text heavy, pictures or inconsistent formats, and omics, is common in many complex agricultural systems.

[24] Gene drive systems are being developed in tested in many different kinds of organisms, including mammals. The controversial systems are discussed for use as prest and predator control (Esvelt and Gemmell, 2017; Neves and Druml, 2017).

Big data platforms shift the use of data from analytics to prediction, where patterns in the data inform new questions rather than questions that are answered by finding patterns in the data. There are many potential uses of the information separate from the source (or destination) GRFA and therefore potential new value chains. Already, these approaches are being used to find lead drug candidates based on *in silico* screens, that is, uncoupled from screening using enzymes or tissue cultures. A candidate cystic fibrosis therapeutic specific to patients with a SNP in their cystic fibrosis transmembrane conductance regulator (CFTR) protein was identified from over 500 000 compounds using only computers (Costa, 2014).

The utility and availability of DSI might decrease the need to exchange GRFA for some uses. For example, biochemists looking for a particular substrate-enzyme pair will have reduced need to access biological material from which to extract enzymes. However, they may increase their demand for biological material that is currently under represented in databases, such as biomes of deep sea sulfur vents. It would be premature to conclude that as DSI grows in relevance, biological materials will decrease in relevance or be used less.

What the actual future balance is between using DSI or biological material will also be influenced by cost and convenience where the uses are substitutable. Interest in a research project or product development might be influenced *a priori* depending on perceptions of cost and convenience. Thus, the balance between DSI and biological material could in part be determined by how many scientists choose questions that are specialist to one or other resource.

Different entrepreneurial opportunities emerge when data collection is distributed, because of the scale of data that can be collected and the value that may be extracted from such collections. As the technologies of miniaturization improve and costs decline, collecting DSI likely will become more distributed, as has happened with access to mobile phones and the Internet and is happening with human genome sequencing (Contreras and Deshmukh, 2017). Retail consumer goods, phone "apps" and Internet e-mail providers either provide discounts or free services in exchange for monitoring consumer behavior. Analyzing the behavior of individuals produces custom advertising, maximizing consumer willingness to spend; analyzing the behavior of groups of people has even more diverse benefits.

Collection and collation of microbiomes of livestock to soil ecosystems and waterways could soon be possible as machines capable of real time metagenome profiling are developed. Sequencers the size of a mobile phone and powered by a USB slot in a computer have been available for several years, prompting predictions that "sequencers will become like telescopes: a formerly boutique scientific instrument that you can now buy from a toy store" (Yong, 2016).

In-field real time pathogen diagnosis for livestock and plant diseases could improve treatment outcomes and provide epidemiological data for country-wide surveillance of GRFA. A credit card-size cartridge holding a microfluidic chip that couples with a mobile phone for imaging has been developed to detect target DNA sequences in samples taken from animals (Chen *et al.*, 2017). The prototype was used to detect pathogens that cause equine respiratory infections to demonstrate its general use by veterinarians monitoring livestock. The prototype could monitor for eight different pathogen-specific DNA sequences simultaneously, and thus will detect multiple causative agents if present in a sick animal. Its detection limits were reportedly as good as commercially available laboratory-based equipment (Chen *et al.*, 2017).

These technologies could also have various law enforcement uses. They could be used to monitor infringement of intellectual property rights on crops and livestock, find raided livestock or verify product label information.

4.4 Conclusion

DSI can and is being used disconnected from biological material to support conservation and sustainable use of GRFA. In some cases, the value comes from re-coupling the DSI with biological material, such as in the use of metagenomics to identify probiotic bacteria to add as soil amendments. The inability to detect a species of bacteria in soil might be the trigger for seeking a probiotic amendment, or to add a nutrient that could amplify the species from undetectable to biologically relevant levels.

In other cases, value is created from DSI without using biological material (Sonka, 2014). An example of this was the invention of a device that used DSI to detect horse pathogens. Another example, from Chapter III, was using DSI to discover the gene editing tool CRISPR/Cas9. This discovery has led to over 600 patent claims since 2004 (Egelie *et al.*, 2016).

V. ACTORS AND ACCESS

5.1 Actors

Diverse actors are involved with DSI on GRFA. For the purposes of this study, they were identified from a variety of sources including scientific literature, the 2017 online forum on synthetic biology for the Convention on Biodiversity, and patent database searches. These actors were clustered into groups to inform an assessment of the relevance of DSI for food security and nutrition.

5.1.1 Governments and public sector institutions

The names of various databases holding DSI have been mentioned throughout this report. In the main, these databases are maintained by public and dedicated stewardship of public institutions. There are about 1 700 databases (NAR) estimated to cost USD300 million per annum to maintain (Editor, 2016). They sit alongside mainly public investment in among other things *ex situ* storage of biological and germplasm materials in the form of gene banks, medical specimen and museum collections (CGRFA; DiEuliis *et al.*, 2016; Fowler, 2016). DSI does not replace either *in situ* or *ex situ* genetic resource conservation strategies but provides an important complementary role and adds resilience as a form of redundancy in storage of certain kinds of biological genetic material, for example DNA.[25]

Publicly funded databases have generally been either open access or open license (Mueller, 2003), making them easily accessible to anyone with the correct kind of computing device and Internet access. For example, the International Nucleotide Sequence Database Collaboration (INSCD) consisting of NCBI/GenBank (US), EMBL-EBI (EU) and DDBJ (Japan) has a policy of permanent free and unrestricted access.[26] GISAID EpiFluTM database requires users to register and agree to terms, but anyone may sign up to it. However, pressure on public databases has resulted in some charging subscriptions for access (Hayden, 2016).

Public availability is not the equivalent of public control. The location of servers holding DSI and Internet traffic potentially can be controlled, even blocked, by state actors (Kamen, 2017). It is not possible to say whether all infrastructure and mirror sites are held by the country administering the database, but it is clear that those in the “Golden Set”, for example, are concentrated in the economic North (Table V.1).

Thus, the transfer of information on GRFA as DSI could depend on the will of the country in which the DSI physically resides or that controls the infrastructure or transmission, rather than the country from which the sequence information originated.

Developing countries also may lack the necessary infrastructure for ready access to DSI on GRFA and thus potentially fail to benefit from it. To work with the often unstructured form of biological DSI requires more local storage of information downloaded from international databases. The amount of storage is significant (Marx, 2013). Terabytes (10^{12}) of storage are routinely used for genomic comparisons.

An even larger barrier is transmission time. It can take so long to up or download data that even in developed countries it is not unusual to manually transfer it on portable hard drives rather than through an Internet network. BGI in China has the capacity to sequence six terabytes a day on its

[25] Tissue, cells and DNA as genetic resources: http://www.nies.go.jp/biology/en/aboutus/facility/capsule.html. DNA as a genetic resource: “There is a fifth type of genetic resource that differs from the other four types because it cannot be used to regenerate an organism and usually is not held in genebanks: **Cloned DNA sequences,** or genetic material from other organisms incorporated into crops by molecular techniques (for example, a gene from the bacteria *Bacillus thuringiensis* used for resistance to insects.)” (Heisey and Rubenstein, 2015)

[26] “Specifically, no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party.” http://www.insdc.org/policy.html

157 sequencers, but can transmit only one terabyte of data a day. It takes 20 days to send the raw data of just 50 human genomes (Marx, 2013).

Despite the high speed links between Europe and North America, researchers based in the United States of America were frustrated by download times from Europe's EBI. This was partially resolved by creating mirror sites closer to the US researchers. More mirror sites can and are being made, but mainly in developed countries (Marx, 2013).

Publicly funded databases put minimal to no restrictions on opportunity to access DSI, but this is not the same as saying that they are sufficient for equitable access. The latter requires overcoming significant inequalities among potential users to download and work with the data (Helmy *et al.*, 2016). Indeed, public availability is not *per se* a benefit, but a pre-requisite for any benefits that could arise from use of DSI.

Table V.1. Database owners

Number of databases	Location	Examples
55	Europe	EMBL
45	US	GenBank
6	Canada	DrugBank
6	Japan	DDBJ
2	Republic of Korea	miRGator
1	Taiwan Province of China	miRTarBase

Source: (Galperin et al., 2016).

5.1.2 Journal editors and granting bodies

Many journals require the submission of DSI generated in the course of research to be submitted to a recognized database as a condition of publication (GenBank; Noor *et al.*, 2006). Failure to do so forfeits acceptance and publication. There are many reasons behind journals adopting this policy and strongly held views amongst those in the scientific community to enforce it (Marx, 2012; Noor *et al.*, 2006).

Science magazine says that it "supports the efforts of databases that aggregate published data for the use of the scientific community" (Science). The kind of data includes, but is not limited to, microarray data, protein or DNA sequences, atomic coordinates or electron microscopy maps for macromolecular structures, ecological and climate data. In short, all the forms of data that have been highlighted in the vignette examples of this report. The minimum requirements are in accordance with MIBBI (Minimum Information for Biological and Biomedical Investigations) standards, which is a "curated, informative and educational resource on data and metadata *standards*, inter-related to *databases* and data *policies*" (MIBBI).

Research scientists, especially those in the public sector, depend on public funding in the form of salary, direct research support, grants and contracts to support their research, including generating DNA sequences and omics data. Often publication is necessary to meet employer expectations or conditions for advancement. This is captured in the phrase "publish or perish" (Rawat and Meena, 2014).

Therefore, for many public scientists at least, there is a *de facto* requirement that DSI arising in their work eventually must be made public, unless they have a contractual relationship with a private funder. Private sector actors do not need to disclose similar information unless they wish to publish.

Researchers in developing countries that lack the necessary infrastructure for ready access to DSI on GRFA may be reluctant to release data because of the fear that those with superior infrastructure will use and perhaps publish the information before they can (Elbe and Buckland-Merrett, 2017).

Like journals, public grant bodies can also put requirements on researchers to make public DSI arising from work completely or partially funded through them. For example, the US Department of Energy Joint Genome Institute provides funding for DNA sequencing and synthesis projects, but requires public release of the data including Sanger trace files, improved assemblies and preliminary automated annotations, “or other value-added data”.

These examples of journal and funding agency policies could have two different effects on researchers, especially those in the public sector. On the one hand, if access and use of DSI has fewer legal requirements than does the use of biological genetic material, and thus fewer compliance costs, more researchers may shift their work away from biological material to DSI, where they can. On the other hand, failure to establish access and benefit-sharing provisions for DSI in legislation might shift some private and public sector researchers away from generating or using DSI, because they have no way to meet what they perceive to be their obligations for fair and equitable sharing of benefits from traditional or sovereign knowledge.

A scientist that was part of a consortium of microbiologists from New Zealand revealed in an interview that he withdrew a 2015 proposal, “*The 100 Springs Metagenome Project: An analysis of the microbial community function of New Zealand’s geothermal ecosystems*”, submitted to the JGI Community Science Program because of JGI’s requirement to make all project data public. The project would have completed microbial metagenomic profiles of 100 geothermal springs in New Zealand, gathering unique information of international scientific, conservation and cultural value. The principle scientific investigator of the proposal was philosophically aligned with JGI’s requirements because he was comfortable with the notion that the funding was from the public and the data should be public, and was himself an advocate of the value of freely accessible DSI. However, he had also carefully and fastidiously developed relationships with the Māori, the indigenous peoples of New Zealand, with whom resided cultural and property rights over the majority of hotsprings from which samples were to be taken. His proposal was authorized by tribal leaders, but the JGI policy requirement was inconsistent with his commitment to the partnership. The governing groups for various Māori tribes were also reluctant to release such a vast description of indigenous knowledge without any specific provision for how benefits could be shared or without having a clear role in governance/use of DSI such that it aligned with their cultural values.

5.1.3 Private sector and private acting public institutions

Other users and generators of DSI are private companies and technology transfer or company incubator arms of public institutions, for example universities. A few companies dominate the market of building sequencing machines, but their business model has generally been focused on hardware supply, not service (Farr, 2016). There is speculation that this might change (Farr, 2016). However, even if it were to, it would be premature to speculate that these companies would also become oligopoly suppliers of DNA sequences much less other forms of DSI.

The goal of others generating DSI may be to establish intellectual property (IP) including for licensing income (Bagley, 2016). This may come from proprietary information or a mix of private and public information. For example, a genetic linkage map of *Coffea arabica* L. (coffee) including QTL for yield, plant height and bean size was developed using a combination of public (SOL Genomics Network) and proprietary (Cenicafe) databases (Moncada *et al.*, 2015). In addition to database content, tools for accessing or using the content can be proprietary. Private or proprietary databases that could hold critical information necessary to extract maximum value from public databases, are growing (Welch *et al.*, 2017).

Some instruments of IP can encourage transparency in DSI. However, transparency is not a requirement of all IP instruments, such as trade secrets (Tvedt and Young, 2007). Thus, accessibility and sharing of benefits may depend on current and future types and uses of IP instruments.

The entrepreneurial activities of public institutions including universities and government agencies might in general also threaten their abilities to provide public good through the use of

DSI. Historically, for example, universities could conduct research on inventions claimed by patents under the research exemption rule. However, court cases in the United States of America have begun to put restrictions on the exemption, noting that the business of a university was in some cases actually business, with the potential to use the exemption to gain unfair advantage (Heinemann, 2009; Mueller, 2003).

5.1.4 The public

“A farmer finds a high-tech solution: DNA testing. Thanks to his background in advanced biology, Thomas recognized that DNA testing could hold the key to distinguishing between Chinese and black truffles….By analyzing a small piece of truffle with it, Thomas can look for a telltale area of DNA that is present in the black but not the Chinese truffle. Now, besides cultivating his own truffles, he also acts as a consultant to other growers, selling them seedlings that have been tested to ensure they’re inoculated with the desirable fungus, assessing and sampling the growing truffles, and helping distribute and verify the valuable harvest” (Barnett, 2016).

The public are private actors but may engage with DSI for a variety of reasons not primarily including income or IP. They may be school children learning about bioinformatics or competing in science fairs (iGEM). They may be people satisfying their curiosity, wish to take a role in science communication, or may be looking for do-it-yourself options in the garden or on the farm.

The public is transforming from a user of DSI as an *in silico* product to a manufacturer of biological materials, including new genetic resources and DNA.

A number of products have already been developed for the public. BioBricks Foundation describes itself as “biotechnology in the public interest” (**BioBricks**). The company provides biological material derived from DSI in the form of DNA encoding functional modules that may be combined by users in different arrangements. An assembly kit can be purchased online, for example from the biotechnology company New England Biolabs (NEB). This company also sells other tools related to the use of BioBricks, such as bacteria into which the recombinant DNA can be inserted. Bacteria that receive the DNA reveal themselves because they are transformed into being resistant to an antibiotic through a linked marker (iGEM). Widescale adoption by people untrained in the use and proper disposal of antibiotic-resistant bacteria could also have biosafety implications. Sewage water treatment, for example, does not destroy antibiotic resistance genes that have the potential to be acquired by other bacteria (Wang *et al.*, 2017a).

One use of machines such as the digital-to-biological converter is to satisfy an anticipated market for home-based manufacturing (Boles *et al.*, 2017). Such machines are expected to shrink in size and cost, aided by developments in microfluidics. Inputting DSI, the home user would be able to synthesize not just DNA, but potentially a variety of different macromolecules including everything needed to print their own virus or antibodies (Boles *et al.*, 2017). However, any checks and balances that might be put in place when ordering synthesized DNA molecules might be lost when home synthesis becomes available.

5.2 Access

“Access to online data has become a basic requirement for conducting scientific research, but the growth in data, databases, websites and resources has outpaced the development of mechanisms and models to fund the necessary cyberinfrastructure, curation and long-term stewardship of these resources...a single, viable framework for sustainable and long-term stewardship of data and resources has not emerged” (Bastow and Leonelli, 2010).

DSI is a fundamental component of the characterization, conservation and sustainable use of GRFA. Any instability in the infrastructure or impediment of access to DSI could therefore undermine efforts of nations, private and public institutions and multinational agencies to fulfill provisions of conservation and sustainable use of GRFA or other provisions to the benefit of human well-being.

Unless DSI on GRFA is maintained and remains easily accessible to all, it will be difficult to ensure the fair and equitable sharing of the benefits arising from its use. Without adequate confidence in the systems of data preservation, dissemination and maintenance, those who create DSI may be increasingly reluctant to share it.

An important warning on this point comes from the development of GISAID. Although GISAID content is not GRFA, the issues faced are not special to GISAID. “While the genetic makeup and the necessary associated data of the different viruses are distinct requiring separate databases/compartments for unambiguous analysis, *the modi operandi for sharing genetic data are generic*” (emphasis added to Shu and McCauley, 2017). The GISAID database provides a global public good, but initially there were problems getting countries and individuals to trust in it. “In 2006, the reluctance of data sharing, in particular of avian H5N1 influenza viruses, created an emergency bringing into focus certain limitations and inequities, such that the World Health Organization (WHO)’s Global Influenza Surveillance Network (now the Global Influenza Surveillance and Response System (GISRS)) was criticised on several fronts, including limited global access to H5N1 sequence data that were stored in a database hosted by the Los Alamos National Laboratories in the United States” (Shu and McCauley, 2017).

Researchers (Elbe and Buckland-Merrett, 2017) have identified three obstacles to sharing data via GISAID, which can be extrapolated to participation in other databases for sharing genetic data on GRFA.

- Science, publications and recognition:

The competive nature of science and winning funding creates a perverse incentive to withhold data. “In a context where the standing of scientists, and the research income they can generate, is heavily linked to their publications, citations, and scientific reputations, there is pressure to be the “first” to publish findings…scientists are concerned that sharing such information in an open and timely manner might enable others to publish findings with their data more quickly than they themselves could” (Elbe and Buckland-Merrett, 2017).

- Governments and trade and access to medicine (or veterinary medicines or vaccines):

Governments can have a variety of reasons to withhold data (Elbe and Buckland-Merrett, 2017). These include such things as concerns about reptuational damage (e.g. as the source of a pathogen, perhaps the home of a company that uses an endangered species in a product) and IP considerations.

- Practical challenges such as confidence in database viability:

Existing and future databases require support (Bastow and Leonelli, 2010; Elbe and Buckland-Merrett, 2017; Hayden, 2016). Presently this comes from those who generate DSI and choose to send it to a database, and actors providing financial resources to maintain infrastructure, reliability and protect against tampering. Future support is not guaranteed (Editor, 2016; Elbe and Buckland-Merrett, 2017).

The 16-year-old Kyoto Encyclopedia of Genes and Genomes (KEGG) was under threat of closing in 2011 due to funding insecurity. The US National Science Foundation withdrew funding from the primary database specific to information from the plant family *Arabidopsis* (TAIR) that had accumulated data for 14 years. In 2016 the Human Genome Research Institute announced plans to reduce support for five “model organism” databases from USD17.6 million by 30–40 percent (Hayden, 2016).

Optimistically, DSI can be salvaged by transfer to other more financially viable platforms as necessary. Fiscal uncertainty is therefore not in principle a threat to the information or an existential threat to databases. However, neither the hardware nor the software of DSI is static. Just like it is now impossible to use a floppy disk in any presently common computer, the way in which DSI is archived, transmitted and manipulated changes over time. Because different databases and individual DSI providers may use different standards, resurrecting information across systems may not always be possible (Editor, 2016). Loss of continuity might put some DSI at risk of extinction.

5.3 Conclusion

DSI has become an intensely debated topic in the context of fair and equitable access and the sharing of benefits of GRFA. The study attempts to inform the debate by clarifying the diversity of relevant actors. This includes those groups most connected to the operational use and generation of DSI.

The different future ways that DSI is envisioned to be used will be by more people, and be used in more ways. Many of the new ways will be disconnected from the biological *raw material*. In particular, a larger role for the public will emerge as technologies combine to increase distributed data gathering capability (Jain *et al.*, 2018), and customized commercial products can be directly marketed to individuals collecting data.

Whether or not ABS legislation includes the use of DSI on GRFA for the conservation and sustainable use of GRFA, including exchange, access and the fair and equitable sharing of the benefits arising from their use, the research and commercial sectors may react in unexpected ways. Some fear that any additional compliance or subscription costs may slow the distribution of assets for upstream research (Manzella, 2016) as the information is either not generated or it is kept secret.

However, as recorded in this study, failure of legislation to provide a framework for ABS might have similar effects. Researchers and businesses may prioritize the value of DSI to country of origin, or to indigenous peoples, and either keep the data secret or abandon its collection altogether.

VI. CONCLUSION

To get value from a genetic resource once required possessing it, or at least its DNA. This is no longer the case.

With increases in the global capacity to generate DNA sequences and to share them, DSI is sufficient to make significant use of the information from GRFA for characterization of biodiversity and improvement of GRFA through breeding. For many, access to this information does not require access to GRFA.

Moreover, using only DSI it is possible today or in the foreseeable future to obtain, transfer and re-construct significant parts of the genetic information from organisms without having possessed the organism or transferred its DNA from one place to another.

The objective of this report is to assist with discussions about whether, and in what ways, DSI now or in the future may substitute for genetic material to do research and development on genetic resources for food and agriculture, and to create value from genetic resources through DSI.

The purpose of the scoping study was to bring together the terminology being used in what is referred to as DSI, reflect on how it is being used in biotechnology or on genetic resources for food and agriculture or may be used in the foreseeable future, and what DSI might look like in the future.

(1) A variety of terms that differ between actors are used in discussions on DSI. Terms include but are not limited to genetic sequence data, genetic information, dematerialized use of genetic resources, *in silico* utilization, and bioinformatics. Actors are those participating in discussions within international agreements, different kinds of biology-based scientists, the biotechnology and agricultural industries, farmers and the public.

What emerges from the landscape of terms and associated examples is that what an actor may mean when referring to DSI on GRFA is much more than an electronic equivalent of a sequence of nucleotides as could be written on a whiteboard. Seemingly different kinds of data are routinely combined to provide information on GRFA. Even the most basic and ubiquitous conception of DSI as an electronic DNA database already combines multiple forms of data that in everyday practice must be used together.

DSI on GRFA, whether it be DNA sequence information or any of the other kinds covered in this report, have characteristics in common. In some situations special to each, DSI that is not DNA sequence information may have the power to replace DNA sequence information. Almost certainly, these other kinds of information are indispensible for making full use of DNA sequence information and are part of the value adding activities of DSI.

This links the view of DSI that it is DNA sequences only, to views about it being genetic information, tangible or intangible, or it being that which can be utilized. Optimistically, the characteristics that all terms have in common might both form the basis for determining what DSI is relevant to GRFA, and for DSI to be defined in a way that allows it to be managed, constructively and appropriately, as envisioned in international agreements.

(2) What is evident from the series of vignettes is that DSI can be and is being generated from all kinds of GRFA: microbes, plants and animals. It contributes to conservation and a variety of value chain activities that rely upon taxonomic description, trait identification, breeding, certification, raw materials and new products. The examples are distributed across microbes, plants and animals to reveal the sector independence of how much DSI is applied.

(3) The value of a genetic resource is also no longer restricted to its biology. The scale and speed of information gathering about organisms can generate future uses and revenue independently of the organism that originally provided the genetic material. DSI has achieved the scale of what is called “big data”. The use of big data can link value to GRFA quite distinct from possession, use or management of GRFA, while also providing assistance and products to those who do possess, use and manage GRFA.

(4) DSI will be used by more people and organizations in the future. However, that may not ensure either equitable access to benefits, or fair opportunity to benefit from DSI. In this evolving role, DSI would augment capacity in GRFA management, but at least some of the technologies it supports, such as gene drives and certain kinds of pesticides, are controversial and potentially damaging too. The use of DSI will also require considerable scientific and infrastructure capacity. Those lacking such capacity might therefore, at least in the short term, benefit less from DSI.

REFERENCES

Abdelrahman, M., Burritt, D.J., and Tran, L.-S.P. 2017. The use of metabolomic quantitative trait locus mapping and osmotic adjustment traits for the improvement of crop yields under environmental stresses. Seminars Cell Develop Biol, S1084-9521.

Abels, S.G., and Khisamutdinov, E.F. 2015. Nucleic acid computing and its potential to transform silicon-based technology. DNA and RNA Nanotechnology *2*, 13-22.

Ader, D., Li, Z., Shah, H.R., Tao, M., Wang, D., and Yang, H. 2011. Methods and compositions for weed control. https://patents.google.com/patent/US20130288895A1/en.

Al-Aamri, A., Taha, K., Al-Hammadi, Y., Maalouf, M., and Domouz, D. 2017. Constructing genetic networks using biomedical literature and rare event classification. Sci Rep *7*, 15784.

Alonso, C., Perez, R., Bazaga, P., Medrano, M., and Herrera, C.M. 2017. Within-plant variation in seed size and inflorescence fecundity is associated with epigenetic mosaicism in the shrub Lavandula latifolia (Lamiaceae). Ann Bot *121*, 153–160.

Altamura, L., Horvath, C., Rengaraj, S., Rongier, A., Elouarzaki, K., Gondran, C., Maçon, A.L.B., Vendrely, C., Bouchiat, V., Fontecave, M., *et al.* 2017. A synthetic redox biofilm made from metalloprotein–prion domain chimera nanowires. Nat Chem *9*, 157-163.

Amarnarth, V., and Broom, A.D. 1977. Chemical synthesis of oligonucleotides. Chem Rev *77*, 183-217.

Anonymous. English Oxford Living Dictionaries https://en.oxforddictionaries.com/definition/information.

Antonets, K.S., and Nizhnikov, A.A. 2017. Amyloids and prions in plants: facts and perspectives. Prion *11*, 300-312.

Avery, O.T., MacLeod, C.M., and McCarty, M. 1944. Studies on the chemical nature of the substance inducing transformation of Pneumococcal types. J Exp Med *79*, 137-158.

Bagley, M.A. 2016. Digital DNA: The Nagoya Protocol and synthetic biology research. Public Law and Legal Theory Research Paper Series 2016 University of Virginia Law School 11

Bagley, M.A., and Rai, A.K. 2013. The Nagoya Protocol and synthetic biology research: a look at the potential impacts. (Washington, DC) T Wilson Center http://scholarship.law.duke.edu/faculty_scholarship/3230.

Bah, C.S.F., Bekhit, A.E.-D., A., C., and McConnell, M.A. 2013. Slaughterhouse blood: an emerging source of bioactive compounds. Compr Rev Food Sci Food Saf *12*, 314-331.

Barbieri, M. 2018. What is code biology? Biosys *164*, 1-10.

Barnett, R. One unexpected answer: grow better truffles. https://ideas.ted.com/what-can-you-do-with-a-home-dna-machine-one-unexpected-answer-grow-better-truffles/. Access date, 3 February 2018

Bastow, R., and Leonelli, S. 2010. Sustainable digital infrastructure. EMBO Rep *11*, 730-734.

BioBricks. https://biobricks.org. Access date, 29 November 2017

Boggess, M.V., Lippolis, J.D., Hurkman, W.J., Fagerquist, C.K., Briggs, S.P., Gomes, A.V., Righetti, P.G., and Bala, K. 2013. The need for agriculture phenotyping: "Moving from genotype to phenotype". J Proteomics *93*, 20-39.

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

Bräutigam, K., Vining, K.J., Lafon-Placette, C., Fossdal, C.G., Mirouze, M., Marcos, J.G., Fluch, S., Fraga, M.F., Guevara, M.Á., Abarca, D., *et al.* 2013. Epigenetic regulation of adaptive responses of forest tree species to the environment. Ecol Evol *3*, 399-415.

Brbić, M., Piškorec, M., Vidulin, V., Kriško, A., Šmuc, T., and Supek, F. 2016. The landscape of microbial phenotypic traits and associated genes. Nuc Acid Res *44*, 10074-10090.

Brouard, J.-S., Boyle, B., Ibeagha-Awemu, E.M., and Bissonnette, N. 2017. Low-depth genotyping-by-sequencing (GBS) in a bovine population: strategies to maximize the selection of high quality genotypes and the accuracy of imputation. BMC Genet *18*, 32.

Buckeridge, M.S. 2018. The evolution of the Glycomic Codes of extracellular matrices. Biosys *164*, 112-120.

Burnham, C.-A.D., Leeds, J., Nordmann, P., O'Grady, J., and Patel, J. 2017. Diagnosing antimicrobial resistance. Nat Rev Microbiol *15*, 697-703.

Busby, P.E., Soman, C., Wagner, M.R., Friesen, M.L., Kremer, J., Bennett, A., Morsy, M., Eisen, J.A., Leach, J.E., and Dangl, J.L. 2017. Research priorities for harnessing plant microbiomes in sustainable agriculture. PLoS Biol *15*, e2001793.

Campanaro, S., Treu, L., Cattani, M., Kougias, P.G., Vendramin, V., Schiavon, S., Tagliapietra, F., Giacomini, A., and Corich, V. 2017. In vitro fermentation of key dietary compounds with rumen fluid: A genome-centric perspective. Sci Total Environ *584-585*, 683-691.

Carrico, J.A., Sabat, A.J., Friedrich, A.W., and Ramirez, M. 2013. Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the next-generation sequencing revolution. Euro Surveill *18*, 20382.

CBD. 2018. 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. AHTEG on Digital Sequence Information on Genetic Resources (Montreal)CBD/DSI/AHTEG/2018/1/2/Add.1

CGRFA. The International Network of Ex Situ Collections under the Auspices of FAO. http://www.fao.org/nr/cgrfa/cgrfa-about/cgrfa-history/cgrfa-internnet/en/. Access date, 3 December 2017

CGRFA. 2012. Cryoconservation of Animal Genetic Resources. FAO Animal Production and Health Guidelines No 12 UN FAO http://www.fao.org/docrep/016/i3017e/i3017e00.pdf.

CGRFA. 2014. Genebank Standards for Plant Genetic Resources for Food and Agriculture. UN FAO http://www.fao.org/3/a-i3704e.pdf.

CGRFA. 2017. Sixteenth Regular Session of the Commission on Genetic Resources for Food and Agriculture. CGRFA-16/17/Report

Chari, R., and Church, G.M. 2017. Beyond editing to writing large genomes. Nat Rev Genet.

Chaudhry, Q., and Castle, L. 2011. Food applications of nanotechnologies: an overview of opportunities and challenges for developing countries. Trend Food Sci Technol *22*, 595-603.

Chen, H., and Yada, R. 2011. Nanotechnologies in agriculture: new tools for sustainable development. Trend Food Sci Technol *22*, 585-594.

Chen, W., Yu, H., Sun, F., Ornob, A., Brisbin, R., Ganguli, A., Vemuri, V., Strzebonski, P., Cui, G., Allen, K.J., *et al.* 2017. Mobile platform for multiplexed detection and differentiation of disease-specific nucleic acid sequences, using microfluidic loop-mediated isothermal amplification and smartphone detection. Anal Chem *89*, 11219-11226.

Chernoff, Y.O. 2004. Replication vehicles of protein-based inheritance. Trends Biotechnol *22*, 549-552.

Chin, J.W. 2017. Expanding and reprogramming the genetic code. Nature *550*, 53-60.

Chu, H. 2017. Host gene–microbiome interactions: molecular mechanisms in inflammatory bowel disease. Genome Med *9*, 69.

Clarke, R. 2010. Private food safety standards: their role in food safety regulation and their impact. (U. FAO) UN FAO http://www.fao.org/docrep/016/ap236e/ap236e.pdf.

Contreras, J.L., and Deshmukh, V.G. 2017. Development of the personal genomics industry. In Genetics, Ethics and Education, S.e.a. Bouregy, ed. (Cambridge University Press).

Costa, F.F. 2014. Big data in biomedicine. Drug Dis Today *19*, 433-440.

De Luca, D., Cennamo, P., Del Guacchio, E., Di Novella, R., and Caputo, P. 2017. Conservation and genetic characterisation of common bean landraces from Cilento region (southern Italy): high di erentiation in spite of low genetic diversity. Genetica *146*, 29-44.

de Tarso, S.G.S., Oliveira, D., and Afonso, J.A.B. 2016. Ruminants as part of the global food system: how evolutionary adaptations and diversity of the digestive system brought them to the future. J Dairy Vet Anim Res *3*, 00094.

Di Pinto, A., Marchetti, P., Mottola, A., Bozzo, G., Bonerba, E., Ceci, E., Bottaro, M., and Tantillo, G. 2015. Species identification in fish fillet products using DNA barcoding. Fish Res *170*, 9-13.

DiEuliis, D., Johnson, K.R., Morse, S.S., and Schindel, D.E. 2016. Specimen collections should have a much bigger role in infectious disease research and response. Proc Natl Acad Sci USA *113*, 4-7.

Dillon, P., Choi, G., and Welch, R. 2004. Nucleotide sequence of Escherichia coli pathogenicity islands. https://www.google.com/patents/US6787643.

Donner, A. 2013. Synthetic influenza seeds. SciBX, 509.

Edgar, R.S., and Wood, W.B. 1966. Morphogenesis of bacteriophage T4 in extracts of mutant-infected cells. Proc Natl Acad Sci USA *55*, 498-505.

Editor. 2016. Database under maintenance. Nat Methods *13*, 699.

Egelie, K.J., Graff, G.D., Strand, S.P., and Johansen, B. 2016. The emerging patent landscape of CRISPR-Cas gene editing technology. Nat Biotechnol *34*, 1025-1031.

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

Esvelt, K.M., and Gemmell, N.J. 2017. Conservation demands safe gene drive. PLoS Biol *15*, e2003850.

European Food Safety, A., European Centre for Disease, P., and Control. 2017. The European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2015. EFSA J *15*, e04694-n/a.

FAO. 2013. Report of the Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture Fifth Session. (Rome) UN FAO IT/GB-5/13/4

FAO. 2017. FAO and the SDGs. Indicators: measuring up to the 2030 Agenda for Sustainable Development. UN FAO http://www.fao.org/3/a-i6919e.pdf.

Farr, C. 2016. Illumina, secret giant of DNA sequencing, is bringing tis tech to the masses. In FastCompany.

Fierer, N. 2017. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat Rev Microbiol *15*, 579-590.

FIFRA. 2014. RNAi Technology as a Pesticide: Program Formulation for Human Health and Ecological Risk Assessment. United States Environmental Protection Agency http://www.epa.gov/scipoly/sap/meetings/2014/january/012814minutes.pdf.

Filonzi, L., Chiesa, S., Vaghi, M., and Nonnis Marzano, F. 2010. Molecular barcoding reveals mislabelling of commercial fish products in Italy. Food Res Internatl *43*, 1383-1388.

Fontanesi, L., Martelli, P.L., Beretti, F., Riggio, V., Dall'Olio, S., Colombo, M., Casadio, R., Russo, V., and Portolano, B. 2010. An initial comparative map of copy number variations in the goat (Capra hircus) genome. BMC Genomics *11*, 639.

Fortes, A.M., and Gallusci, P. 2017. Plant Stress Responses and Phenotypic Plasticity in the Epigenomics Era: Perspectives on the Grapevine Scenario, a Model for Perennial Crop Plants. Front Plant Sci *8*, 82.

Fowler, C. 2016. Seeds on Ice (Westport CT and New York, NY: Prospecta Press)

Fromm, M.E. 2017. Similar performance from seeds with epigenetic traits. US 20170223914 A1, https://www.google.com/patents/US20170223914.

Gabius, H.-J. 2018. The sugar code: why glycans are so important. Biosys *164*, 102-111.

Gallagher, R.R., Li, Z., Lewis, A.O., and Isaacs, F.J. 2014. Rapid editing and evolution of bacterial genomes using libraries of synthetic DNA. Nat Protoc *9*, 2301-2316.

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

Galperin, M.Y., Fernández-Suárez, X.M., and Rigden, D.J. 2016. The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. Nucleic Acids Res *45*, D1-D11.

García-Sancho, M. 2015. Genetic information in the age of DNA sequencing. Information and Culture: A Journal of History *50*, 110-142.

GenBank. https://www.ncbi.nlm.nih.gov/genbank/. Access date, 25 November 2017

GenBank. How to submit data to GenBank. https://www.ncbi.nlm.nih.gov/genbank/submit/. Access date, 7 December 2017

Gerber, P.J., Steinfeld, H., Henderson, B., Mottet, A., Opio, C., Dijkman, J., Falcucci, A., and Tempio, G. 2013. Tackling climate change through livestock - a global assessment of emissions and mitigation. UN FAO http://www.fao.org/docrep/018/i3437e/i3437e.pdf.

Ghatak, A., Chaturvedi, P., and Weckwerth, W. 2017. Cereal crop proteomics: systemic analysis of crop drought stress responses towards marker-assisted selection breeding. Front Plant Sci *8*.

Gibson, D.G. 2014. Programming biological operating systems: genome design, assembly and activation. Nat Method *11*, 521-526.

GLIS. 2015. First Meeting of the Expert Consultation on the Global Information System on Plant Genetic Resources for Food and Agriculture. UN FAO http://www.fao.org/3/a-be664e.pdf.

Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C., and Gelbart, W.M. 2000. DNA: The genetic material. In An Introduction to Genetic Analysis (W.H. Freeman).

Hackmann, T.J., and Spain, J.N. 2010. Invited review: ruminant ecology and evolution: perspectives useful to ruminant livestock research and production. J Dairy Sci *93*, 1320-1334.

Haile-Mariam, M., Pryce, J.E., Schrooten, C., and Hayes, B.J. Including overseas performance information in genomic evaluations of Australian dairy cattle. J Dairy Sci *98*, 3443-3459.

Hamada, D., Yanagihara, I., and Tsumoto, K. 2004. Engineering amyloidogenicity towards the development of nanofibrillar materials. Trends Biotechnol *22*, 93-97.

Hammer, K., Knupffer, H., Xhuveli, L., and Perrino, P. 1996. Estimating genetic erosion in landraces — two case studies. Genet Res Crop Evol *Genetic Resources and Crop Evolution*, 329-336.

Hauser, C.A.E., Maurer-Stroh, S., and Martins, I.C. 2014. Amyloid-based nanosensors and nanodevices. Chem Soc Rev *43*, 5326-5345.

Hayden, E.C. 2014. Synthetic-biology firms shift focus. Nature *505*, 598.

Hayden, E.C. 2016. Funding for model-organism databases in trouble. In Nature News (Nature).

Haynes, K.A., Broderick, M.L., Brown, A.D., Butner, T.L., Dickson, J.O., Harden, W.L., Heard, L.H., Jessen, E.L., Malloy, K.J., Ogden, B.J., *et al.* 2008. Engineering bacteria to solve the Burnt Pancake Problem. *2*, 8.

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., and Li, Z. 2014. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Pl Sci *5*.

Heinemann, J.A. 1999. Genetic evidence of protein transfer during bacterial conjugation. Plasmid *41*, 240-247.

Heinemann, J.A. 2004. Challenges to regulating the industrial gene: Views inspired by the New Zealand experience. In Challenging Science: Science and Society Issues in New Zealand, K. Dew, and R. Fitzgerald, eds. (Dunedin: Dunmore).

Heinemann, J.A. 2007. A typology of the effects of (trans)gene flow on the conservation and sustainable use of genetic resources. Background Study Paper (Rome) UN FAO Bsp35rev1, 94 ftp://ftp.fao.org/ag/cgrfa/bsp/bsp35r1e.pdf.

Heinemann, J.A. 2009. Hope not Hype. The future of agriculture guided by the International Assessment of Agricultural Knowledge, Science and Technology for Development (Penang: Third World Network)

Heinemann, J.A., Agapito-Tenfen, S.Z., and Carman, J.A. 2013. A comparative evaluation of the regulation of GM crops or products containing dsRNA and suggested improvements to risk assessments. Environ Int *55*, 43-55.

Heinemann, J.A., Kurenbach, B., and Quist, D. 2011. Molecular profiling — a tool for addressing emerging gaps in the comparative risk assessment of GMOs. Env Int *37*, 1285-1293.

Heinemann, J.A., Sparrow, A.D., and Traavik, T. 2004. Is confidence in monitoring of GE foods justified? Trends Biotechnol *22*, 331-336.

Heisey, P.W., and Rubenstein, K.D. 2015. Using Crop Genetic Resources to Help Agriculture Adapt to Climate Change: Economics and Policy. USDA EIB-139 http://www.ers.usda.gov/publications/eib-economic-information-bulletin/eib139.

Helmy, M., Awad, M., and Mosa, K.A. 2016. Limited resources of genome sequencing in developing countries: challenges and solutions. Appl Transl Genom *9*, 15-19.

Henderson, G., Cox, F., Ganesh, S., Jonker, A., Young, W., Census, G.R., and Janssen, P.H. 2015. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. Sci Rep *5*, 14567.

Higgins, D., Pal, C., Sulaiman, I.M., Jia, C., Zerwekh, T., Dowd, S.E., and Banerjee, P. 2018. Application of high-throughput pyrosequencing in the analysis of microbiota of food commodities procured from small and large retail outlets in a U.S. metropolitan area – A pilot study. Food Res Internatl *105*, 29-40.

Honee, G., Vriezen, W., and Visser, B. 1990. A translation fusion product of two different insecticidal crystal protein genes of Bacillus thuringiensis exhibits an enlarged insecticidal spectrum. Appl Environ Microbiol *56*, 823-825.

Hope, J. 1995. Mice and beef and brain diseases. Nature *378*, 761-762.

IAASTD, ed. 2009. International Assessment of Agricultural Knowledge, Science and Technology for Development (Washington, D.C.: Island Press).

Ibeagha-Awemu, E.M., Peters, S.O., Akwanji, K.A., Imumorin, I.G., and Zhao, X. 2016. High density genome wide genotyping-by-sequencing and association identifies common and low frequency SNPs, and novel candidate genes influencing cow milk traits. Sci Rep *6*, 31109.

iGEM. About iGEM. http://2017.igem.org/About. Access date, 18 February 2018

iGEM. Registry of standard biological parts. http://parts.igem.org/Help:An_Introduction_to_BioBricks. Access date, 29 November 2017

ITPGRFA. International Treaty on Plant Genetic Resources in Food and Agriculture. ftp://ftp.fao.org/docrep/fao/011/i0510e/i0510e01.pdf.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., ***et al.*** 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol.

Jayawardena, N., Kaur, M., S., N., Malmstrom, J., Goldstone, D., Negron, L., Gerrard, J.A., and Domigan, L.J. 2017. Amyloid fibrils fro hemoglobin. Biomolecules *7*, 37.

Jenuwein, T., and Allis, C.D. 2001. Translating the histone code. Science *293*, 1074-1080.

Jewett, M.C., and Forster, A.C. 2010. Update on designing and building minimal cells. Curr Opin Biotechnol *21*, 697-703.

Jin, L., Mao, K., Li, J., Huang, W., Che, T., Fu, Y., Tang, Q., Liu, P., Song, Y., Liu, R., ***et al.*** 2018. Genome-wide profiling of gene expression and DNA methylation provides insight into low-altitude acclimation in Tibetan pigs. Gene *642*, 522-532.

Kaiser, D., and Masuda, T. 1973. In vitro assembly of bacteriophage Lambda heads. Proc Natl Acad Sci USA *70*, 260-264.

Kamen, M. Governments shut down the internet more than 50 times in 2016. http://www.wired.co.uk/article/over-50-internet-shutdowns-2016

Khalil, A.S., and Collins, J.J. 2010. Synthetic biology: applications come of age. Nat Rev Genet *11*, 367-379.

Kim, J., Hanotte, O., Mwai, O.A., Dessie, T., Bashir, S., Diallo, B., Agaba, M., Kim, K., Kwak, W., Sung, S., ***et al.*** 2017. The genome landscape of indigenous African cattle. Genome Biol *18*, 34.

King, C.-Y., and Diaz-Avalos. 2004. Protein-only transmission of three yeast prion strains. Nature *428*, 319-323.

King, T., Cole, M., Farber, J.M., Eisenbrand, G., Zabaras, D., Fox, E.M., and Hill, J.P. 2017. Food safety for food security: relationship between global megatrends and developments in food safety. Trend Food Sci Technol *68*, 160-175.

Knight-Jones, T.J.D., and Rushton, J. 2013. The economic impacts of foot and mouth disease – what are they, how big are they and where do they occur? Prev Vet Med *112*, 161-173.

Knowles, T.P.J., and Buehler, M.J. 2011. Nanomechanics of functional and pathological amyloid materials. Nat Nanotechnol *6*, 469-479.

Koopaee, H.K., and Koshkoiyeh, A.E. 2014. SNPs genotyping technologies and their applications in farm animals breeding programs: review. Braz Arch Biol Technol *57*, 87-95.

Kotula, J.W., Kerns, S.J., Shaket, L.A., Siraj, L., Collins, J.J., Way, J.C., and Silver, P.A. 2014. Programmable bacteria detect and record an environmental signal in the mammalian gut. Proc Natl Acad Sci USA *111*, 4838-4843.

Lamas-Toranzo, I., Guerrero-Sánchez, J., Miralles-Bover, H., Alegre-Cid, G., Pericuesta, E., and Bermejo-Álvarez, P. 2017. CRISPR is knocking on barn door. Reprod Dom Anim *52(Suppl. 4)*, 39-47.

Larson, W.A., Seeb, L.W., Everett, M.V., Waples, R.K., Templin, W.D., and Seeb, J.E. 2014. Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (Oncorhynchus tshawytscha). Evol Appl *7*, 355-369.

Lau, J.A., Lennon, J.T., and Heath, K.D. 2017. Trees harness the power of microbes to survive climate change. Proc Natl Acad Sci USA *114*, 11009-11011.

Le Duc, D., Renaud, G., Krishnan, A., Almén, M.S., Huynen, L., Prohaska, S.J., Ongyerth, M., Bitarello, B.D., Schiöth, H.B., Hofreiter, M., *et al.* 2015. Kiwi genome provides insights into evolution of a nocturnal lifestyle. Genome Biol *16*, 147.

Li, E., Beard, C., and Jaenisch, R. 1993. Role for DNA methylation in genomic imprinting. Nature *366*, 362-365.

Li, M., Wu, H, Wang, T, Xia, Y, Jin, L, Jiang, A, Zhu, L, Chen, L, Li, R, Li, X. 2012. Co-methylated genes in different adipose depots of pig are associated with metabolic, inflammatory and immune processes. Int J Biol Sci *8*, 831-837.

Lidder, P., and Sonnino, A. 2011. Background study paper no. 52. Biotechnologies for the management of genetic resources for food and agricuture. Commission on Genetic Resources for Food and Agriculture

López-Uribe, M.M., Soro, A., and Jha, S. 2017. Conservation genetics of bees: advances in the application of molecular tools to guide bee pollinator conservation. Conserv Genet *18*, 501-506.

Lozier, J.D., and Zayed, A. 2017. Bee conservation in the age of genomics. Conserv Genet *18*, 713-729.

Luca, F., Kupfer, S.S., Knights, D., Khoruts, A., and Blekhman, R. 2018. Functional genomics of host-microbiome interactions in humans. Trend Genet *34*, 30-40.

Lundgren, J.G., and Duan, J.J. 2013. RNAi-based insecticidal crops: potential effects on nontarget species. Biosci *63*, 657-665.

Manheim, B.S. 2016. Regulation of synthetic biology under the Nagoya Protocol. Nat Biotechnol *34*, 1104-1105.

Manzella, D. 2016. Background study paper n. 10. The global information system and genomic information: transparency of rights and obligations. IT/GB7/SAC-1/16/ BSP 10

Mardis, E.R. 2017. DNA sequencing technologies: 2006-2016. Nat Protoc *12*, 213-218.

Marijuán, P.C., Navarro, J., and del Moral, R. 2018. How prokaryotes ‘encode’ their environment: Systemic tools for organizing the information flow. Biosys *164*, 26-38.

Marraffini, L.A., and Sontheimer, E.J. 2010. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet *11*, 181-190.

Marx, V. 2012. My data are your data. Nat Biotechnol *30*, 509-511.

Marx, V. 2013. The big challenges of big data. Nature *255*, 256-260.

Mauchline, T.H., and Malone, J.G. 2017. Life in earth – the root microbiome to the rescue? Curr Opin Microbiol *37*, 23-28.

McCann, H.C., Rikkerink, E.H.A., Bertels, F., Fiers, M., Lu, A., Rees-George, J., Andersen, M.T., Gleave, A.P., Haubold, B., Wohlers, M.W., *et al.* 2013. Genomic analysis of the kiwifruit pathogen Pseudomonas syringae pv. actinidiae provides insight into the origins of an emergent plant disease. PLOS Path *9*, e1003503.

McSweeney, C., and Mackie, R. 2012. Micro-organisms and ruminant digestion: state of knowledge, trends and future prospects. Background Study Paper CGRFA

Mdladla, K., Dzomba, E.F., and Muchadeyi, F.C. 2017. The potential of landscape genomics approach in the characterization of adaptive genetic diversity in indigenous goat genetic resources: a South African perspective. Small Rumin Res *150*, 87-92.

MIBBI. http://mibbi.sourceforge.net/foundry.shtml. Access date, 7 December 2017

Mogren, C.L., and Lundgren, J.G. 2017. In silico identification of off-target pesticidal dsRNA binding in honey bees (Apis mellifera). PeerJ *5*, e4131.

Moncada, M.D.P., Tovar, E., Montoya, J.C., González, A., Spindel, J., and McCouch, S. 2015. A genetic linkage map of coffee (Coffea arabica L.) and QTL for yield, plant height, and bean size. Tree Genet Genome *12*, 5.

Moore, A. 2000. Owning genetic information and gene enhancement techniques: why privacy and property rights may undermine social control of the human genome. Bioethics *14*, 97-119.

Morimoto, J., Simpson, S.J., and Ponton, F. 2017. Direct and trans-generational effects of male and female gut microbiota in Drosophila melanogaster. Biol Lett *13*, 20160966.

Moriya, Y., Yamada, T., Okuda, S., Nakagawa, Z., Kotera, M., Tokimatsu, T., Kanehisa, M., and Goto, S. 2016. Identification of Enzyme Genes Using Chemical Structure Alignments of Substrate–Product Pairs. J Chem Inf Model *56*, 510-516.

Mueller, J.M. 2003. Public access versus proprietary rights in genomic information: what is the proper role of intellectual property rights? J Health Care L & Pol'y *222*, 240.

Muller, H.J., and Prokofyeva, A.A. 1935. The individual gene in relation to the chromomere and the chromosome. Proc Natl Acad Sci USA *21*, 16-26.

NAR. NAR Database Summary Paper Alphabetic List. http://www.oxfordjournals.org/our_journals/nar/database/a/. Access date, 28 November 2017

Navarro, S.X. 2014. Methods and Compositions for Weed Control Using EPSPS Polynucleotides. WO2015108982A8, https://patents.google.com/patent/US20160330967A1/en.

NCB. 2016. Genome editing an ethical review. (N.C.o. Bioethics) Nuffield Council on Bioethics

NEB. BioBrick Assembly Kit. https://www.neb.com/products/e0546-biobrick-assembly-kit - Product%20Information. Access date, 29 November 2017

Neves, M.P., and Druml, C. 2017. Ethical implications of fighting malaria with CRISPR/Cas9. BMJ Glob Health *2*, e000396.

NIFA-NSF. 2011. Phenomics: genotype to phenotype. USDA and NSF https://www.nsf.gov/bio/pubs/reports/phenomics_workshop_report.pdf.

NIH. The Cost of Sequencing a Human Genome. https://www.genome.gov/sequencingcosts/. Access date, 23 November 2017

Niu, Y., Jin, M., Li, Y., Li, P., Zhou, J., Wang, X., Petersen, B., Huang, X., Kou, Q. and Chen, Y. . 2017. Biallelic β-carotene oxygenase 2 knockout results in yellow fat in sheep via CRISPR/Cas9. Anim Genet *48*, 242–244.

NLM. Bioinformatics. https://www.nlm.nih.gov/tsd/acquisitions/cdm/subjects12.html. Access date, 26 November 2017

Noor, M.A.F., Zimmerman, K.J., and Teeter, K.C. 2006. Data Sharing: How Much Doesn't Get Submitted to GenBank? PLoS Biol *4*, e228.

Norelli, J.L., Wisniewski, M., Fazio, G., Burchard, E., Gutierrez, B., Levin, E., and Droby, S. 2017. Genotyping-by-sequencing markers facilitate the identification of quantitative trait loci controlling resistance to Penicillium expansum in Malus sieversii. PLoS One *12*, e0172949.

Novick, A., and Weiner, M. 1957. Enzyme induction as an all-or-none phenomenon. Proc Natl Acad Sci USA *43*, 553-566.

Ofir, G., and Sorek, R. 2017. Vesicles spread susceptibility to phages. Cell *168*, 189-199.

Oldham, P.D. 2009. Global status and trends in intellectual property claims: genomics proteomics and biotechnology. SSRN (original date 2004).

Orwin, K.H., Dickie, I.A., Holdaway, R., and Wood, J.R. 2018. A comparison of the ability of PLFA and 16S rRNA gene metabarcoding to resolve soil community change and predict ecosystem functions. Soil Biol Biochem *117*, 27-35.

Overall, C.M. 2014. Can proteomics fill the gap between genomics and phenotypes? J Proteomics *100*, 1-2.

Palminteri, S. Scientists sequence plant DNA in the field to identify species within hours. https://news.mongabay.com/wildtech/2017/09/scientists-id-plant-species-in-the-field-within-hours/. Access date, 3 February 2018

Pan, J., Wang, B., Pei, Z.-Y., Zhao, W., Gao, J., Mao, J.-F., and Wang, X.-R. 2015. Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. Mol Ecol Resour *15*, 711-722.

Periyannan, S., Bansal, U., Bariana, H., Deal, K., Luo, M.-C., Dvorak, J., and Lagudah, E. 2014. Identification of a robust molecular marker for the detection of the stem rust resistance gene Sr45 in common wheat. Theor Appl Genet *127*, 947-955.

Perkel, J.M. 2017. How to hack the genome. Nature *547*, 477-479.

Petrone, J. 2016. DNA writers attract investors. Nat Biotechnol *34*, 363-364.

Piaggio, A.J., Segelbacher, G., Seddon, P.J., Alphey, L., Bennett, E.L., Carlson, R.H., Friedman, R.M., Kanavy, D., Phelan, R., Redford, K.H., *et al.* 2017. Is it time for synthetic biodiversity conservation? Trend Ecol Evol *32*, 97-107.

Popp, J., Pető, K., and Nagy, J. 2013. Pesticide productivity and food security. A review. Agron Sustain Dev *33*, 243-255.

Prakash, K., and Fournier, D. 2018. Evidence for the implication of the histone code in building the genome structure. Biosys *164*, 49-59.

Ptashne, M., Johnson, A.D., and Pabo, C.O. 1982. A genetic switch in a bacterial virus. Sci Amer *247*, 128-140.

Rajan, K. 2015. Materials informatics: the materials 'gene' and big data. Annu Rev Mater Res *45*, 153-169.

Rando, O.J. 2012. Combinatorial complexity in chromatin structure and function: revisiting the histone code. Curr Opin Genet Develop *22*, 148-155.

Rantsiou, K., Kathariou, S., Winkler, A., Skandamis, P., Saint-Cyr, M.J., Rouzeau-Szynalski, K., and Amézquita, A. in press. Next generation microbiological risk assessment: opportunities of whole genome sequencing (WGS) for foodborne pathogen surveillance, source tracking and risk assessment. Internatl J Food Microbiol.

Rawat, S., and Meena, S. 2014. Publish or perish: Where are we heading? J Res Med Sci *19*, 87-89.

Reardon, S. 2016. Welcome to the CRISPR zoo. Nature *531*, 160-163.

Regalado, A. 2016. 23andMe sells data for drug search. In MIT Technology Review.

Remington, K.A., Heidelberg, K., and Venter, J.C. 2005. Taking metagenomic studies in context. Trend Microbiol *13*, 404.

ResFinder. https://cge.cbs.dtu.dk/services/ResFinder/. Access date, 28 November 2017

Richards, R.A., Rebetzke, G.J., Watt, M., Condon, A.G., Spielmeyer, W., and Dolferus, R. 2010. Breeding for improved water productivity in temperate cereals: phenotyping, quantitative trait loci, markers and the selection environment. Funct Plant Biol *37*, 85-97.

Richardson, S.M., Mitchell, L.A., Stracquadanio, G., Yang, K., Dymond, J.S., DiCarlo, J.E., Lee, D., Huang, C.L.V., Chandrasegaran, S., Cai, Y., *et al.* 2017. Design of a synthetic yeast genome. Science *355*, 1040-1044.

Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., Altmann, T., Stitt, M., Willmitzer, L., and Melchinger, A.E. 2012. Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci USA *109*, 8872-8877.

Rodríguez López, C.M., and Wilkinson, M.J. 2015. Epi-fingerprinting and epi-interventions for improved crop production and food quality. Front Plant Sci *6*.

Rodríguez-Leal, D., Lemmon, Z.H., Man, J., Bartlett, M.E., and Lippman, Z.B. 2017. Engineering quantitative trait variation for crop improvement by genome editing. Cell *171*, 470-480.e478.

Sainz de Murieta, I., and Rodríguez-Patón, A. 2012. DNA biosensors that reason. Biosys *109*, 91-104.

Sammons, R.D., Ivashuta, S., Liu, H., Wang, D., Feng, P.C.C., Kouranov, A.Y., and Andersen, S.E. 2015a. Method for controlling herbicide-resistant plants. WO2011112570A1, https://patents.google.com/patent/WO2011112570A1/en.

Sammons, R.D., Ivashuta, S., Liu, H., Wang, D., Feng, P.C.C., Kouranov, A.Y., and Andersen, S.E. 2015b. Method for controlling herbicide-resistant plants http://www.google.com/patents/US9121022.

San Miguel, K., and Scott, J.G. 2016. The next generation of insecticides: dsRNA is stable as a foliar-applied insecticide. Pest Manag Sci *72*, 801-809.

Schafer, S., and Nadeau, J.H. 2015. The genetics of epigenetic inheritance: modes, molecules, and mechanisms. Q Rev Biol *90*, 381-415.

Schomburg, I., Chang, A., and Schomburg, D. 2002. BRENDA, enzyme data and metabolic information. Nucleic Acids Res *30*, 47-49.

Science. http://www.sciencemag.org/authors/science-editorial-policies - unpublished-data-and-personal-communications. Access date, 7 December 2017

Shaner, D.L., and Beckie, H.J. 2013. The future for weed control and technology. Pest Manag Sci *70*, 1329-1339.

Shu, Y., and McCauley, J. 2017. GISAID: global initiative on sharing all influenza data - from vision to reality. Euro Surveill *22*, 30494.

Songstad, D.D., Petolino, J.F., Voytas, D.F., and Reichert, N.A. 2017. Genome editing of plants. Crit Rev Pl Sci *36*, 1-23.

Sonka, S. 2014. Big data and the Ag sector: more than lots of numbers. Int Food Agribus Man *17*, 1-19.

Speck-Planche, A., Kleandrova, V.V., Ruso, J.M., and D. S. Cordeiro, M.N. 2016. First multitarget chemo-bioinformatic model to enable the discovery of antibacterial peptides against multiple Gram-positive pathogens. J Chem Inf Model *56*, 588-598.

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

Stegemann, S., and Bock, R. 2009. Exchange of genetic material between cells in plant tissue grafts. Science *324*, 650-651.

Steinfeld, H., Gerber, P., Wassenaar, T., Castel, V., Rosales, M., and de Haan, C. 2006. Livestock's long shadow. Environmental issues and options. (Rome) UN FAO, 391

Stephens, Z., Lee, S., Faghri, F., Campbell, R., Zhai, C., Efron, M., Iyer, R., Schatz, M., Sinha, S., and Robinson, G. 2015. Big data: astronomical or genomical? PLoS Biol *13*, e1002195.

Strohman, R.C. 1997. The coming Kuhnian revolution in biology. Nature Biotech *15*, 194-200.

Strzyz, P. 2017. Getting instructions from mum. Nat Rev Genet *18*, 513.

Sultana, S., Ali, M.E., Hossain, M.A.M., Asing, Naquiah, N., and Zaidul, I.S.M. 2018. Universal mini COI barcode for the identification of fish species in processed products. Food Res Internatl *105*, 19-28.

Suni, S.S., Scott, Z., Averill, A., and Whiteley, A. 2017. Population genetics of wild and managed pollinators: implications for crop pollination and the genetic integrity of wild bees. Conserv Genet *18*, 667-677.

Taberlet, P., Valentini, A., Rezaei, H.R., Naderi, S., Pompanon, F., Negrini, R., and Ajmone-Marsan, P. 2008. Are cattle, sheep, and goats endangered species? Mol Ecol *17*, 275-284.

Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., Yoshida, K., Uemura, A., Oikawa, K., Abe, A., *et al.* 2017. Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination. BMC Biol *15*, 86.

Tao, Z., Shen, L., Gu, X., Wang, Y., Yu, H., and He, Y. 2017. Embryonic epigenetic reprogramming by a pioneer transcription factor in plants. Nature *551*, 124-128.

Tapio, I., Snelling, T.J., Storzzi, F., and Wallace, R.J. 2017. The ruminal microbiome associated with methane emissions from ruminant livestock. J Anim Sci Biotechnol *8*, 7.

The Economist. 2012. Computing with soup. The Economist *Q1*.

Toombs, J.A., Petri, M., Paul, K.R., Kan, G.Y., Ben-Hur, A., and Ross, E.D. 2012. De novo design of synthetic prion domains. Proc Natl Acad Sci USA *109*, 6519-6524.

Tuttle, C., and Woodfin, L. 2014. Formulations and methods for control of weedy species. http://www.google.com/patents/WO2014167514A1?cl=en.

Tvedt, M.W., and Young, T. 2007. Beyond Access: Exploring Implementation of the Fair and Equitable Sharing Commitment in the CBD. ABS Series IUCN 2

Tyagi, A., Kumar, A., Aparna, S.V., Mallappa, R.H., Grover, S., and Batish, V.K. 2016. Synthetic biology: applications in the food sector. Crit Rev Food Sci Nutr *56*, 1777-1789.

UNEP/UNCTAD. 2008. Organic Agriculture and Food Security in Africa. (New York and Geneva) UNEP-UNCTAD Capacity-building Task Force on Trade, Environment and Development UNCTAD/DITC/TED/2007/15

Van, B.E., Kubler, L., Raemaekers, R., Bogaert, T., and Plaetinck, G. 2011. Methods for controlling pests using RNAi. https://www.google.co.nz/patents/EP2347759A2?cl=en.

Various. 2017. Database issue. Nucleic Acids Res *45*.

Varshney, R.K., Terauchi, R., and McCouch, S.R. 2014. Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. PLoS Biol *12*, e1001883.

Waldrop, M.M. 2016. The chips are down for Moore's law. Nature *530*, 144-147.
Wang, M., Shen, W., Yan, L., Wang, X.-H., and Xu, H. 2017a. Stepwise impact of urban wastewater treatment on the bacterial community structure, antibiotic contents, and prevalence of antimicrobial resistance. Environ Pollut *231*, 1578-1585.
Wang, Y., Cao, X., Zhao, Y., Fei, J., Hu, X., and Li, N. 2017b. Optimized double-digest genotyping by sequencing (ddGBS) method with high-density SNP markers and high genotyping accuracy for chickens. PLoS One *12*, e0179073.
Watson, J.D., Baker, T.A., Bell, S.P., Gann, A., Levine, M., and Losick, R. 2014. Molecular Biology of the Gene, 7 edn (Boston: Pearson and Cold Spring Harbor Laboraroty Press)
Watson, J.D., and Crick, F. 1953. Molecular structure of nucleic acids. Nature *171*, 737-738.
Weigle, J. 1966. Assembly of phage Lambda in vitro. Proc Natl Acad Sci USA *55*, 1462-1466.
Weiss, M.G. 1943. Inheritance and physiology of efficiency in iron utilization in soybeans. Genetics *28*, 253-268.
Welch, E.W., Bagley, M., and Kuiken, T. 2017. Potential implications of new synthetic biology and genomic research trajectories on the International Treaty for Plant Genetic Resources for Food and Agriculture (ITPGRFA or 'Treaty'). UN FAO
http://www.fao.org/fileadmin/user_upload/faoweb/plant-treaty/GB7/gb7_90.pdf
Weller, J.I., Ezra, E., and Ron, M. 2017. A perspective on the future of genomic selection in dairy cattle. J Dairy Sci *100*, 8633-8644.
WHO. Avian and other zoonotic influenza.
http://www.who.int/mediacentre/factsheets/avian_influenza/en/. Access date, 28 November 2017
Wickner, R.B., Edskes, H.K., Roberts, B.T., Baxa, U., Pierce, M.M., Ross, E.D., and Brachmann, A. 2004. Prions: proteins as genes and infectious entities. Genes & Develop *18*, 470-485.
Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., *et al.* 2014. Mass-spectrometry-based draft of the human proteome. Nature *509*, 582.
Wimmer, E. 2006. The test-tube synthesis of a chemical called poliovirus. EMBO Rep *7*, S3-S9.
Woolliams, J.A., and Oldenbroek, J.K. 2017. Genetic diversity issues in animal populations in the genomic era. In Genomic Management of Animal Genetic Diversity, J.K. Oldenbroek, ed. (Wageningen: Wageningen Academic).
Yang, W., Kang, X., Yang, Q., Lin, Y., and Fang, M. 2013. Review on the development of genotyping methods for assessing farm animal diversity. J Anim Sci Biotechnol *4*, 2-2.
Yong, E. A DNA sequence in every pocket.
https://www.theatlantic.com/science/archive/2016/04/this-technology-will-allow-anyone-to-sequence-dna-anywhere/479625/. Access date, 27 November 2017
Zhang, S. 2016. Farmers are manipulating microbiomes to help crops grow. In Wired.
Zhu, K.Y., Zhang, X., and Zhang, J. 2014. Double-stranded RNA-based nanoparticles for insect gene silencing. https://www.google.com/patents/US8841272.
Zimkus, B.M., and Ford, L.S. 2014. Genetic Resource Collections Associated with Natural History Museums: A Survey and Analysis to Establish a Benchmark of Standards. In DNA Banking for the 21st Century Proceedings of the US Workshop on DNA Banking, pp. 9-44.

GLOSSARY

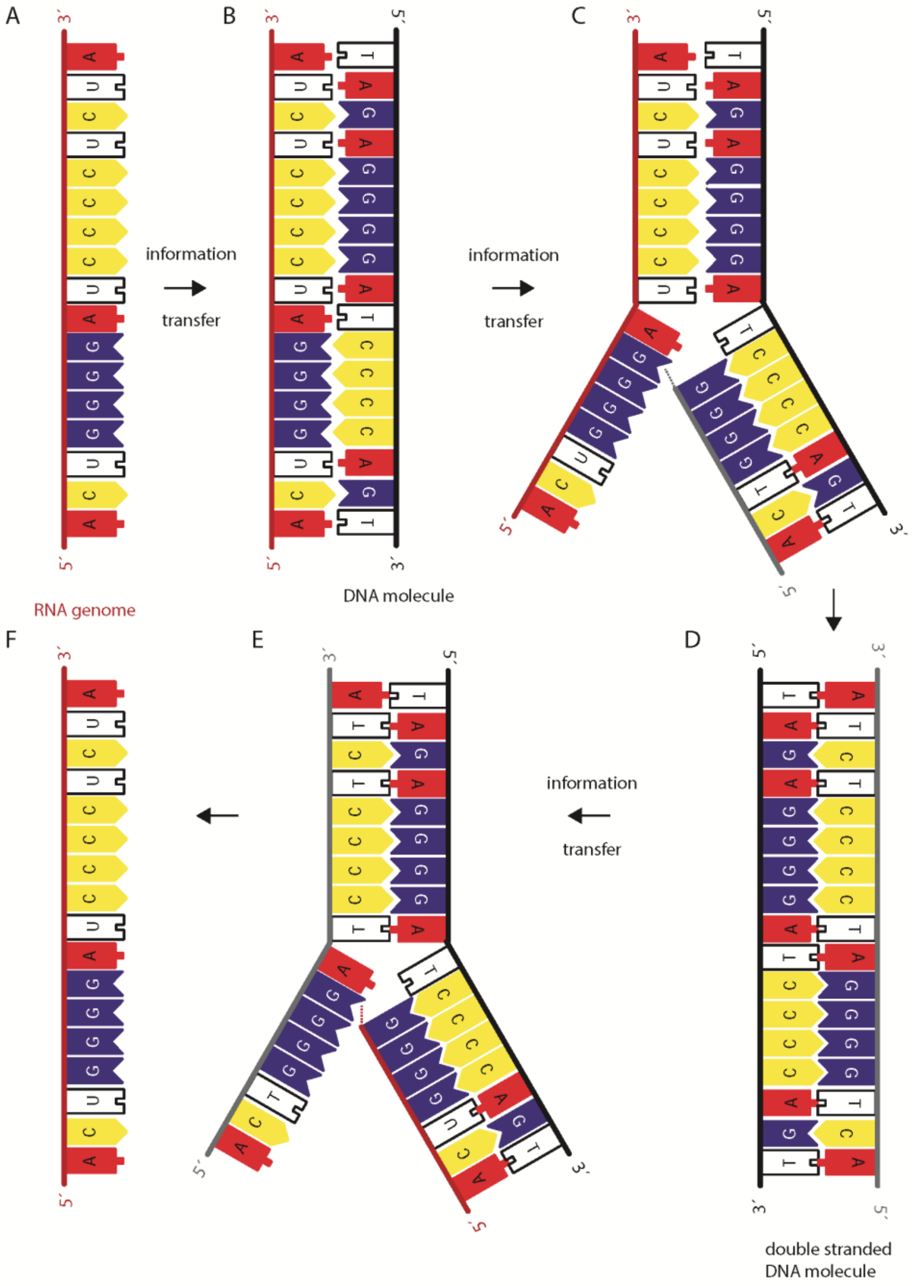
Alleles	Mutants or variants in the DNA sequence at a single gene or locus.
Amyloids	Aggregates of certain proteins folded into a shape that allows many copies of that protein to form fibrils.
Big data	Data of a very large size (structured or unstructured), typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data.
Bioinformatics	The branch of science concerned with information and information flow in biological systems, especially the use of computational methods in genetics and genomics.
Database	Usually a set of data held in computer storage and typically accessed or manipulated by means of specialized software.
Epiallele	Alleles that are DNA identical (no sequence difference) but present distinct epigenetic profiles due to differences in chromatin marks, such as DNA methylation that may be associated with changes in gene expression depending on the location of the modification.
Epigenetics	Heritable states that are not due to a change in DNA sequence. Different molecular mechanisms may apply including: (a) covalent modification of DNA such as methylation; (b) covalent modification of DNA binding proteins such as histones whose modification state has a hereditary component; and (c) genetic switches of stable regulatory states.
Epigenomics	A large set of epigenetic information from a single organism.
Gene drive systems	Alleles that when they are heterozygous convert the alternative allele into another copy of the gene drive allele.
Gene silencing	RNA or DNA directed biochemical pathway that inhibits either transcription or translation.
Genomic selection	The use of genetic markers that are spread throughout the genome to select individuals with desired predicted breeding values.
Histone	Any of a class of small basic proteins that are found in association with DNA in chromatin, and play an important role in the regulation of gene activity.
Histone modification	A histone modification is a covalent post-translational modification to histone proteins that includes methylation, phosphorylation, acetylation, ubiquitylation and sumoylation.
Landscape genomics	The study of genomes and their changes as a function of geography.
Metabolomics	A large set of information about metabolic intermediaries in an organism.
Metagenomics	The culture-independent genomic analysis of all the micro-organisms in a particular environmental sample.
Metadata	Data about other data.
omics	Refers to genomics, transcriptomics, proteomics, metabolomics and so on.

Phenomics	The use of large-scale approaches to study how genetic instructions from a single gene or the whole genome translate into the full set of phenotypic traits of an organism.
Prion	A protein with alternative conformations where one conformation is contagious.
Proteomics	A large set of information about the proteins in an organism.
Quantitative trait loci (QTL)	Some traits can vary continuously in a population, for example the size of individuals. These traits are influenced by many different genes, also called genetic loci, with the allelic state of each gene contributing a small amount.
Transcriptomics	A large set of all RNA in an organism whose informational content originated in a DNA sequence.
Traits	Any measurable aspect of an organism, including morphological, biochemical and molecular properties.

ANNEX

Interchangeability of information between media in nature.

The retrovirus lifecycle illustrates how information can be “dematerialized” from a source biological material. Information is transmissible between different media (Figure). Retroviruses have genomes built from ribonucleotides (RNA) (Panel A). The order of ribonucleotides guides the construction of a strand of a non-identical deoxyribonucleic acid (DNA) (Panel B). This strand in turn guides the construction of a non-identical second strand of DNA (Panel C) to form a double-stranded DNA molecule (Panel D). By this time, the information in the RNA genome has been transferred by biological processes two times resulting in both chemically different molecules and base sequences. The information is sequentially returned to a strand of RNA (Panels E-F). The descendants of a retrovirus contain no atoms from the RNA molecule of the original infecting virus, but they do inherit the same genes.



Arguably the conservation of information in the digitized representation of the physical order of nucleotides in a strand of DNA is no more dramatic a contrast with the strand of DNA itself than is the conservation of information transferring between two different nucleic acids, as it does in nature.

The transfer of information using multi-subunit multi-protein complexes as required by retroviruses would be no easier to design than would be the computer transferring such information. Natural systems through which information is conserved through physical transformation of medium are analogous to human-built systems that transform the sequence information from chemical to digital, and increasingly back through the capacity to synthesize a DNA molecule. They serve to highlight the observation that there is growing recognition “that ‘plant genetic resources’, i.e. genetic material of plant origin with actual or potential value for food and agriculture containing ‘functional units of heredity’, are subject to new characterization techniques in genomics that translate both the function and the physical unit into digital data sets” (Manzella, 2016).