My name is Paul Oldham and I am a Senior Visiting Fellow at the Institute for Advanced Study of Sustainability (IAS) at United Nations University. I was previously a researcher at the ESRC Centre for Economic and Social Aspects of Genomics (Cesagen) in the UK.

My comments are directed towards strengthening the fact finding and scoping study prepared for the Secretariat on digital sequence information.

Peer Review Comments

Given the short time frame that the authors have had to prepare a study on an issue of huge scale I think the authors deserve to be commended for the work that they have done. However, I think that the study could be improved in a variety of ways. Partly this is a matter or the organisation of the sections and their emphasis (notably with respect to synthetic biology) and partly an issue of strengthening aspects of the substantive content.

I think that readers of the report might benefit from a straightforward chunk describing the key developments linked to sequence data that would include:

a) Identification of the structure of DNA

b) The rise of Sanger Sequencing, followed by Shotgun Sequencing (that is for the human genome and rice genome etc).

c) The shift from short segments to whole genome sequencing (WGS) as enabled by shotgun sequencing and Next Generation Sequencing and the corresponding massive change in the scale of sequence data;

d) A shift from classic 'cut and splice' genetic engineering to chemical synthesis and engineering of DNA and whole genomes (with the Craig Venter group creating many of the landmarks for that)

e) The role of computational tools (massive parallel sequencing) and bioinformatics tools (such as visualization and modeling tools… which are critical for comprehending data at this kind of scale)

f) The opening up of engineering approaches to biology enabled by the above represented by the rise of synthetic biology, whole genome engineering and metabolic engineering. This includes emerging work on the use of artificial dna bases to express non-natural proteins.

g) Genome editing in health and agriculture as the most recent emerging technology.

That is, a more straightforward narrative account spelling out key landmarks and technologies (possibly in a box) could greatly assist the reader in understanding the

historical development of sequence data in the right order.

In this regard I would mention that UNEP/CBD/WG-ABS/3/INF/4 (https://www.cbd.int/doc/meetings/abs/abswg-03/information/abswg-03-inf-04-en.pdf) that was developed to inform the early debates on the Nagoya Protocol covers some of the history and issues addressed in the DSI paper at a time prior to the rise of synthetic biology.

Synthetic Biology

The COP decision on dsi arose in the context of the debates on synthetic biology. However, in my view it is important to recognize that synthetic biology is a relatively recent and small but growing field that is *only part of the story* of the rise of sequence data and its uses. While it is important to pay attention to synthetic biology (along with whole genome engineering, molecular engineering, genome editing etc.) in my view the paper presently forefronts synthetic biology in inappropriate ways.

Thus, section 3.1 on How is digital sequence information used and by whom starts with synthetic biology when synthetic biology logically appears towards the end of that section before section 3.1.5 on community laboratories etc.

While synthetic biology clearly merits attention I think the history of the rise of biotechnology and its relationship with sequence data should logically come first.

This type of problem crops up periodically throughout the paper and I think could in many cases be remedied through some reorganization. That is I strongly suggest that wherever possible a logical historical sequence is followed to give the reader a better sense of the emergence of developments over time.

Other comments on the treatment of synthetic biology

4.2 Registry of Standard Parts

This section states that shared repositories such as the Registry of Standard Biological Parts are common sources of genetic sequence data. I do not think that this is accurate because the registry of standard biological parts by definition covers standardized biological parts that have been created by researchers and is tiny and specific when compared with databases. This is not an argument that the repository of Standard Biological Parts should excluded, as it is important to synthetic biologists, but that it needs to be treated in a manner proportionate to the scale of wider use of sequence data.

I do not therefore think that the phrase used in the paper that ‘most researchers access sequence data through databases and parts registries’ is correct given that the standard biological parts registry consists of engineered parts. The use of the term repository

may be more appropriate here.

I note that the Nucleic Acids Research list of databases is referenced on page 30 and it may be possible to include other examples from that list to balance out this section. This 2012 article may assist in strengthening the treatment of protein sequences and structures. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3265122/

Open Science

It is important to highlight the issue of Open Science and its focus on open access to data and the reproducibility of research results. However, it is I think a mistake to conflate synthetic biology with the wider open science movement. Again, some parts of the synthetic biology community emphasise open science and open standards, on the other hand others do not. Focusing on synthetic biology misses the wider story of researchers who are dedicating themselves to promoting access to taxonomic, biological and sequence data and scientific literature through the development of software tools such as those from ROpenSci. Once again synthetic biology is part of the story but not the story.

Other comments

The evolution of approaches to licensing DNA data

In the discussion of the no restriction policy adopted by the INSDC I think the paper would benefit from a consideration of some of the history behind the rise of sequence data and databases and policies around openness. For example, the 1990s and early parts of the 21st Century were characterized by debates around who owned sequence data (notably the issues of patent rights and to a degree copyright as discussed in UNEP/CBD/WG-ABS/3/INF/4). So there was an important tension, notably in human genome sequencing and sequencing of food crops, between making the data open and keeping it secret. In some cases, notably the sequencing of the Rice genome by Syngenta upon completion of the mapping the company initially used a restrictive licensing model to access the data (see INF/4). However, as I understand it the data was later made available as part of wider international research on the rice genome.

Another example is the Single Nucleotide Polymorphism (SNP Consortium) referenced in INF/4 where companies concerned with the privatization of the SNPs through patent rights created the consortium to deliberately make the sequences publicly available. As has been widely pointed out this reflected a common interest in preventing problems with patent thickets around SNPs.

The issue here is that there has been a long running debate within the scientific and commercial communities about open vs closed proprietary approaches that needs to be highlighted as existing approaches to licensing or public domain cannot be

understood without that background. Sabrina Safrins outstanding 2004 paper discussed in INF4 and available here (https://ssrn.com/abstract=658421) discusses the implications of "hyperownership" when states also begin to assert sovereign rights. My own work built on this as does work by Manuel Ruiz and colleagues on natural information.

In connection with the policies adopted by INSDC focusing on unrestricted access, I would suggest that there is a need to articulate the tensions and debates that lie behind this. If Antoine Danchin remains as an EMBL advisor he could be a good contact to understand the reasoning and debates behind the INSDC policy. As someone who is also a thoughtful contributor to debates on synthetic biology Prof. Danchin might also be able to assist with finding a better balance in the treatment of synthetic biology in this paper.

Proteins

In general the paper focuses on DNA and amino acid sequences. However, I think more attention also needs to be paid to proteins, protein structures and the study of protein–protein interactions as an important outcome of the availability of DNA and amino acid sequences. Specifically, this is important in terms of product development and also links to the importance of visualization and modeling tools. In particular, in the discussion on line 13 of page 21 I am struck that techniques such as Nuclear Magnetic Resonance are important in the study of proteins and proteomics while modeling of protein folding and protein structures is also very important.

4.1.3 Standards for digital sequence information

This discussion also raises the important question of the quality of available sequence data in the absence of reference information. In particular, it is conceivable that the promotion of standards by the CBD could be one outcome of the discussion. Issues around data standards also came to the fore in a 2016 paper in Genome Biology that found that gene names in a review of supplementary material for journal publications contained errors because the authors used Excel (https://doi.org/10.1186/s13059-016-1044-7). At issue here is the integrity of the sequence data and therefore its broader utility to the scientific community.

6.2 Open source agreements

Some of this ground is covered in UNEP/CBD/WG-ABS/8/INF/3 which considered the potential use of open source agreements for ABS. It may provide some additional background. I would note that a lot of the software used in modern biology is open source and thus readily accessible to researchers in developing countries. Well known examples would be the bioconductor suite in R (https://www.bioconductor.org/) while the previously mentioned rOpenSci (https://ropensci.org/) is making important

contributions to improve free access to a wide range of taxonomic and related databases.

As noted in the comments on synthetic biology, while biobricks is certainly interesting it refers to engineered parts and there is a risk of missing the wider ecosystem of open source tools and licensing agreements that have accompanied the emergence of sequence data and bioinformatics.

Page 61. Patent applications line 31 onwards

I am afraid that this is very weak. As noted above, the history of the development of approaches to sequence data cannot be understood without considering debates on openness vs. proprietary rights. Concerns around patents are right in the middle of that and it is naïve to pretend otherwise.

The description of the Oldham et al article is incorrect as the article actually used informatics techniques to mine patent databases for 6 million species names to identify the international landscape of patent activity involving genetic resources across sectors.

[As an aside more recent work animal genetic resources and patent activity for WIPO and FAO is here: http://www.wipo.int/edocs/pubdocs/en/wipo_pub_947_3.pdf and includes discussion of sequencing]

The discussion on patent search engines could be better balanced. Patent data does reveal sequences of commercial interest (and sometimes epic controversy) precipitating the 2013 US Supreme Court decision. The original judgement from the lower court in the Myriad case is required reading here by focusing attention on the argument that "DNA represents the physical embodiment of biological information, distinct in its essential characteristics from any other chemical found in nature" (page 3 of the judgment available here: http://graphics8.nytimes.com/packages/pdf/national/20100329_patent_opinion.pdf?scp=3&sq=Myriad%20Genetics&st=cse). The Supreme Court judgement is important but it should be noted that it permits the patentability of cDNA. I would suggest here that the authors look up more recent literature on the actual impact and significance of the Supreme Court judgement.

The assertions on page 62 between lines 4-13 are questionable in terms of legal accuracy and I suggest that the authors or secretariat seek advice from lawyers at WIPO on the precise legal status of DNA sequences in different jurisdictions. I do not believe the assertions in the text to be true in Europe. For example European Directive Article 5.1 of Directive 98/44/EC specifies that: "...an element isolated from the human body or otherwise produced by means of a technical process, including the sequence or partial sequence of a gene, may constitute a patentable invention".

UNEP/CBD/WG-ABS/3/INF/4: 21 . As far as I am aware this has not changed.

I would therefore suggest seeking further clarification from WIPO to ensure the accuracy of the information in this paragraph as the laws vary by jurisdiction. The US Supreme Court decision does not have legal effect on patent laws outside the United States.

In connection with sentences 12-13 note that the World Intellectual Property Organisation maintains a list of sequence data submitted under the Patent Cooperation Treaty. The quote from Jefferson is somewhat peculiar given that the Lens provides access to sequence data in patents from a range of jurisdictions. Is the data not meaningful and in what sense?

Note also that discussions at the WIPO-IGC on disclosure of origin are ongoing and there are various studies associated with that process. The WIPO Secretariat may be able to assist here with the latest information on issues such as disclosure of origin.

Indigenous Peoples and Local Communities

My final observation concerns indigenous peoples and local communities who do not appear in the scoping study. I would observe in connection with sequence data that the collection of human DNA from indigenous peoples has been a focus of considerable debate over the last 20 years. Jenny Reardon, Kim TallBear and Rebecca Tsosie, among others, have documented and considered these issues and in the United States and Canada governance structures have been put in place by tribal and first nations governments to address issues around samples and sequences.

As discussed in UNEP/CBD/WG-ABS/8/INF/3: 40-41 in Canada the concept of DNA on loan from indigenous peoples gained in importance and I refer the authors in particular to Arbour, L & Cook, D (2006) 'DNA on Loan: Issues to Consider when Carrying out Genetic Research with Aboriginal Families and Communities'. *Community Genet* 2006; 9:153-160.

I am not aware of the latest status of this in Canada. However, Article 13 of the CIHR guidelines (http://www.cihr-irsc.gc.ca/e/29134.html) continues to state that

"Article 13 Biological samples should be considered "on loan" to the researcher unless otherwise specified in the research agreement.

Subject to the terms of the research agreement with their community, biological samples from Aboriginal participants should be considered "on loan" to the researcher, analogous to a licensing arrangement, and this should be detailed in the research agreement."

While the main focus of debate in connection with indigenous peoples has been targeted at health research I would note the widespread adoption of the UNDRIP by Parties and also observe that developments in the health sector could inform best practice in circumstances involving indigenous peoples and local communities under the Nagoya Protocol.

I would also note that while the paper is logically focused on the wider issues involved with sequence data, in fields such as genomics issues relating to the rights and interests of indigenous peoples played an important role in highlighting issues of ethics and equity in relation to genetic research. It would in my view therefore be a mistake to inadvertently exclude them.

Ends