

GigaScience

Myth-busting the provider-user relationship for digital sequence information

--Manuscript Draft--

Manuscript Number:	GIGA-D-21-00228R1	
Full Title:	Myth-busting the provider-user relationship for digital sequence information	
Article Type:	Research	
Funding Information:	bundesministerium für bildung und forschung (031B0862)	Dr. Amber Hartman Scholz
Abstract:	<p>The United Nations Convention on Biological Diversity (CBD) formally recognized the sovereign rights of nations over their biological diversity. Implicit within the treaty is the idea that mega-biodiverse countries will provide genetic resources and grant access to them and scientists in high-income countries will use these resources and share back benefits. However, little research has been conducted on how this framework is reflected in real-life scientific practice. Currently, parties to the CBD are debating whether digital sequence information (DSI) should be regulated under a new benefit-sharing framework. At this critical time point in the upcoming international negotiations, we test the fundamental hypothesis of provision and use by looking at the global patterns of access and use in scientific publications. Our data reject the provider-user relationship and suggest far more complex information flow for digital sequence information. Therefore, any new policy decisions on digital sequence information should be aware of the high level of use of DSI across low- and middle-income countries and seek to preserve open access to this crucial common good.</p>	
Corresponding Author:	Amber Hartman Scholz, Ph.D. Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures GmbH: Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH Braunschweig, GERMANY	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures GmbH: Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH	
Corresponding Author's Secondary Institution:		
First Author:	Amber Hartman Scholz, Ph.D.	
First Author Secondary Information:		
Order of Authors:	Amber Hartman Scholz, Ph.D. Matthias Lange Pia Habekost Paul Oldham Ibon Cancio Guy Cochrane Jens Freitag	
Order of Authors Secondary Information:		
Response to Reviewers:	Response to Reviewers Dear Dr. Scott Edmunds and anonymous reviewers, Thank you for the positive response to our manuscript, the helpful comments, and	

suggested improvements. We believe we have been able to address all of the reviewer's comments, which we respond to individually below. Most substantially, we have re-designed figure 1 and improved figure 2 per the suggestions of reviewer 2. Additionally, we have added several new sentences to address reviewer 1's concerns and conducted further typo checks and proofreading as recommended by reviewer 2.

We hope, with these edits, the manuscript will be appropriate for tandem publication with the "dataset paper" by Lange et al.

With our best regards,
Amber Hartman Scholz on behalf of the co-authors

Reviewer reports and responses:

Reviewer #1: The authors extract country names from gene sequence data in "traditional" ENA database and compare them with the country names of submitters to provide data-driven proof of the myth about the relationship between providers-users relationships for digital sequence information. I also implicitly believe in that myth. This verification has important implications for future handling of DSI. There is much room for debate in this manuscript, but I think it is important to ask the world as soon as possible. I make some points about this manuscript.

1.The Nagoya Protocol has prompted a major shift in the use of genetic resources. I think authors should also discuss the year of registration for digital sequence information. (Specifically, before and after the Nagoya Protocol) (It is desirable to provide actual data, but this time I will only seek the views of the author.)

At the present moment, there is no international consensus on whether digital sequence information (DSI) (at least that which is available from public databases) falls under an international legal framework for access and benefit-sharing. Therefore, there is no date that we can cite in the manuscript. However, as discussed in the introduction, a potential new access and benefit-sharing framework designed for DSI is actively being considered by Parties to the CBD. We have lightly edited the text to make this point 100% clear and added a new phrase "This decision is contested but widely expected to be resolved in some form.." and updated the dates for COP15 to April 2022.

2.The authors focus only on ENA's "traditional" gene sequence information (data corresponding to NCBI's GenBank). Regarding digital sequence information, NGS is currently being actively used, and data is also being accumulated in SRA (sequence read archive) at a tremendous pace. In addition, recent MinION devices can acquire digital sequence information on the spot without taking genetic resources out of the country. The lack of this data is a flaw in this manuscript, but at least the authors can discuss it in the manuscript, I think.

This paper builds off research done in Rohden et al. in which the "traditional" INSDC sequence dataset was analyzed. Thus, the same starting point for that study was used in this manuscript for consistency and comparison purposes which we hoped would be particularly useful for the policymakers that are already familiar with the Rohden et al. study which was requested by and produced for the Parties to the CBD as part of the inter-sessional process. Furthermore, in the interest of transparency, this project was conducted on a shoestring budget as a pilot project to see whether we could make new contributions to the policy process.

However, the reviewer rightly points out that inclusion of the SRA dataset would provide more information and statistical power. However, the data in the "traditional" sequence dataset are y more likely to actually be the dataset that policymakers are interested in because the SRA dataset is more targeted towards re-sequencing of genomes and metagenomes. While this is rapidly evolving, our thinking was that the traditional dataset was most representative of biological diversity and thus the most relevant for the policy topic presented here.

For future work, we are considering incorporating the much larger SRA dataset. And, we are developing new methods that will allow for much larger publication datasets in a future analysis. As this will require table merges with billions of records, there are some technical limitations that are still being explored.

The sentence describing future work has been edited to address this point on p.11 "Furthermore, future assessments will expand the baseline datasets to include larger sequence dataset such as from the INSDC Short Read Archive, include a more expansive set of open access publications, and provide first analyses on the field of study and taxonomic patterns."

3. In recent years, there has been a movement called "museomics" that extracts DNA from museum specimens and obtains digital sequence information. Museomics contributes to Ancient DNA and Taxonomic clarification. ENA/GenBank/DDBJ also has a "specimen_voucher" field, which already has hundreds of thousands of digital sequence information with this data (<https://doi.org/10.3897/biss.5.73787>). Does this have any effect on this study? Should DSI from museum specimen be excluded from the statistical processing in this study? (Authors don't have to mention this in the manuscript)

In Rohden et al., we noted that three types of metadata fields are available to connect sequence data to the original genetic resource: specimen_voucher, culture_collection, and bio_material. At present, only 4% of sequence entries use any of these three metadata fields, thus a relatively small fraction of the sequence dataset.

Furthermore, specimens that use these metadata fields might or might not fall out of the legal scope of the CBD's requirements for benefit-sharing. It would be inappropriate to, as the reviewer implies, exclude all specimen_voucher-associated sequences as these genetic resources could just as easily be from ancient DNA or have been acquired post-1992 (effective date of the CBD) or post-2014 (effective date of the Nagoya Protocol). In other words, excluding these or other sequence data would neither have a significant statistical effect (less than 4% of the dataset) nor be legally appropriate given that sequences could come from Nagoya/CBD-relevant material.

Reviewer #2: The paper is interesting and easy to follow, and the interactive charts are useful and make it easy to interact with the data. Moreover, the implications and contributions of the study are clear. However, the paper needs to be carefully be edit and proofread in order to make it ready for publication.

Thank you for the helpful editing and proofreading suggestions here and below which we have implemented in the manuscript in track changes.

Minor Issues:

The paper needs to be carefully proofread and edited

Done.

The layout of Figure 1 isn't good. I recommend using subplots with subtitles. Remove of countries with very low percentage. In addition, the caption of this figure needs to be improved.

Figure 1 has been re-designed as a distribution diagram and now replaces the three pie charts. The caption has also been revised.

Figure 2. Beside the countries colored with pink. It is hard to distinguish between the values of different countries. Although this figure emphasis that the ratio is mostly even, I personally think, it would be better to use different color scheme, or normalize the values, so the difference among countries will more noticeable.

As the reviewer correctly points out, the graph is intended to emphasize that the ratio is mostly even. We have re-made the figure with a new color scheme to make the differences (although still minimal) more pronounced and hope this improves the visual interpretation. Indeed, this new color scheme particularly enables more coloring in the slightly darker blue (corn blue) countries such as Bolivia or southwestern African

countries.

The paper's layout has unused free space in pages 5 and 6.

We believe this can be corrected during the subsequent editorial and layout process after re-submission and defer to the editorial staff.

The quality of figures throughout the paper need to be improve.

The figures were originally generated as .png files. We have significantly re-worked Figure 1 and 2 as requested by the reviewer and replaced all figures (except figure 4) with vector graphics which will be uploaded and submitted as separate files. Figure 4 is generated by "hovering over" the data for Malaysia and is thus the result of a user-experience interaction. Thus, unfortunately, for this figure, it is not possible to generate a vector graphic.

In the track-changes version of the re-submitted manuscript we have still included low-quality screenshots of the revised figures for convenience and ease of reviewing the manuscript. Please note these low-quality screenshots embedded in the MS Word document are intended to be removed in the copyediting process.

Typos (some examples):

All typos have been corrected. Many thanks to reviewer 2 for the keen eyes! Unless otherwise noted, they have been addressed in track changes.

-Abstract: CBD) -> CBD,
 -P. 2: "for user checks [4]. (Countires"
 -P. 3: originated. (Note
 This seems, in our opinion to be grammatically correct as-is.

-P. 3: "use" -> "use."
 -Figure 2: graph 3.4 -> Graph 3.4 (why call this Graph?)
 Graph 3.4 refers to the numbering of the graphs/figures available through the online data platform, which differ from the figure numbering in the manuscript. In fact, every figure in the manuscript has a similar sentence at the figure legend to refer the reader to the online data platform and indicate the numbering on that platform.

@Editorial staff: Would it be possible to embed a hyperlink to the data platform in the figure legend? We have added hyperlinks in the text and hope they will function still for the online manuscript.

-P. 8 "DSI (/country "
 /country is not a typo. This "/" is indicative that this is a formal metadata field associated with sequence entries. We have attempted to show this more clearly in the caption. This is fully explained earlier in the manuscript on p.3.

-P. 8: figure 3 -> Figure 3
 -Figure captions are with different formats.
 -Figure 4 caption is in a mislocated
 -P. 12: Figure 2 and 5 -> Figures 2 and 5

Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes
Full details of the experimental design and statistical methods used should be given	

<p>in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	
<p>Resources</p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>
<p>Availability of data and materials</p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>

Myth-busting the provider-user relationship for digital sequence information

Amber Hartman Scholz^{1*}, Matthias Lange², Pia Habekost², Paul Oldham³, Ibon Cancio⁴, Guy Cochrane⁵, Jens Freitag²

1. Leibniz Institute DSMZ German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany
2. Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany
3. Manchester Institute of Innovation Research, Alliance Manchester Business School, Manchester University, Manchester UK
4. Research Centre for Experimental Marine Biology and Biotechnology of Plentzia (PiE-UPV/EHU), University of the Basque Country (UPV/EHU), EMBRC-Spain, Plentzia, Spain
5. European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI), Hinxton, UK

*To whom correspondence should be addressed: amber.h.scholz@dsmz.de

ORCID IDs:

Amber Hartman Scholz: 0000-0002-3461-0881; Matthias Lange: 0000-0002-4316-078X; Pia Habekost: 0000-0002-6067-2322; Paul Oldham: 0000-0002-1013-4390; Ibon Cancio: 0000-0003-4841-0079; Guy Cochrane: 0000-0001-7954-7057; Jens Freitag: 0000-0001-6905-5497

Abstract

The United Nations Convention on Biological Diversity (CBD) formally recognized the sovereign rights of nations over their biological diversity. Implicit within the treaty is the idea that mega-biodiverse countries will provide genetic resources and grant access to them and scientists in high-income countries will use these resources and share back benefits. However, little research has been conducted on how this framework is reflected in real-life scientific practice. Currently, parties to the CBD are debating whether digital sequence information (DSI) should be regulated under a new benefit-sharing framework. At this critical time point in the upcoming international negotiations, we test the fundamental hypothesis of provision and use by looking at the global patterns of access and use in scientific publications. Our data reject the provider-user relationship and suggest far more complex information flow for digital sequence information. Therefore, any new policy decisions on digital sequence information should be aware of the high level of use of DSI across low- and middle-income countries and seek to preserve open access to this crucial common good.

Keywords

UN Convention on Biological Diversity; Access and benefit sharing; digital sequence information; provider user; Nagoya Protocol; INSDC

Background

The Convention on Biological Diversity is the international policy mechanism to reduce species, habitat, and ecosystem loss on this planet. The three overarching goals of the CBD, agreed upon in 1992, are conservation of biodiversity, sustainable use of this biodiversity, and fair and equitable benefit sharing from genetic resources. The third goal represents a political “balancing act” with the first two goals because it is intended to incentivize access and use of genetic resources (GR) so that benefits from use of biodiversity will flow back to the providing country, thus encouraging conservation, and supporting the first two goals.

Although it is officially recognized by parties to the CBD that all countries are both users and providers of GRs, in practice, most low- and middle-income countries (LMICs) see themselves predominantly as providers and, conversely, many high-income countries (HICs) view themselves as users [1]. While the CBD originally envisioned a facilitation mechanism for access to GR, the Nagoya Protocol (negotiated in 2010) codified a bilateral system in which a single country gives permission to a single user, which has perpetuated the provider-user paradigm [2]. In fact, the complex legal landscape that has resulted from the post-2010 implementation of the Nagoya Protocol reflects this. HICs often focus on user compliance [3] and LMICs focus on access laws even though every country should theoretically be responsible for user checks [4]. (Countries are not bound by the Nagoya Protocol to regulate access.) For example, to our knowledge, to date only developed countries have implemented user compliance mechanisms (i.e. laws that check if users have complied) with provider country laws, most notably the European Union [5] and Japan [6].

However, whether patterns of scientific use of GR actually follow these user-provider assumptions is not a question that has received much attention [7]. GR provision and use is difficult to follow since GR sampling and exchanges are not centrally administered or recorded. However, the use and citation of sequence data from GR in scientific publications enables a “proxy” view on provider-user relationships and happens to be itself highly relevant to the current CBD discussions.

In CBD policy circles, nucleotide sequence data (as well as potentially other data types) are known as “digital sequence information” (DSI) [8]. Because of the exponential growth and widespread use and reliance on DSI in the biological sciences, the political question of the hour is whether and how benefit-sharing from DSI should be required. This decision is contested but widely expected to be resolved in some form at the 15th Conference of the Parties (COP15). COP15 has been delayed by the pandemic but tentatively scheduled for late April 2022, will answer whether DSI should be treated like GR, whether monetary and/or non-monetary access

and benefit-sharing will be required and documented, and, if so, whether the policy framework for benefit-sharing will be bilateral or multilateral [9]. Thus, COP15 will be an important milestone for policymakers and scientists alike, making the question of patterns of use of DSI, presented here, quite timely.

Because many negotiators at the COP15 will be familiar with the bilateral mechanisms of the CBD and its Nagoya Protocol, it is likely that the default pre-conception around DSI for most negotiators will be the “provider-user dichotomy” assuming a primarily uni-directional (roughly global south to north) provision and use relationship. This is actually a hypothesis that can be tested with data from open access public DSI databases, in which the country of origin for the DSI can be found, and via publication databases, where use of DSI can be assessed by proxy through the affiliations of the authors, which can be parsed into geographical locations. While keeping in mind the potential shortcomings and accuracy issues [10], here we test this hypothesis and display the results in a free and open data analysis platform with the aim of analyzing whether a real directionality exists from provider country to DSI user country, with LMICs in one side and HICs on the other. The data and their implications are intended to support evidence-based policymaking.

Data Description

This article is a companion paper to a data note submitted in the same issue of this journal, which was made available as a pre-print in BioRxiv [11]. A web application is provided to explore and visualize the dataset at <http://wildsi.ipk-gatersleben.de> and, for clarity, we have compiled here a brief excerpt of the dataset described in the data note.

The vast majority of scientific journals and most funding agencies require that DSI be made freely available, at the latest, by the time of publication. Submissions of sequence data are required by journals as a condition of publication and rely upon the use of unique identifier(s) (called accession numbers, ANs) generated by a member of the International Nucleotide Sequence Database Collaboration (INSDC). During the sequence submission process, metadata associated with the DSI is also submitted including, where appropriate, the country field (data field “/country) which is defined as “locality of isolation of the sequenced sample indicated in terms of political names for nations, oceans or seas, followed by regions and localities.”

At the time of these analyses, there were 17,816,729 sequences in the INSDC with a country tag. Generally, sequences with country information come from a natural environment. For this study, we did not perform subsequent analyses on the taxonomic distribution of these sequences, which was previously assessed in [12]. Access to GR is needed to produce DSI. In this paper we use the term “provider” to designate the /country information found in the INSDC and indicate the geographical location from where the GR, and thus indirectly the DSI, originated. (Note that “provider” does not reflect where the sequencing was done or the entity that made the research/funding investment.)

For each of the >17.8 million ANs, if a publication was listed in the sequence entry page (within the INSDC database), this was added to the dataset as a “primary” publication. In a parallel step, the European PubMedCentral (ePMC) open-access publication database was text-mined for all >17.8 million sequence ANs. If a publication listed any of these sequences, it was added to the dataset as a “secondary publication”. A total of 117,483 primary and/or secondary publications were included in this analysis. Publications citing the use of DSI are representative of DSI scientific “use.” The associated author metadata from the primary and secondary publications was machine read and parsed. The geographical location of the first author was identified where data quality was sufficient. We note that first author information presents a restricted view of author networks that reflects limitations in the availability of full author information. As more author data becomes available, we anticipate that it will be possible to engage in analysis of author networks. This dataset forms the basis of the DSI “user” geographical locations. Additional quality control, data parsing, table merging, and data visualization steps were required that are further explained in the corresponding data note [11]. We make no further classification under the term “use”, as our methods at this stage cannot distinguish among the different types of use of DSI, for example, commercial versus non-commercial. On average, we expect that many peer-reviewed publications are more likely to derive from non-commercial research.

Analyses

The first question addressed is which countries are currently providing DSI to and using DSI from the global dataset available through the INSDC (Figure 1). The largest providers of DSI are currently not LMICs but are the United States, China, Canada, and Japan, providing roughly half of the global dataset. Large middle-income countries such as India and Brazil are in the next wave of providers (Fig. 1, blue bars). In the two years since Rohden et al. [12] described this trend for a CBD-commissioned study, the pattern has not significantly changed. Although, it is important to note that only 14.6% of all sequences in this INSDC release have country information available. This is down slightly from the 16% observed in the April 2019 global sequence dataset analyzed by Rohden et al. However, this does not necessarily represent a statistical trend. There are multiple factors that could cause this seeming decrease. For example, large deposits of sequences that are not appropriate for country labelling, e.g., human data (which would only rarely have country information associated with it) could grow the dataset and thus decrease the percent of country-labelled DSI. This question (cause of the 16% vs. 14.6%) was not further investigated.

Once these DSI are made available by provider countries, the natural question is to ask “who” is using the DSI, i.e., scientists sitting in which countries are publishing (which we call “using”) and citing DSI in a publication (Fig.1, purple bars). In order to begin to understand the provider-user relationship, it is also important to understand where the DSI that was *actually* used in a publication comes from, i.e., which countries provided access to the GR (Fig.1, red bars). To summarize, most DSI is being provided and used by HICs and DSI use and provision often occurs in roughly similar proportions. However, some outliers such as Costa Rica and Argentina

do show larger differences between providing and using DSI. These data do not, however, support the idea of a uni-directional provider-user dichotomy.

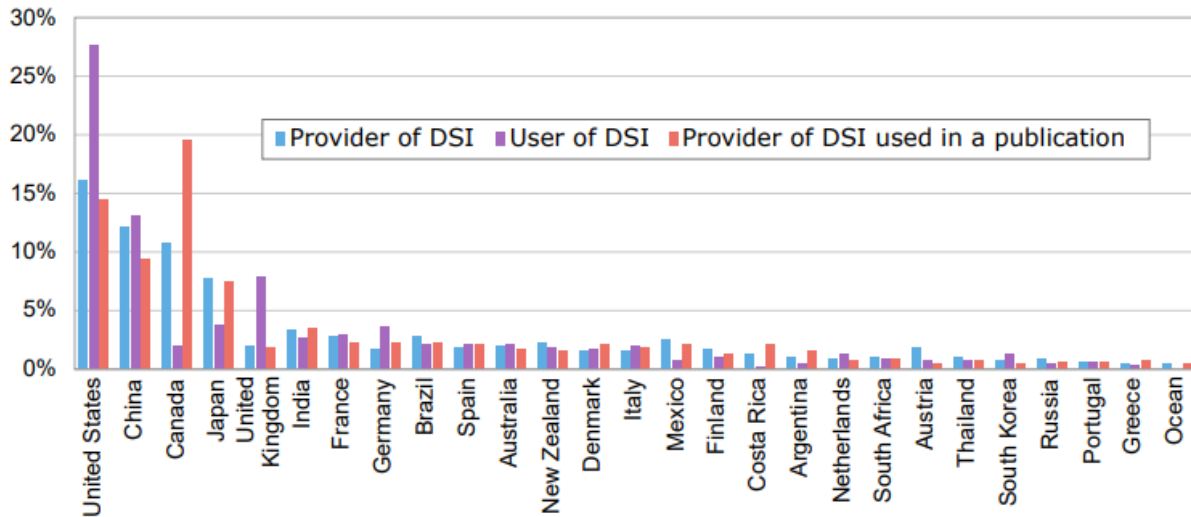


Figure 1. Bar graph comparison of each countries provision of DSI (that is, where it is the country of origin) for DSI (left bar) relative to its proportion of users (authors in scientific publications that reference DSI) (middle bar), compared with the proportion of its DSI cited in the middle bar. An interactive chart is available on the web platform under [graph 1.6](#).

To further understand the real-world provider-user relationship, another angle to examine is the relationship between a country’s use of its own “national DSI” (where it was the provider country of the GR) versus its use of DSI from other countries. We call this the “self:world” use ratio where self-use of DSI is divided by use of DSI from all other countries (Figure 2).

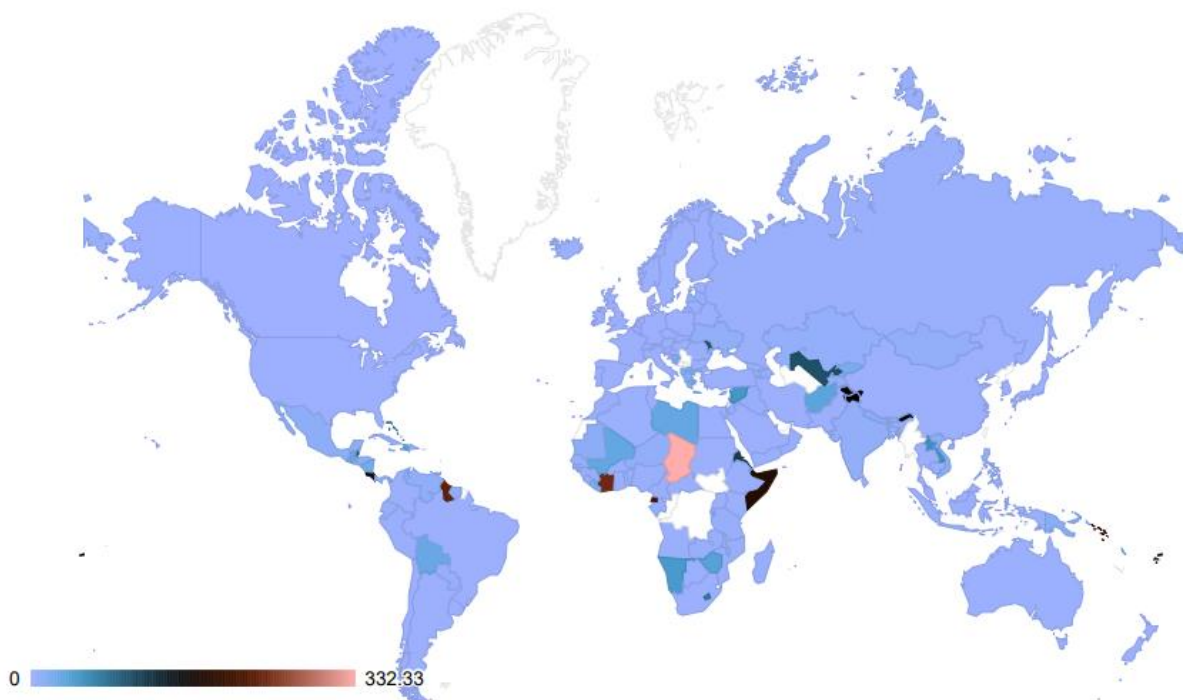


Figure 2. The “self:world” use ratio which is the relationship between the foreign use (use by non-domestic scientists) of a country’s DSI (nominator) and the use by domestic scientists of a country’s DSI. Light blue is a balanced use of DSI between foreign and domestic DSI. Pink is an indication of strong foreign use of DSI with less domestic use. Dark colors (black and red) indicate An interactive chart is available on the web platform under [graph 3.4](#).

Although there are significant differences in the number of scientists and the volume of use and re-use between LMICs and HICs, the relationship between a country’s use of its own DSI and use of the global DSI dataset (all non-national DSI) is relatively homogenous. A few countries have more “foreign use”, i.e. their national scientists use less relative to non-national scientists, such as Chad and Guyana. However, the vast swaths of green suggest that the ratio between national and foreign use of DSI is relatively even.

At international meetings, including the CBD Conference of the Parties, countries can form negotiating blocs that enable coordination between similar perspectives and sharing of preparation work when developing negotiating positions. These blocs often represent underlying economic similarities between countries. To understand broad trends through a similar lens used in these political discussions, we grouped countries into three overarching groups: low-income countries in a group called G-77, middle-income countries known as BRICS (Brazil, Russia, India, China, South Africa), and high-income countries under OECD (Organization of Economic Cooperation and Development). These broad groupings, although imperfect, allow for visual representations that could proxy common trends within UN political discussions.

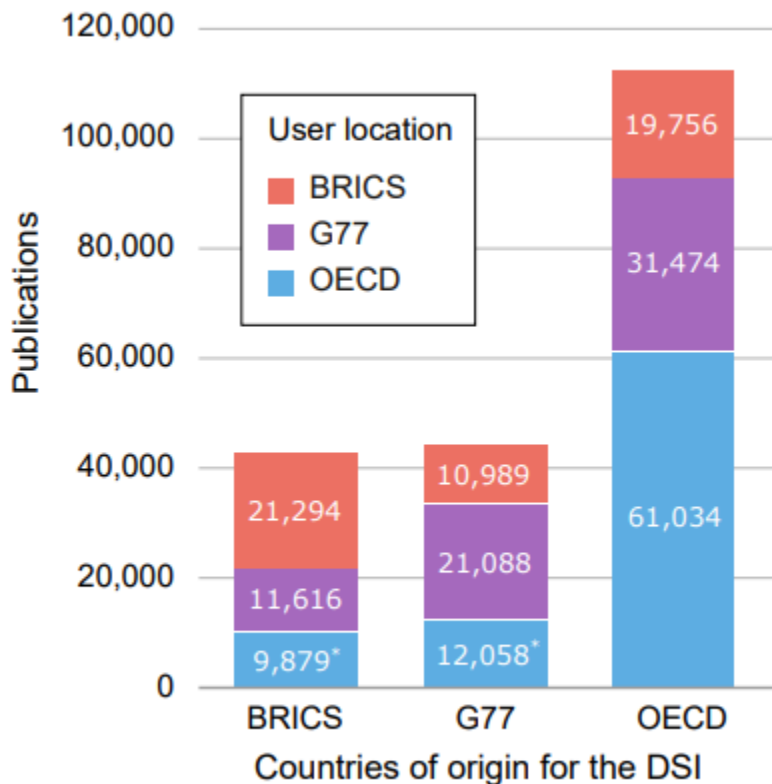


Figure 3. DSI use by economic blocs. On the x-axis is the country of origin of the DSI (i.e., the “/country” field in the sequence database). On the y-axis are the total number of DSI-using publications from where the color blocks represent the geographical location of the users (authors) grouped by economic blocs (i.e., OECD (proxy for developed countries; G77 proxy for low-income countries; BRICS (Brazil, Russia, China, India, and South Africa) proxy for large middle-income countries).. The * indicates the subset of the DSI-based publications where the international call for benefit-sharing is most focused and where concerns are heightened. An interactive chart is available on the web platform under [graph 5.1](#).

Three significant trends can be observed in Figure 3. First, the largest group of users (counted by a publication not an individual) in each bar always matches the economic bloc of the country of origin of the DSI. For example, the biggest group of users using G77-sourced DSI is G77-located users. This suggests that users tend to use DSI from their own country/bloc rather than from outside. They research and publish more with “locally-sourced” DSI. This data again rejects the unidirectional provider-use hypothesis. If that hypothesis were true, then OECD users (from HICs) should be the biggest user group in all three bars. Instead, OECD-located users are the smallest group of users of BRICS-sourced DSI and the second smallest group of users of G77-sourced DSI.

Second, OECD DSI is used nearly three times more than either BRICS-sourced DSI and G77-sourced DSI. This is shown by the difference in the height of the bars. This is likely due mostly

to the fact that there is simply a lot more OECD-sourced DSI in the databases than DSI from the other blocs. Many negotiators feel strongly that DSI sourced from mega-biodiverse countries is inherently more usable and valuable than DSI from other sources. The data shown here do not support this; otherwise G77 or BRICS DSI, even though fewer in the database, should be used more than OECD DSI.

Finally, the graph also shows that there are important gaps between the different blocs. There are fewer total DSI-related publications coming from users (authors) located in G77 and BRICS compared to OECD-based users (roughly 30-40% fewer). The use of DSI and, likely biological research in general, has lower total output as compared to OECD countries. However, the scale of the data also shows that G77- and BRICS-based authors are still quite scientifically productive. The * indicates the use of G77 or BRICS-sourced DSI by OECD-located users. These users are the primary intended targets of the CBD discussions on DSI and benefit-sharing even though these are the smallest blocks of use. It is critical that efforts to try to capture benefit-sharing from these two small blocks does not disrupt the “self-use” of BRICS and G77-based authors, which could potentially inhibit the biological research community and, with it, opportunities for sustainable economic development in these countries. In other words, policymakers would be wise to ensure that a shot aimed at benefit-sharing does not backfire on their own scientists.

The fourth question we were able to investigate with this dataset is the geographical interconnectedness between DSI researchers. To this end, two network diagrams were built. The “providing” network displays which countries are using a given country’s DSI (i.e., the countries to which country X is providing”). And the “using” network displays the countries whose DSI are being used by country X’s scientists. These data are also helpful to show that both neighboring and distant countries use DSI from many countries.

In Figure 4, DSI provisioning and use for Malaysia, which is both a G77 member and a mega-biodiverse country, is shown as an example. In Figure 4a, many LMICs (and not just HICs) use DSI from Malaysia, for example, Zambia, India, Peru, and Mexico to name a few. Conversely, in Figure 4b, scientists in Malaysia use DSI from 68 countries. Again, here there is no evidence of a provider-use relationship in DSI usage. Rather, Malaysian scientists use (cite in publications) DSI from a wide variety of countries and economic settings including Germany, Norway, Costa Rica, and Ghana to name a few. This data complements the data presented in Figure 3a which suggests that, although scientists use their own national DSI more frequently, when they use foreign DSI they do not appear to be primarily using DSI from biodiversity-rich countries but rather DSI from across the world without any clear geographical or economic clustering patterns.

Figure 4. Networks diagrams displaying country-based DSI provision and use patterns. In both graphs, data from Malaysia was selected as an example. A) 59 countries are using data from Malaysia. b) Malaysian scientists are using DSI from 68 countries. Neither the length of the connecting arrows nor the clustering reflects a statistical or quantitative relationship because the clustering algorithm is based on a random distribution. An interactive chart is available on the web platform under [graph 6.1 and 6.3](#).

A final question we addressed is what the overall providing-use relationship is for every country and whether there is a global trend to this relationship. Indeed, given the linear trend displayed in figure 5, it seems that many countries provide and use DSI from a roughly equal number of countries. In other words, if scientists from a given country are providing DSI, they are often using DSI at a similar level. However, for small countries, especially LMICS, with accordingly smaller datasets and scientific communities, tend to cluster in the bottom left of the graph meaning they have little provisioning of DSI and even less scientific use of DSI.

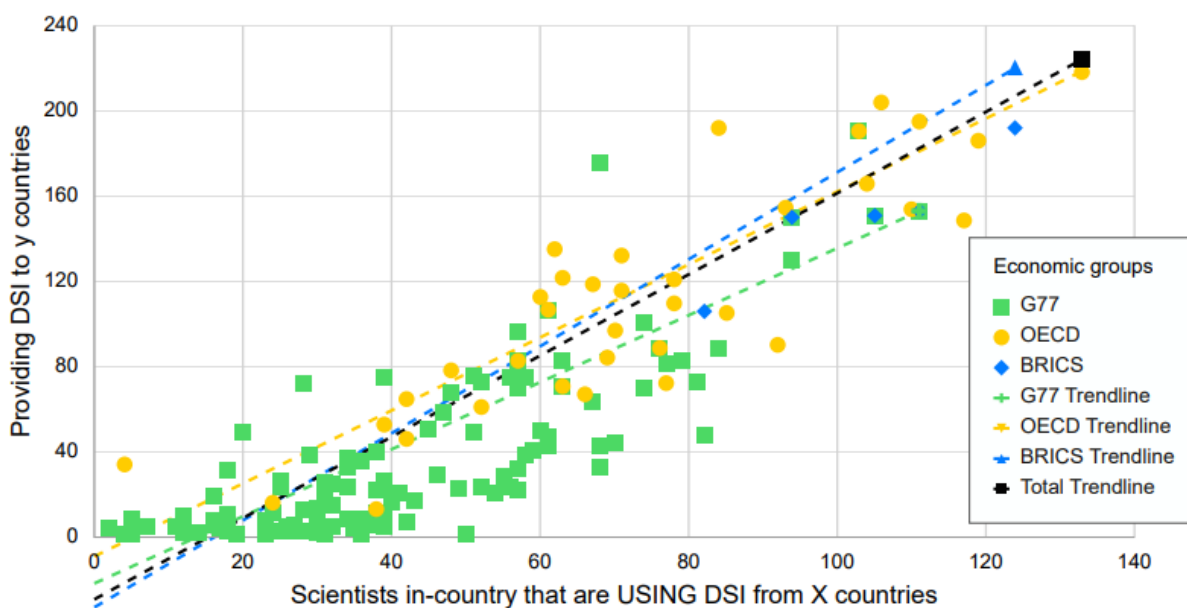


Figure 5. Relationship between use and provision of DSI for every country. The x-axis displays the use of DSI by a country's scientists and the y-axis displays the provisioning of DSI by a given country. An interactive chart is available on the web platform under [graph 6.3](#).

Discussion

Preconceptions in policymaking can be tested by looking at empirical data. Because there is a central repository for DSI, namely the International Nucleotide Database Collaboration, global

analyses can be conducted to inform the debate around benefit-sharing from DSI and test hypotheses. The data presented here show that the concept of a user-provider dichotomy from provision by low- and middle-income countries to use high-income countries for DSI is rejected.

This suggests that if an ABS policy mechanism for DSI incorrectly assumes a uni-directional provider-user relationship in which benefits (non-monetary and monetary) flow from HICs to LMICs, this will only occur in some instances and, indeed, based on figure 3, is the least frequent type of DSI use. Given current political discussions, LMIC-sourced DSI would be the most likely to fall under an ABS regime, but it is LMIC researchers that are the predominant users of this DSI (Figure 3). Thus, in a bilateral DSI-ABS system (where benefit-sharing is based on individual sequence records from a given country and benefit-sharing is based only on an individual sequence), scientists working in LMICs will have benefit-sharing obligations for their use of other LMIC- and HIC-sourced DSI. This could have unintended consequences and negatively impact scientists in LMICs because the DSI they use the most will require benefit-sharing (Figure 3, two left bars) while other DSI will likely not require benefit-sharing.

Scientists in LMICs are often resource-limited and have more personnel and infrastructure constraints than scientists working in HICs. If DSI policymakers do not recognize the “self-use” of DSI, that is the use of that country’s own DSI by in-country scientists, then they could potentially do great harm to scientists in LMICs which could have long-term implications for their domestic bioeconomy strategies and broader research and innovation goals. There are indeed important inequalities across the globe, but a DSI ABS system should try to *reduce* these inequalities rather than exacerbate them.

The global goal, in our view, ultimately should be to increase the scientific output and generation of DSI from G77 and BRICS countries to similar levels as those observed within the OECD and shown in Figure 3 (i.e., bring the two left bars up to similar levels as the right bar). Increased research capacity in LMICs would have global benefits and global biodiversity knowledge gaps, including those identified by the Global Biodiversity Framework, could be better filled. In order to do this, any DSI policy mechanism should recognize the existing divide and encourage DSI use, publication, and collaboration, perhaps explicitly dedicating significant capacity building to scientifically leveling the DSI playing field. This would be a much different approach than a “lock-it-up and control it” approach to DSI, which would negatively affect researchers everywhere on the globe and especially those in LMICs.

Furthermore, provisioning, use, and re-use of DSI as interpreted through DSI citation in publications is not a one-way street but rather a multi-directional traffic circle of data flowing in many different directions amongst all countries of the world. DSI is used by neighboring countries and distant countries, by LMICs and HICs alike without any clear patterns, regional, or economic trends (Figure 4). Additionally, the network diagrams remind us that scientists working in developing countries as well as the biological diversity of developed countries are often overlooked in political discussions that oversimplify provision and use of biological data.

The one clear trend observed in this analysis is that providing and using tend to go hand in hand (Figs. 2 and 5). Large countries tend to use and provide a relatively large amount of DSI and smaller countries use and provide less but trends based on development status (or, indirectly, the presence of mega-biodiversity) were not detected. In general, the relationship between use and provision seems to be relatively linear and not biased towards HICs but, instead, slightly biased towards LMICs (see LMIC trend line, Fig.5).

The dataset presented here is not a comprehensive dataset of all publications citing DSI but is limited to the open-access publications available for text-mining in ePMC as well as other dataset limitations explained in [11]. Furthermore, we note the dataset would be further improved if the country of origin information (14.6% of sequences in this dataset had such information) provided by scientists submitting sequence data to INSDC were more consistent and compliance with the requirement to submit this information for relevant DSI were bolstered. With this information, clearer references to regional conditions would be possible and thus more valid scientific statements and analyses would be possible. For example, gene-function relationships could be mapped more precisely with climatic, geological or atmospheric features.

However, acknowledging the above limitations, this analysis is the largest and only comparison of this size and perspective to-date and represents a novel attempt to bring data and a new perspective on DSI to the policymaking process. We encourage other groups to expand and build upon this dataset for other policy environments and to re-use and complement these data with additional perspectives. Future studies are planned that will expand this dataset to include closed-access publications and the patent system [13], where greater insights on potential commercial use of DSI can be assessed. Furthermore, future assessments will expand the baseline datasets to include larger sequence dataset such as from the INSDC Short Read Archive, include a more expansive set of open access publications, and provide first analyses on the field of study and taxonomic patterns.

Potential Implications

Future political decisions around how to handle DSI should account for the complexity of geographical provision and use trends presented here. A DSI framework that requires benefit-sharing from individual sequences, should anticipate the high level of LMIC-LMIC provision and use and high level of HIC-HIC provision and use. Benefit-sharing expectations should be adjusted accordingly. If policymakers do not want to require LMICs to provide benefits to other LMICs then a simple, de-coupled multilateral mechanism should be considered which decouples access and use of individual sequences from benefit-sharing requirements and instead requires benefits further downstream in the value chain.

Policymakers also need to appreciate the tremendous contribution towards non-monetary benefit sharing that these global biological and publication databases make towards broader CBD goals and towards the SDGs and create incentives in any DSI ABS framework to support these contributions. When policymakers meet in May 2022 to make a decision on DSI and the

Global Biodiversity Framework, we hope these data will make a constructive contribution to an evidence based DSI policymaking process.

These data also raise a fundamental question about current ABS frameworks already in place for GR, especially the Nagoya Protocol. For GR there is no central repository for movement across national borders as there is for DSI, but these data suggest that the provider-user relationship for GR could follow similar patterns to those observed for DSI. If so, this could suggest the existing bilateral system and the predominance of user checks in HICs (rather than globally) is perhaps not the most appropriate way to ensure benefit-sharing. While provocative, this could suggest that policymakers, in the future, ought to revisit ABS frameworks from the bottom up.

Methods

The methods used in this article are described in the companion paper, which is a Data Note in the same issue of this journal [11].

Data Availability

The dataset, figures, supplemental figures, and web application are available at <http://wildsi.ipk-gatersleben.de>. SQL queries to generate the figures are available in the GigaScience GigaDB repository [].

Abbreviations

ABS, access and benefit sharing
AN, accession number
BRICS, Brazil, Russia, India, China, South Africa
CBD, Convention on Biological Diversity
COP15, 15th Conference of the Parties to the Convention on Biological Diversity
DSI, digital sequence information
ePMC, European PubMed Central
G77, the Group of 77 representing mostly low-income countries
GR, genetic resources
HICs, high-income countries
INSDC, International Nucleotide Database Collaboration
LMICs, low- and middle-income countries
OECD, Organization of Economic Cooperation and Development
SDGs, Sustainable Development Goals
UN, United Nations

Competing Interests

The authors declare that they have no competing interests.

Funding

This publication was made possible by the research project WiLDSI (Wissenschaftliche Lösungsansätze für Digitale Sequenzinformation) funded by the German Federal Ministry of Education and Research (BMBF) under funding code 031B0862.

Author's contributions

AHS wrote the manuscript. AHS, GC, JF, and ML designed the research with critical input from IC and PO. ML and PH designed and programmed the data platform. All authors reviewed and edited the manuscript.

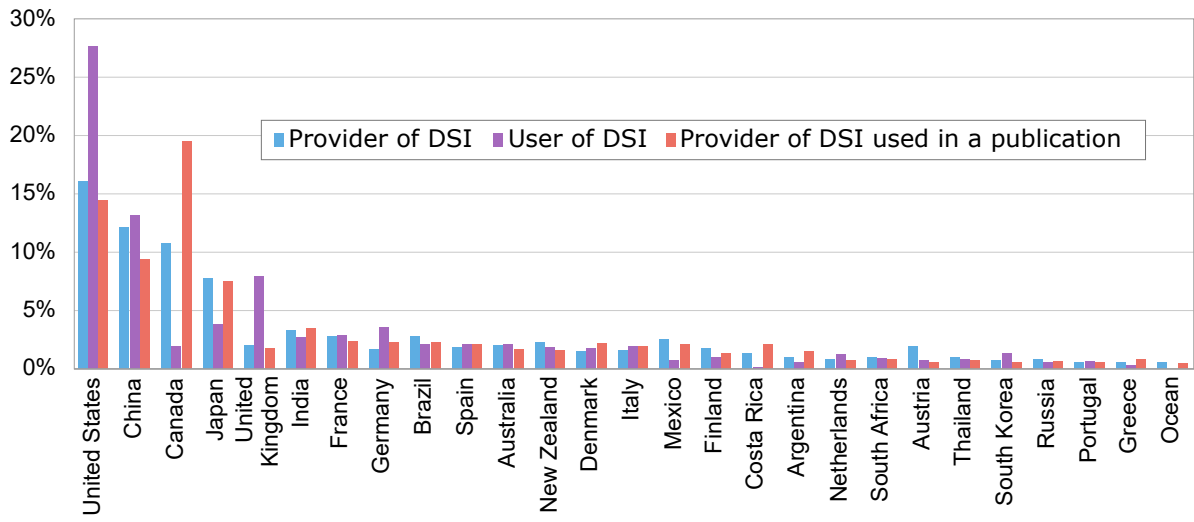
Acknowledgements

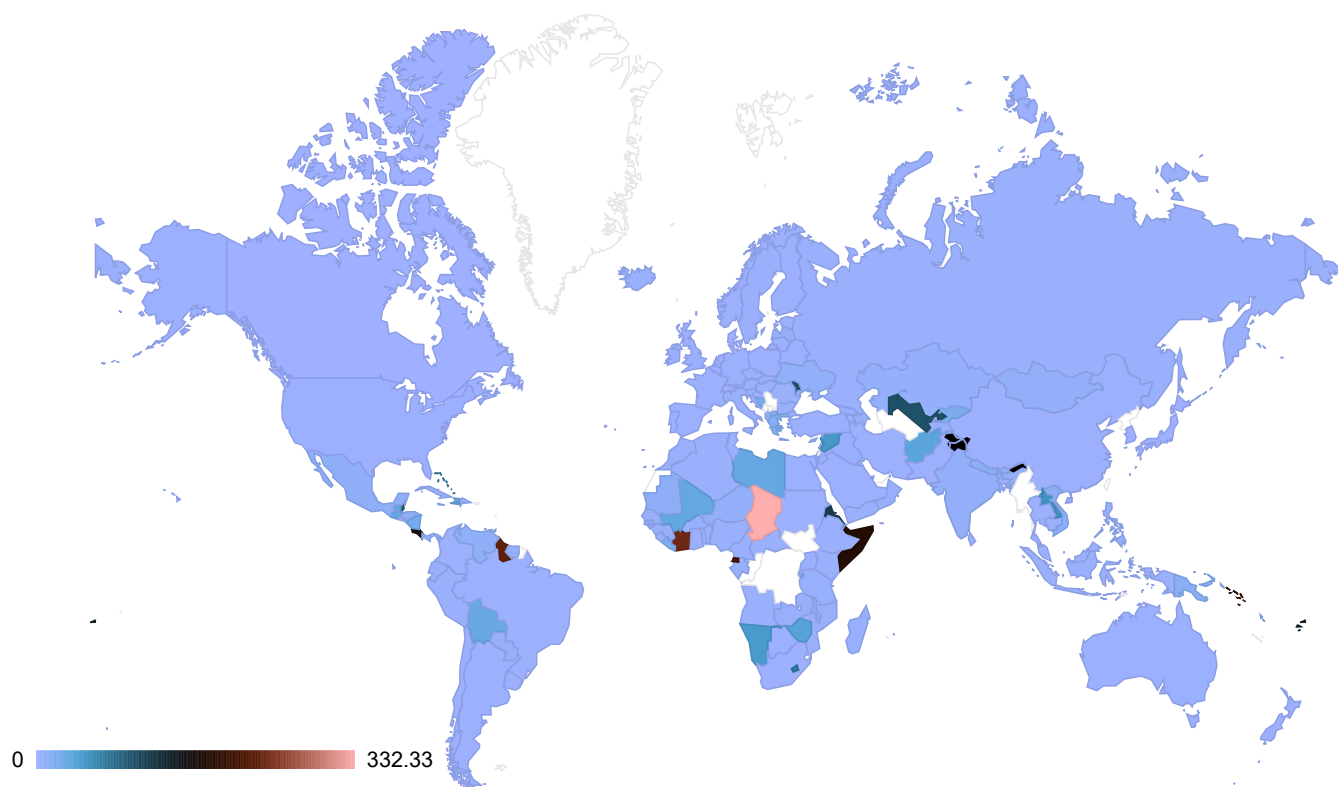
We gratefully acknowledge the important technical input of our co-authors, Upneet Hillebrand, Mehmood Ghaffar, Blaise Alako, and Florian Zunder, in the related Data Note publication [11]. Their work enabled this focused policy analysis. We also acknowledge the technical guidance in improving the figures and presentation from Andrew Hufton.

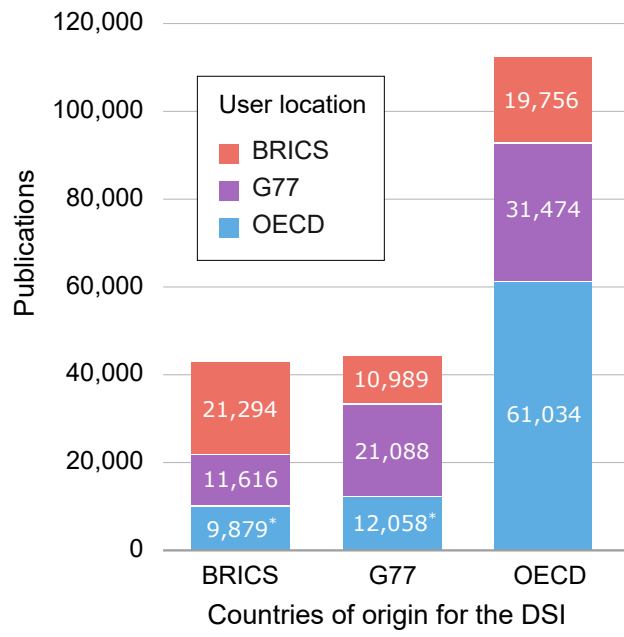
References

- [1] E. C. Kamau, B. Fedder, and G. Winter, 'The Nagoya Protocol on Access to Genetic Resources and Benefit Sharing: What is New and What are the Implications for Provider and User Countries and the Scientific Community', *Law Environ. Dev. J.*, vol. 6, p. 246, 2010.
- [2] M. Buck and C. Hamilton, 'The Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity: THE NAGOYA PROTOCOL', *Rev. Eur. Community Int. Environ. Law*, vol. 20, no. 1, pp. 47–61, Apr. 2011, doi: 10.1111/j.1467-9388.2011.00703.x.
- [3] F. A. Bockmann *et al.*, 'Brazil's government attacks biodiversity', *Science*, vol. 360, no. 6391, pp. 865.1-865, May 2018, doi: 10.1126/science.aat7540.
- [4] T. Greiber, 'Implementation of the Nagoya Protocol in the European Union and in Germany', *Phytomedicine*, vol. 53, pp. 313–318, Feb. 2019, doi: 10.1016/j.phymed.2018.10.020.
- [5] P. O. of the E. Union, 'CELEX1, Regulation (EU) No 511/2014 of the European Parliament and of the Council of 16 April 2014 on compliance measures for users from the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization in the Union Text with EEA relevance', Apr. 16, 2014. <http://op.europa.eu/en/publication-detail/-/publication/6b16d48a-dff0-11e3-8cd4-01aa75ed71a1/language-en> (accessed Jul. 29, 2021).
- [6] 'Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from

- their Utilization (ABS)'. <http://abs.env.go.jp/english.html> (accessed Jul. 29, 2021).
- [7] J. Overmann and A. H. Scholz, 'Microbiological Research Under the Nagoya Protocol: Facts and Fiction', *Trends Microbiol.*, vol. 25, no. 2, pp. 85–88, Feb. 2017, doi: 10.1016/j.tim.2016.11.001.
- [8] M. Brink and T. Hintum, 'Practical consequences of digital sequence information (DSI) definitions and access and benefit- sharing scenarios from a plant genebank's perspective', *PLANTS PEOPLE PLANET*, p. ppp3.10201, May 2021, doi: 10.1002/ppp3.10201.
- [9] F. Rohden and A. H. Scholz, 'The international political process around Digital Sequence Information under the Convention on Biological Diversity and the 2018–2020 intersessional period', *PLANTS PEOPLE PLANET*, p. ppp3.10198, May 2021, doi: 10.1002/ppp3.10198.
- [10] P. Sebo, S. de Lucia, and N. Vernaz, 'Accuracy of PubMed-based author lists of publications and use of author identifiers to address author name ambiguity: a cross-sectional study', *Scientometrics*, vol. 126, no. 5, pp. 4121–4135, May 2021, doi: 10.1007/s11192-020-03845-3.
- [11] M. Lange *et al.*, 'Quantitative monitoring of nucleotide sequence data from genetic resources in context of their citation in the scientific literature', *GigaScience* 2021. doi:10.1093/gigascience/giaxxx.
- [12] Ad hoc Technical Expert Group on Digital Sequence Information on Genetic Resources, 'Combined study on digital sequence information in public and private databases and traceability'. <https://www.cbd.int/doc/c/1f8f/d793/57cb114ca40cb6468f479584/dsi-ahteg-2020-01-04-en.pdf>
- [13] P. Oldham, S. Hall, and O. Forero, 'Biological Diversity in the Patent System', *PLoS ONE*, vol. 8, no. 11, p. e78737, Nov. 2013, doi: 10.1371/journal.pone.0078737.
- [14] Scholz AH; Lange M; Habekost P; Oldham P; Cancio I; Cochrane G; Freitag J (2021): Supporting data for "Myth-busting the provider-user relationship for digital sequence information" GigaScience Database. <http://dx.doi.org/10.5524/100945>







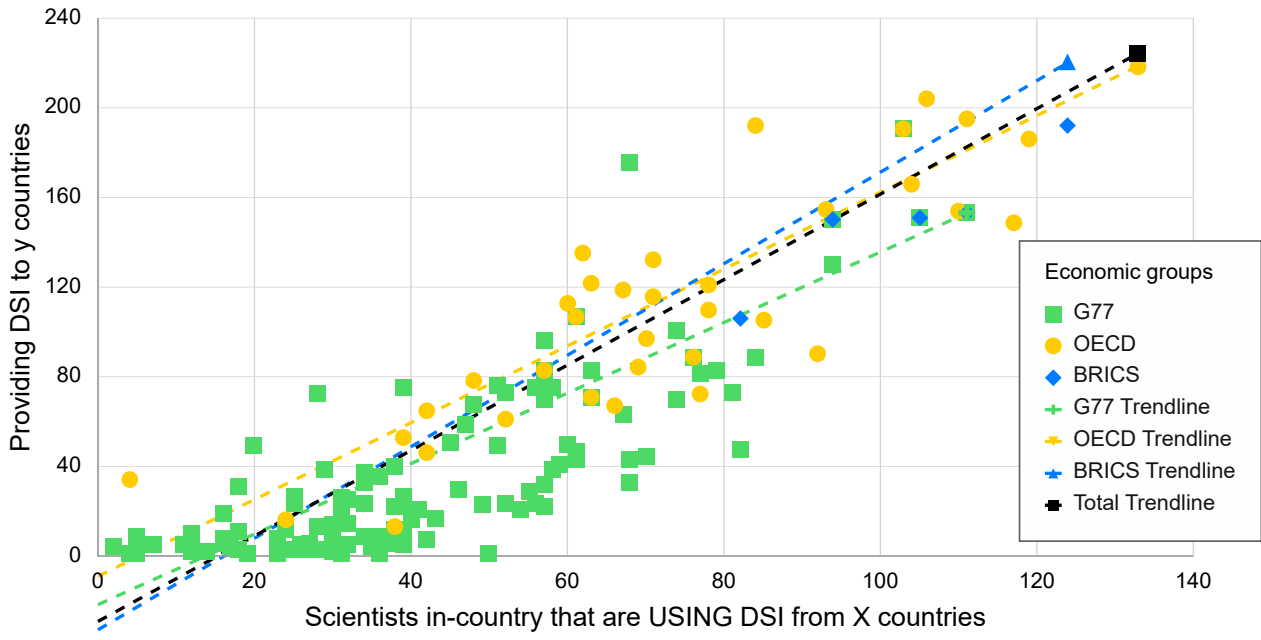
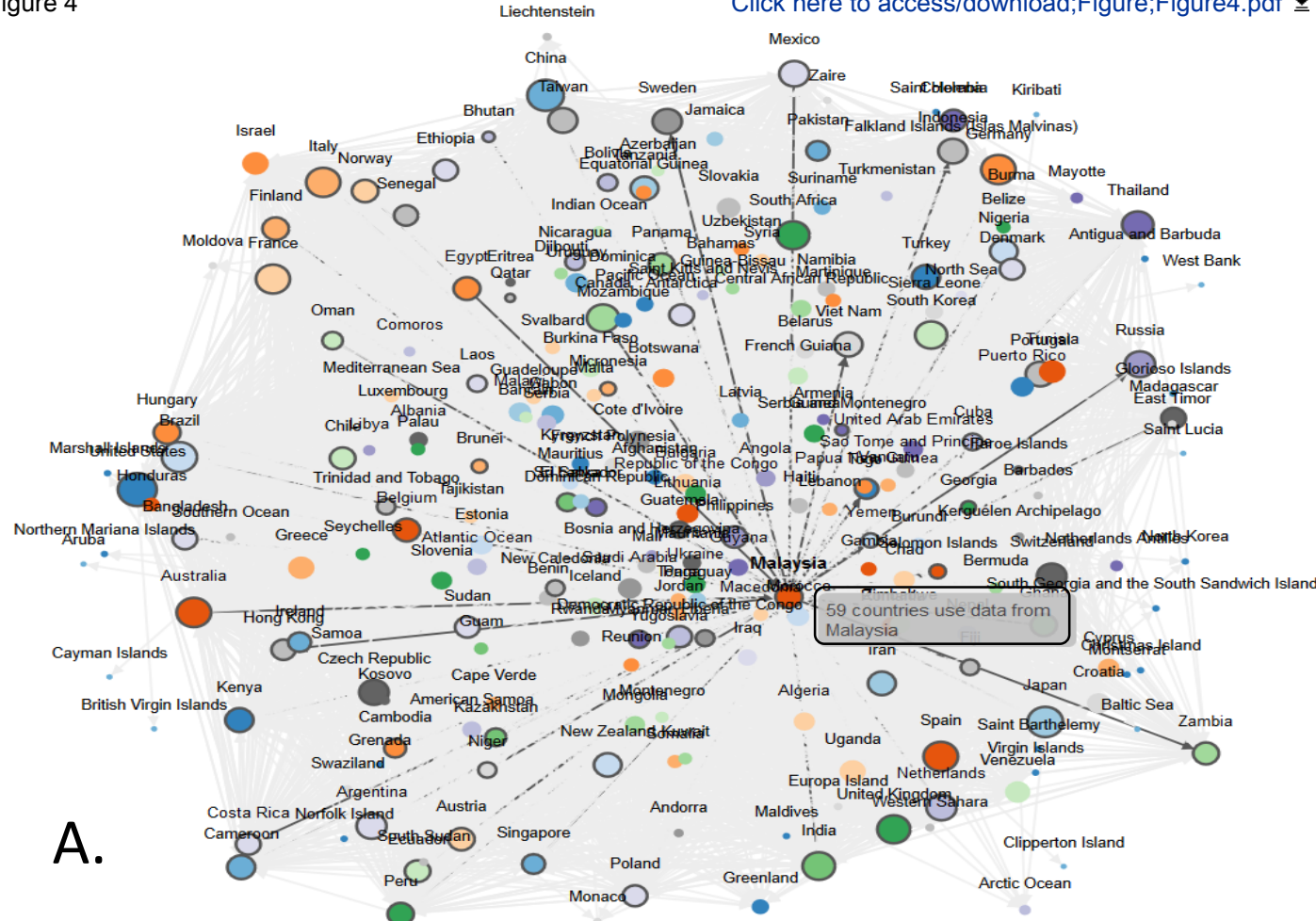
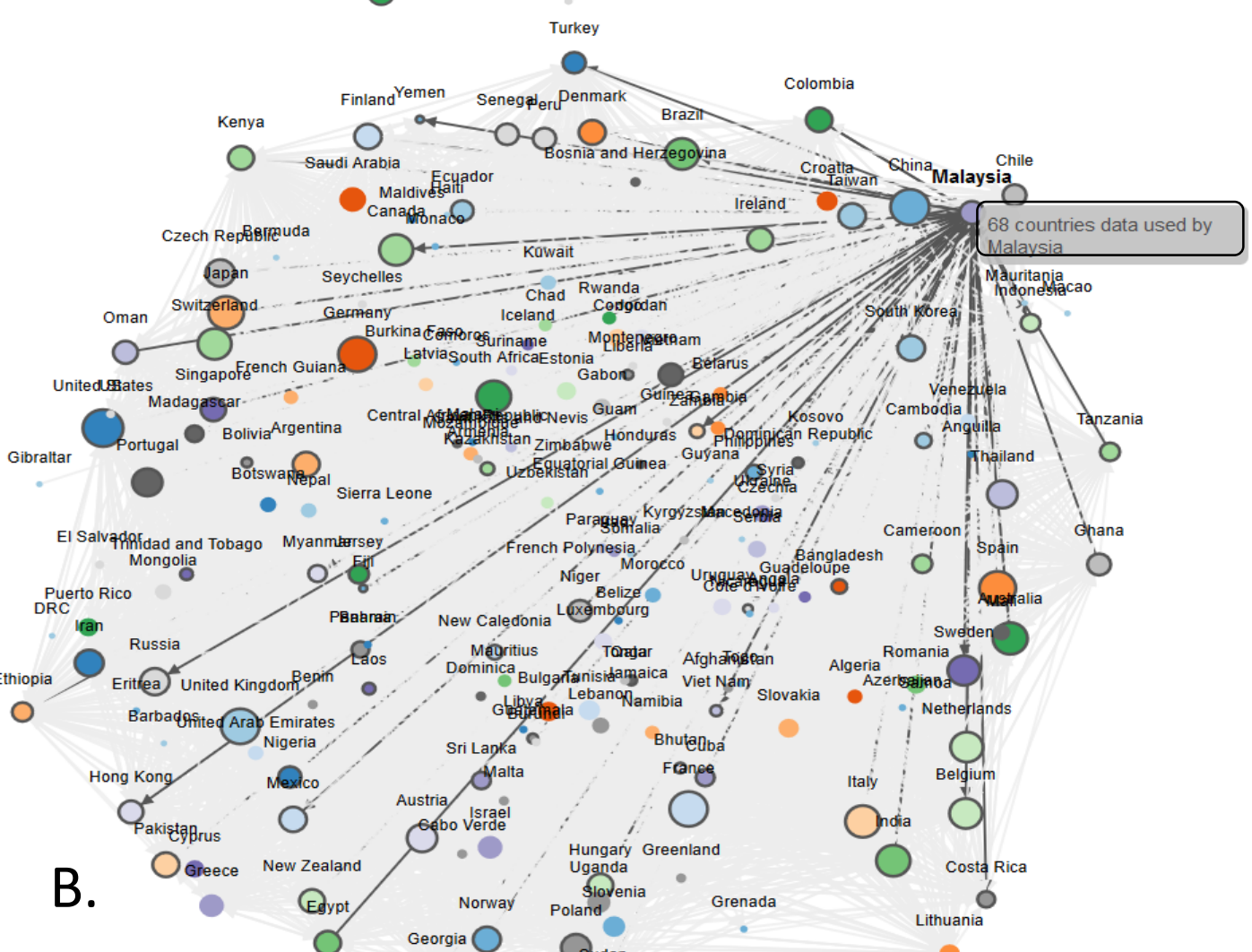


Figure 4

[Click here to access/download;Figure;Figure4.pdf](#)



A.



B.

Response to Reviewers

Dear Dr. Scott Edmunds and anonymous reviewers,

Thank you for the positive response to our manuscript and the helpful comments and suggested improvements. We have been able to address all of the reviewer's comments, which we respond to individually below. Most substantially, we have re-designed figure 1 and improved figure 2 per the suggestions of reviewer 2. Additionally, we have added several new sentences to address reviewer 1's concerns and conducted further typo checks and proofreading as recommended by reviewer 2.

We hope, with these edits, the manuscript will be appropriate for tandem publication with the "dataset paper" by Lange et al.

With our best regards,
Amber Hartman Scholz on behalf of the co-authors

Reviewer reports and responses:

Reviewer #1: The authors extract country names from gene sequence data in "traditional" ENA database and compare them with the country names of submitters to provide data-driven proof of the myth about the relationship between providers-users relationships for digital sequence information. I also implicitly believe in that myth. This verification has important implications for future handling of DSI. There is much room for debate in this manuscript, but I think it is important to ask the world as soon as possible. I make some points about this manuscript.

- 1. The Nagoya Protocol has prompted a major shift in the use of genetic resources. I think authors should also discuss the year of registration for digital sequence information. (Specifically, before and after the Nagoya Protocol) (It is desirable to provide actual data, but this time I will only seek the views of the author.)**

At the present moment, there is no international consensus on whether digital sequence information (DSI) (at least that which is available from public databases) falls under an international legal framework for access and benefit-sharing. Therefore, there is no date that we can cite in the manuscript. However, as discussed in the introduction, a potential new access and benefit-sharing framework designed for DSI is actively being *considered* by Parties to the CBD. We have lightly edited the text to make this point 100% clear and added a new phrase "This decision is contested but widely expected to be resolved in some form.." and updated the dates for COP15 to April 2022.

- 2. The authors focus only on ENA's "traditional" gene sequence information (data corresponding to NCBI's GenBank). Regarding digital sequence information, NGS is currently being actively used, and data is also being accumulated in SRA (sequence read archive) at a tremendous pace. In addition, recent MinION devices can acquire digital sequence information on the spot without taking genetic resources out of the country. The lack of this data is a flaw in this manuscript, but at least the authors can discuss it in the manuscript, I think.**

This paper builds off research done in Rohden et al. in which the “traditional” INSDC sequence dataset was analyzed. Thus, the same starting point for that study was used in this manuscript for consistency and comparison purposes which we hoped would be particularly useful for the policymakers that are already familiar with the Rohden et al. study which was requested by and produced for the Parties to the CBD as part of the inter-sessional process. Furthermore, in the interest of transparency, this project was conducted on a shoestring budget as a pilot project to see whether we could make new contributions to the policy process.

However, the reviewer rightly points out that inclusion of the SRA dataset would provide more information and statistical power. However, the data in the “traditional” sequence dataset are y more likely to actually be the dataset that policymakers are interested in because the SRA dataset is more targeted towards re-sequencing of genomes and metagenomes. While this is rapidly evolving, our thinking was that the traditional dataset was most representative of biological diversity and thus the most relevant for the policy topic presented here.

For future work, we are considering incorporating the much larger SRA dataset. And, we are developing new methods that will allow for much larger publication datasets in a future analysis. As this will require table merges with billions of records, there are some technical limitations that are still being explored.

The sentence describing future work has been edited to address this point on p.11
“Furthermore, future assessments will expand the baseline datasets to include larger sequence dataset such as from the INSDC Short Read Archive, include a more expansive set of open access publications, and provide first analyses on the field of study and taxonomic patterns.”

3. In recent years, there has been a movement called "museomics" that extracts DNA from museum specimens and obtains digital sequence information. Museomics contributes to Ancient DNA and Taxonomic clarification. ENA/GenBank/DDBJ also has a "specimen_voucher" field, which already has hundreds of thousands of digital sequence information with this data (<https://doi.org/10.3897/biss.5.73787>). Does this have any effect on this study? Should DSI from museum specimen be excluded from the statistical processing in this study? (Authors don't have to mention this in the manuscript)

In Rohden et al., we noted that three types of metadata fields are available to connect sequence data to the original genetic resource: specimen_voucher, culture_collection, and bio_material. At present, only 4% of sequence entries use any of these three metadata fields, thus a relatively small fraction of the sequence dataset.

Furthermore, specimens that use these metadata fields might or might not fall out of the legal scope of the CBD's requirements for benefit-sharing. It would be inappropriate to, as the reviewer implies, exclude all specimen_voucher-associated sequences as these genetic resources could just as easily be from ancient DNA or have been acquired post-1992 (effective date of the CBD) or post-2014 (effective date of the Nagoya Protocol). In other words, excluding these or other sequence data would neither have a significant statistical effect (less than 4% of the dataset) nor be legally appropriate given that sequences could come from Nagoya/CBD-relevant material.

Reviewer #2: The paper is interesting and easy to follow, and the interactive charts are useful and make it easy to interact with the data. Moreover, the implications and contributions of the study are clear. However, the paper needs to be carefully be edit and proofread in order to make it ready for publication.

Thank you for the helpful editing and proofreading suggestions here and below which we have implemented in the manuscript in track changes.

Minor Issues:

The paper needs to be carefully proofread and edited

Done.

The layout of Figure 1 isn't good. I recommend using subplots with subtitles. Remove of countries with very low percentage. In addition, the caption of this figure needs to be improved.

Figure 1 has been re-designed as a distribution diagram and now replaces the three pie charts. The caption has also been revised.

Figure 2. Beside the countries colored with pink. It is hard to distinguish between the values of different countries. Although this figure emphasis that the ratio is mostly even, I personally think, it would be better to use different color scheme, or normalize the values, so the difference among countries will more noticeable.

As the reviewer correctly points out, the graph is intended to emphasize that the ratio is mostly even. We have re-made the figure with a new color scheme to make the differences (although still minimal) more pronounced and hope this improves the visual interpretation. Indeed, this new color scheme particularly enables more coloring in the slightly darker blue (corn blue) countries such as Bolivia or southwestern African countries.

The paper's layout has unused free space in pages 5 and 6.

We believe this can be corrected during the subsequent editorial and layout process after re-submission and defer to the editorial staff.

The quality of figures throughout the paper need to be improve.

The figures were originally generated as .png files. We have significantly re-worked Figure 1 and 2 as requested by the reviewer and replaced all figures (except figure 4) with vector graphics which will be uploaded and submitted as separate files. Figure 4 is generated by "hovering over" the data for Malaysia and is thus the result of a user-experience interaction. Thus, unfortunately, for this figure, it is not possible to generate a vector graphic.

In the track-changes version of the re-submitted manuscript we have still included low-quality screenshots of the revised figures for convenience and ease of reviewing the manuscript. **Please note these low-quality screenshots embedded in the MS Word document are intended to be removed in the copyediting process.**

Typos (some examples):

All typos have been corrected. Many thanks to reviewer 2!

- Abstract: CBD) -> CBD,
- P. 2: "for user checks [4]. (Countires"
- P. 3: originated. (Note

This seems, in our opinion to be grammatically correct as-is.

- P. 3: "use" -> "use."

- Figure 2: graph 3.4 -> Graph 3.4 (why call this Graph?)

Graph 3.4 refers to the numbering of the graphs/figures available through the online data platform which differ from the figure numbering in the manuscript.

- P. 8 "DSI (/country "

/country is not a typo. This "/" is indicative that this is a formal metadata field associated with sequence entries. We have attempted to show this more clearly in the caption. This is fully explained earlier in the manuscript on p.3.

- P. 8: figure 3 -> Figure 3

- Figure captions are with different formats.

- Figure 4 caption is in a mislocated

- P. 12: Figure 2 and 5 -> Figures 2 and 5