Introduction

1. The Wellcome Sanger Institute uses genomic sequences to advance the understanding of human and pathogen biology to improve human health. We use science at scale to tackle the most challenging global health research questions.

2. The Wellcome Sanger Institute, based on the Wellcome Genome Campus, Hinxton, UK, is a world-leading hub for genomes and biodata research. The campus is also home to EMBL-EBI, Connecting Science, Sanger Institute spin-out companies, start-up companies and Genomics England.

Key messages

3. We strongly disagree with the proposal to include digital sequence information (DSI) in the scope of the Convention on Biological Diversity (CBD) and the Nagoya Protocol.

4. We fully agree that countries should share equitably in the benefits of research and development which utilises sovereign genetic resources, but it is our view that the inclusion of DSI in the Nagoya Protocol would fail to achieve this goal and has the potential to do more harm than good.

5. Science is a global endeavour and the ability to tackle complex global challenges depends on international collaboration and the ability to share and freely access research findings and data. International research, development and surveillance activities could be seriously threatened by the inclusion of DSI in the Nagoya Protocol.

6. Globally, researchers are sharing and utilising DSI on animals, plants and pathogens via open access databases. It is absolutely vital that free and unrestricted sharing of DSI continues. To impose terms and conditions or access agreements would seriously undermine the value, accessibility and sustainability of DSI.

The value of DSI

7. DSI holds tremendous value for understanding biology and evolution. The development of genomics is a direct result of DSI being freely available and shared globally in easy-to-access public databases. This widespread sharing of DSI is fundamental for advancing research and driving innovation, for example, by addressing conservation challenges, delivering actionable public health strategies and responding rapidly to global public health emergencies.

8. Public health emergencies require rapid responses. The rapid and unconstrained availability of DSI from genetic resources is critical for quickly determining how the disease is spreading, how the pathogen is evolving and identifying the genes involved in disease onset. This information is vital for informing emergency response tactics and public health strategies. This process is time-sensitive and delays due to the establishment and negotiation of licencing agreements on a bilateral basis could be disastrous for public health.

Global accessibility of DSI in easy-to-access databases

9. The Sanger Institute shares non-human DSI openly via the European Nucleotide Archive (ENA), which is maintained by EMBL-EBI. As of 2018, the ENA comprised a comprehensive database of 1.5 billion sequences and this number continues to grow[1]. Databases like the ENA collate, integrate, curate and make freely available scientific data from around the world. DSI within these databases is easily findable (via accession numbers) and accessible to researchers and publics around the world. The ENA offers a variety of support programmes and training packages to facilitate the use of DSI. This allows researchers who do not have access to the infrastructure and resources required for large-scale sequencing projects to still benefit from the resource.

10. Open access databases are founded upon the principle of open science. They typically have an intricate and integrated network of services, tools and data resources that exist to support the discovery, analysis and interpretation of DSI. Genome sequences are compared to thousands of other genomic sequences from different organisms, using tools such as BLAST, Pfam and InterPro, to identify features such as genetic regions, conserved genomic domains and protein structures. These comparisons of thousands of sequences provide insights into characteristics including virulence, metabolism and genetic lineage. The true value of DSI comes from aggregated datasets where patterns of conserved features or differences can be easily identified. The unconstrained aggregation of thousands of genome sequences is fundamental to providing biological insights in genetic resources.

11. **Restrictions on the open sharing of DSI in databases like the ENA and its affiliated network could create a major barrier to research and innovation across a whole array of research fields.**

Open access science

12. Since the 1990's, there has been a push towards open access science and genomics has been a global leader[2]. Initiatives like Plan S (a framework for transitioning to fully open access publishing)[3] are bringing funders together to ensure that research findings and associated data are freely and openly available to other researchers and the general public.

13. Open access to research data brings a far greater return on investment for taxpayers, provides researchers with the opportunity to work with data they may not have the capacity to generate themselves and circumvents the unnecessary duplication and substantial cost of reproducing DSI.

14. **Restrictions imposed on sharing datasets openly, by incorporating DSI into the Nagoya Protocol, will reduce accessibility to science, increase the cost of science and reduce the return on investment to the public.**

[1] Harrison, P.W. et al. (2019). The European Nucleotide Archive in 2018. *Nucleic Acids Research*, 47, D84-D88. http://europepmc.org/abstract/MED/30395270

[2] Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility, Report of a meeting organised by the Wellcome Trust and held on 14–15 January 2003 at Fort Lauderdale, USA.

[3] https://www.coalition-s.org/

Restricted data

15. Researchers are more likely to use resources, such as DSI, when access is not constrained by payment or access agreements.

16. The Sanger Institute shares data free of charge, but access to the majority of our human DSI requires a data access agreement because of the sensitivities around human genomic data. Anecdotally, we know that the requirement to sign a data access agreement can be a barrier to access, particularly for researchers in some low and middle income countries.

17. Development of data access processes is non-trivial, requiring dedicated sustainable resource to maintain them and more sophisticated researcher environments to use them.

18. **Imposing additional regulation and/or requirements for access agreements will discourage and limit use of that genetic resource, risking failure to achieve the goals and ambitions of the CBD.**

Access and benefit sharing arrangements

19. Existing Nagoya Protocol mechanisms are already impacting research at the Sanger Institute. Established research projects with long-term collaborators in Ethiopia and Gabon have a track record of training and capacity building, but have stalled due to uncertainties around the existing Nagoya Protocol. The inclusion of DSI within scope of the Nagoya Protocol will exacerbate these issues, cause delays to other research projects and hinder scientific progress.

20. International collaborations are imperative to our ability to tackle complex global challenges. We frequently collaborate with researchers in host countries when generating DSI from their sovereign genetic resources. These successful collaborations are beneficial to both parties for advancing research and scientific understanding as well as boosting recognition and capacity building efforts in host countries.

21. Countries exercising sovereign rights over their genetic resources are able to benefit from access and benefit sharing (ABS) arrangements in accordance with the existing Nagoya Protocol. Given that the generation of DSI requires physical samples of genetic resources, ABS can be achieved using existing Nagoya mechanisms.

22. The inclusion of DSI within the scope of the Nagoya Protocol will impose significant barriers to research and innovation and could ultimately risk inequitable ABS. For example, if DSI from a genetic resource is released freely and openly by one country, but the DSI from a very closely related genetic resource is constrained by its respective country, then investment and research is more likely to focus on the unconstrained, rather than constrained, genetic resource. In turn, resulting research outputs might only be relevant to specific geographical regions where DSI is freely available.

23. **Equitable access and benefit sharing can be achieved under the existing framework of the Nagoya protocol.**

Case Study 1 - Darwin Tree of Life Project

Reference genomes provide a template for assembling and comparing individual genomes of the same species. As of October 2017, the National Centre for Biotechnology Information (NCBI) database held 2,534 sequenced eukaryotic species genomes, which represents less than 0.2% of all known eukaryotes[4].

The Darwin Tree of Life project, launched in November 2018, seeks to sequence and create reference genomes for the 66,000 known eukaryotic species in the British Isles over the next 10 years. Led by the Wellcome Sanger Institute, this extensive project will draw on the expertise of several UK organisations, including the Natural History Museum, Royal Botanic Gardens Edinburgh, Earlham Institute, Universities of Cambridge, Oxford and Edinburgh, EMBL-EBI, Marine Biology Association and others for sample collection, DNA sequencing, genome assembly and annotation, and data storage. The project will feed in to the wider Earth BioGenome Project - a global collaborative effort to sequence the genomes of all 1.5 million known species of animals, plants, protozoa and fungi on Earth.

The Darwin Tree of Life and Earth BioGenome projects will revolutionise our understanding of biology and evolution. However, building such an urgently needed reference dataset for the planet's dwindling species list will require a multi-billion dollar global effort with massive automation of sample processing and sequencing. The subsequent sequence data needs to be stored in public domain databases and made freely available for wider research use. The project can only achieve its scientific goals of protecting biodiversity, supporting conservation and understanding the ecosystems around us if the data are used globally and by future generations.

The inclusion of DSI within scope of the Nagoya Protocol risks hindering the development of comparable projects in other countries and thus impeding access to the benefits of such endeavours. Any attempt to prohibit rapid and widespread sharing of invaluable DSI would prevent researchers and conservationists globally from advancing their research and responding promptly to existing and emerging threats to public health and biodiversity.

Case study 2 – Tracking the spread of cholera at the household level – Prof Nicholas Thomson

Every year, cholera causes up to 143,000 deaths and the disease-causing bacteria, *Vibrio cholerae*, infects up to 4 million people worldwide. Cholera is caused through the consumption of food or water contaminated with *Vibrio cholerae* and the disease remains to be a global public health threat. The capital city of Bangladesh, Dhaka, is a megacity and is hyper-endemic for cholera; experiencing two seasonal outbreaks every year. Very little was previously known about the diversity of the *Vibrio cholerae* strains that circulated around Dhaka and their role in the seasonal outbreaks. For the first time, to understand the diversity and transmission of cholera in an endemic setting, Wellcome Sanger Institute researchers and their collaborators tracked the transmission of cholera at the household level.

Vibrio cholerae was isolated from index patients who had been admitted to the International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b) and over a period of 3 weeks, follow-up samples were collected from household contacts who shared a cooking pot with the index patients.

[4] Lewin, H. A. et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. *PNAS*, 115, 4325-4333. https://www.pnas.org/content/115/17/4325

By sequencing these 303 *Vibrio cholerae* isolates from 224 individuals across 103 households, the researchers were able to determine how the cholera strains from each person were related and then compare them with other strains from around the world.

They found that nearly 80% of secondary infections were linked to the first index case within the same household, which implied that once cholera had entered the household, it spread between household members rather than repeatedly coming in from outside.

These findings strongly reinforce the importance of sanitation and hygiene to prevent the chain of transmission within households. These findings can be used by public health officials to improve cholera control strategies, for example, improving sanitation, water chlorination and vaccinating household members to help reduce the spread of this deadly disease.

The inclusion of DSI in the scope of the Nagoya Protocol could cause prohibitive delays to research projects like this and hinder rapid public health efforts to control disease epidemics.

Case study 3 – Mutational signatures as a tool for health research and conservation - Dr Alex Cagan

Mutational signatures as a tool for health research

Nearly 990 people in the UK are diagnosed with cancer every day and 1 in 2 people will be diagnosed with a form of cancer during their lifetime[5]. Cancer is caused through the accumulation of mutations in the DNA of somatic tissue. It might be expected that larger animals or those with longer lifespans should have a higher incidence of cancer, but this trend is not seen in nature.

In an ongoing project at the Wellcome Sanger Institute, researchers are comparing somatic mutation rates across a variety of different species. In collaboration with the Zoological Society of London (ZSL), UK Cetacean Strandings Investigation Programme (CSIP), the University of Cambridge and others, researchers are using DSI, histology and laser-capture microscopy to estimate mutation rates and identify specific mutational signatures across different species. This research will provide insights into the evolution of mutation rates across different species and help us to further understand cancer and aging in humans.

Researchers are particularly interested in animals with long lifespans and low mutation rates as well as those that have mechanisms for reducing their mutation rate. Studying these animals will help us to understand the somatic mutation rate in humans and could identify ways of reducing these rates to prevent human cancers. Many of these animals, however, are not native to the UK or EU and, as such, researchers have currently delayed their focus on these animals as a result of the uncertainties around the requirements of the existing Nagoya Protocol. Inclusion of DSI risks researchers excluding species from certain countries and exacerbating existing inequalities.

Mutational signatures as a tool for conservation

Polluting and toxic chemicals in the natural environment can cause an array of health concerns in both humans and in animals. Characteristic mutational signatures can provide insights into the threats posed to endangered animals.

[5] https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk

Polychlorinated biphenyl (PCB) is a highly toxic industrial compound known to cause cancer and developmental damage. Although its use is now widely prohibited, it is still persistent in the environment through spills, leaks and improper disposal and storage of industrial equipment. Like other carcinogens, PCB causes unique mutational signatures that enable researchers to identify specific environmental threats endangering different animals.

Biomonitoring of habitats using mutational signatures will help determine the impact on health and the threats posed to species by environmental change. This will help inform specific environmental management and conservational measures to prevent further endangerment of animals. This type of research will only be effective if DSI can be openly shared.

Case study 4 – Pathogen surveillance – Professor David Aanensen

Antimicrobial resistance is an increasingly serious threat to global public health that requires international action. Pathogenic bacteria infect millions of people worldwide every year and are responsible for many diseases including tuberculosis, pneumonia and typhoid. Many pathogenic bacteria have developed drug resistance and subsequently several antibiotics are no longer effective against these life-threatening diseases. Pathogens do not respect geographical borders and antimicrobial resistance is present in every country.

The Centre for Genomic Pathogen Surveillance (CGPS), based at the Wellcome Sanger Institute and led by Professor David Aanensen, performs global surveillance of pathogens using whole genome sequencing to understand the emergence and spread of diseases and drug resistance.

The CGPS combines structured population surveys and whole genome sequencing to identify specific pathogenic strains present in different regions and how they can spread within and between countries. This form of genomic technology is becoming the gold-standard tool for surveillance and is used to inform public health policies.

The surveillance data generated by the CGPS is freely and openly available in public repositories (e.g. the ENA) and the team is continually developing easy-to-use software tools to make data integration, visualisation and interpretation accessible to all. The CGPS invests in capacity-building in low- and middle-income countries to help train the future leaders of new national and emerging surveillance programmes.

Rapid sharing of surveillance DSI is vital for tracking and anticipating the geographical routes of diseases and drug resistance and for informing regional and national patient treatment strategies to prevent drug resistance spreading to other regions.

Pathogen surveillance requires research to be performed across geographical borders.

Case study 5 – Conservation of endangered gorilla populations – Dr Chris Tyler-Smith

Mount Tschiaberimu in the Democratic Republic of Congo (DRC) is home to a highly endangered population of gorillas. With a population of one female, four males and one baby of unknown sex, the colony is no longer viable and requires new gorillas to provide enough genetic variability for the colony to continue.

Living on Mount Tschiaberimu, they were thought to be mountain gorillas, but some unusual characteristics put this in to question. To confirm their species, researchers at the Wellcome Sanger

Institute alongside their US collaborators and conservationists from the DRC took a sample from a gorilla, which was sequenced at the Sanger Institute and then compared to openly available DSI from three gorilla species. The analysis showed that it was, in fact, not a mountain gorilla, but an eastern lowland subspecies. This information was necessary for developing a conservation strategy.

Identifying the gorilla species was only possible because the researchers were able to access DSI of the other gorillas.