

Wellcome Trust Sanger Institute

DATA SHARING POLICY

May 2014

The Wellcome Trust Sanger Institute is dedicated to advancing genetic and genomic science for the benefit of all. Rapid and open data sharing strategically supports this mission by enabling research and accelerating translation. However, such policies are only sustainable if scientific credit is generated for all parties involved, and the Institute will play its part in developing a global research environment which rewards data sharing.

The following principles form the basis for data sharing at the Wellcome Trust Sanger Institute; in addition guidelines for implementing the policy are provided in the associated document, WTSI Data Sharing Guidelines.

Access

The Institute aims to provide rapid access to data sets of use to the research community and will place these in publicly accessible repositories when possible. The Institute will support data and interoperability standards to maximise access and ensure ease of integration with other global resources.

Ethical Considerations

Conducting genetic and genomic research carries responsibilities to protect confidentiality and the privacy of research participants. Access to certain data sets will therefore be carefully managed and granted in a transparent manner to all appropriately qualified researchers.

Rights of Data Providers

The Institute recognises the need for researchers to be appropriately credited for their scientific contribution and investment in data generation. It is therefore expected that all researchers both honour agreements in line with Fort Lauderdale's data sharing principles¹, and appropriately acknowledge the contribution of others.

Optimising Translation

The Institute recognises that, in specific instances, the use of intellectual property protection and attendant potential delays to data sharing may be necessary to prevent inappropriately exclusive claims by others and to ensure health benefits occur.

¹ 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.

Wellcome Trust Sanger Institute
DATA SHARING GUIDELINES
June 2017

These guidelines are intended to provide researchers with practical guidance for implementing the Data Sharing Policy the Sanger Institute.

These guidelines will be kept under review and further developed by the Sanger Institute Human Materials and Data Management Committee (HMDMC). Please contact Hayley Clissold (Policy Officer datasharing@sanger.ac.uk) if you have any questions about the data sharing policy and guidelines.

Table of Contents

1. Release strategy	2
2. Data Types	2
3. Metadata Submission	3
4. Sharing Summary Statistics	3
5. Release Timing strategy.....	4
6. Exceptions to Data Release	5
7. Collaborations.....	6
8. Publication Moratoria.....	6
9. Audit	6
10. Anonymisation.....	6
Glossary	7

1. Release strategy

Data generated at the Sanger Institute is released *via* either open or managed access.

- **Open access data** are data that can be released into the public domain without restriction. The bulk of open data generated by the Sanger Institute is generated by studies using samples from animals or pathogens. These data should be submitted to the European Nucleotide Archive (ENA at EMBL-EBI). Some forms of pre-processed summary statistics can and should also be open access.
- **Managed access data** are data that may be released to researchers under certain conditions with restrictions on use and re-distribution usually related to the terms of consent given by research participants. Managed access datasets should be submitted to the European Genome-phenome Archive (EGA at EMBL-EBI).

2. Data Types

- Sequencing data** – The Sanger Institute will, where appropriate, release the following sequencing file types; raw BAMs, improved BAMs and vcfs (as detailed in Table 1)
- Genotyping data** – The Sanger Institute will, where appropriate, release the following genotype file types; iDAT files and Ped/map files (as detailed in Table 1)
- Functional Analysis Assay Data** - (e.g. obtained from microarray, SAGE (Serial Analysis of Gene Expression) and high-throughput sequencing studies, to describe gene functions and interactions, including gene transcription and translation (transcriptomics, proteomics and metabolomics)). Primary data sets of use to the research community should be submitted to ArrayExpress (EMBL-EBI) as soon as possible after generation, and definitely at publication.
- Mass Spectrometry** – Primary mass spectrometry data should be submitted to PRIDE (EMBL-EBI) as soon as possible after generation, and definitely at publication.
- Annotation Data** - Annotation data should be made available as they are generated via an appropriate browser (e.g. Ensembl, Vega, GeneDB) which allows users to both browse and export annotation to flat files. Where appropriate, annotation should be continuously available via Distributed Annotation System (DAS) sources, and registered in the DAS registry, so they can be displayed by any genome annotation application or website that is a DAS client. Annotations of sequence data generated at the Sanger Institute should be included in the final sequence entry submitted to EMBL-EBI.
- Other Biological/Biochemical Assay Data** - Other biological/biochemical assay data, such as the results of receptor-ligand interaction studies, images of histological assays, etc., should also be shared via the Sanger Institute or other suitable databases as soon as possible after generation, and definitely at publication (e.g., IntAct at EMBL-EBI for protein interaction data).
- Model Organism Phenotypic Information** – should be released as detailed above (i-vi) and links to the data (accession numbers) should be displayed on the project web page. Morphological and

other phenotypic data should be submitted to a the Sanger Institute database or other appropriate database (e.g., Mouse Resources Portal).

- viii. **Summary Statistics** – the Sanger Institute will, where appropriate, release summary statistics in an appropriate repository. Pre-processed summary statistics for GWAS will be made openly available and should be deposited in the Sanger Institute GWAS database or other appropriate repository no later than the time of publication.

3. Metadata Submission

For all human data that is submitted to the EGA (i.e., managed data access studies), the Institute should submit the following metadata as a bare minimum, no later than publication -

- Gender
- Phenotype (e.g., type of cancer sample)
- Ethnic Origin (where available/relevant e.g., population cohort studies, study of a disease in a specific ethnic group)
- Donor/subject ID

When submitting phenotypic data it is best practice to use controlled ontologies, further details of which can be found here <http://www.ebi.ac.uk/efo/>

For all data submitted to the ENA (i.e., open access), Institute researchers should comply with established minimal metadata deposition no later than publication. For further information please check the ENA website http://www.ebi.ac.uk/ena/submit/mixs-checklists#MlxS_shared to ensure that the level of metadata submission is sufficient to meet ENA requirements.

4. Sharing Summary Statistics

Before publishing/releasing human summary statistics, those responsible for the data should assess the risk of re-identification of individual research participants **and** the risk of harm should re-identification happen even if unlikely. This includes considering the nature of the study, the population under study and the type of traits studied, alongside the data fields being shared and whether it would be appropriate to apply any filtering on the values in a dataset in order to reduce the risk of individual re-identification.

For GWAS summary statistics data it is recommended that all the values contained in a dataset be **rounded up to 3 significant digits (e.g. 1.95x10⁵)**.

Data fields released as part of a data set are considered to constitute different levels of risk of re-identification.

Low risk:

- Variant identifier (rsid, chromosome position etc.)
- Effect allele and other allele
- Effect size and measure of uncertainty (e.g. standard error or confidence intervals)
- Association statistics (e.g. p-value, z-score)

Moderate risk – all of the above plus:

- Effect allele frequency
- QC metrics (call rate, imputation accuracy score, etc)

High risk – all of the above plus:

- Genome-wide pairwise LD measures (e.g. r²)

Whether to publish fields designated “moderate” or “high” risk should be considered in the context of each individual study. The risk of re-identification and subsequent harm associated with each data field will vary according to the nature of the study and the data set.

With regards to a dataset as a whole, researchers should consider the following factors that can contribute to lowering the risk of individual re-identification:

- **Allele frequencies** – the risk of identifying individuals increases when the datasets contain rare or very rare alleles.
 - Where data sets are generated from >10,000 samples the risk of re-identifying individuals from rare alleles is minimal.
 - Where data sets were generated from <10,000 samples the datasets should be filtered so that all the minor alleles with a frequency of <0.05 or down to 0.005 (depending on the number of samples in the study) should be excluded¹.
- **Number of variants shared** – In data generated from <10,000 samples the number of variants shared could be reduced². For example sharing variants in linkage equilibrium can reduce the risk of individual participant re-identification.

Note: The guidelines above should not be taken as absolute thresholds. The risk of re-identifying individuals is dependent on a combination of factors and not on one criterion alone.

GWAS Summary Statistics

GWAS summary statistics will be released and will contain, at a minimum, all the fields categorized as “low risk”. Researchers may wish to publish some or all fields from the higher risk categories where they deem it appropriate and they do not pose an increased risk of re-identification or harm as a result of re-identification.

5. Release Timing strategy

Data and File Type	ENA	EGA
Genotyping		
<i>iDAT</i>	12 months	6 months
<i>Ped/map files</i>	15 months	9 months
Sequencing		
<i>Raw BAMs</i>	12 months	6 months
<i>Improved BAMs</i>	15 months	9 months
<i>vcf</i>	To be released as soon as possible <u>and certainly no later than publication</u>	To be released as soon as possible <u>and certainly no later than publication</u>
RNA Seq	To be released as soon as possible <u>and certainly no later than publication</u>	To be released as soon as possible <u>and certainly no later than publication</u>
GWAS Summary Statistics	To be released as soon as possible <u>and certainly no later than publication</u>	To be released as soon as possible <u>and certainly no later than publication</u>

¹ Genomic Privacy and Limits of Individual Detection in a Pool.

<http://www.nature.com/ng/journal/v41/n9/full/ng.436.html>

² Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays.

<http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000167>

Table 1: Summary table for submission of open and managed sequencing and genotyping data.

The timings at which the respective data files attributed to sequencing and genotyping data should be submitted to major public repositories, following core pipeline QC. Summary statistics should be submitted at the same time as their respective datasets and no later than publication.

6. Exceptions to Data Release

Where researchers wish to delay or exempt data from release they should complete the “delay to data release/exemption” form: <https://helix.wtgc.org/node/448> and return it to datasharing@sanger.ac.uk. Delays and exemptions from data release may be granted for the following:

Sensitive Studies

Studies where data are socio-politically sensitive or there is a significant risk of harm to individual participants if they are re-identified, even where the risk of re-identification itself is low, can be delayed or made exempt from data release.

Bioterrorism

Where the release of data, in particular from pathogens, could lead to potential misuse or presents a potential (bio)security threat, data can be exempt from data release.

PhD Data

Data generated as part of a PhD project need only be released at thesis submission.

Capacity Building Projects

Collaborations with researchers in low-to-middle income countries may request a delay to data release to give those researchers the opportunity to develop their own skills and expertise in data generation and analysis.

Intellectual Property (IP)

Data sharing may be delayed to seek IP protection when this is necessary to optimize translation.

Other

If you have any other queries with regards to exemptions/delays to data release, please contact Hayley Clissold (datasharing@sanger.ac.uk).

Additional Considerations

Research in a Resource Project

Whilst data from large resource generating projects should be released rapidly (Table 1), some of these resource projects may contain a research element. In this instance, a delay should not be sought for the research element of the project, but the data should instead be protected by a publication moratorium. For example, the UK10K project released its data rapidly but chose to protect its cohort and exome data sets by the use of two separate publication moratorium dates.

Exemptions not requiring approval

Sequenom and cytogenetic data and data from replication studies of a subset of data previously released (genotyping or sequencing based) to validate initial findings need not be submitted to public archives.

Case Study for exemption to data release

Researchers studying the genomics of complex traits in a founder population requested an exemption to data release of individual level genomic data because personal identifying information and genealogies relating to the study population were publically available. It was agreed that in these circumstances individuals were potentially identifiable from the level genomic data and that only summary data should be made available.

Case Study for delay to data release

Researchers working with sample donors living in a low-to-middle-income country requested and were granted a delay to data release to give researchers in that country the opportunity to learn how to use and interpret the results of the study themselves before the data were made publicly available.

7. Collaborations

The Sanger Institute researchers are responsible for ensuring that collaborations respect the Institute's data sharing policy. For collaborations in which primary data are generated elsewhere, the data/results should be shared in a timely fashion, preferably in line with the Sanger Institute data sharing policy.

8. Publication Moratoria

In general, the Institute considers that the scientific credit gained from sharing data outweighs the risk to researchers of being "scooped" but researchers are allowed to claim "first publication rights" through use of a statement associated with the data access agreement. Where researchers are using a project specific data access agreement they should refer to the standard Sanger Institute data access agreement for wording.

9. Audit

Data sharing practices will be regularly audited by the Sanger Institute's HMDMC to ensure adherence to the data sharing policy.

10. Anonymisation

1. In all but exceptional circumstances, research data sets should be pseudonymised or fully anonymised (unlinked anonymised). Unless appropriate approvals are in place to collect, store and process personal data (information which allows identification of an individual) all data should be (pseudo)anonymised prior to receipt at the Institute. Researchers wishing to collect personal data should speak to the Ethics and Governance Office before starting work.
2. Pseudonymised data may be preferable using unlinked anonymised data sets because retaining the link allows for eventualities such as feedback of individual data, withdrawal of consent, expanding data sets, etc. However, the key to the link should remain with the clinical collaborator or other data provider (i.e., the key should not be held at the Sanger Institute).
3. General rules: Only the first 3 digits of post codes should appear and dates should be approximated to YY, or MM/YY if necessary. All unique identifiers (except for the alphanumeric linking code in the case of linked anonymised information), such as NHS number and National Insurance Number, should have been removed. Remember that nonymization is context-specific

– in particular, for information pertaining to patients with rare genetic disorders, removing the last 3 digits of post codes may not be considered sufficient to render the data anonymised.

List of Abbreviations

EBI	European Bioinformatics Institute
EGA	European Genome-phenome Archive
ENA	European Nucleotide Archive
HMDMC	Human Material and Data Management Committee
IP	Intellectual Property

Glossary

Primary data: refers to the lowest level of data that is considered useful to archive in public primary data repositories.

Managed access data: refers to data that cannot be made publicly accessible and is therefore shared with the research community *via* a managed access procedure.

Appropriately qualified researcher: we define an appropriately qualified researcher either as someone who has authored a relevant peer-reviewed article that we can locate on [PubMed](#), and who is still working in the field, or as a successful applicant to a relevant data access committee

Pseudonymised data (also known as coded or linked anonymised): are anonymous to the people who receive and hold it (e.g. a research team) but contain information or codes that would allow the suppliers of the data, such as a clinician, to identify individuals.

Fully anonymised data (also known as unlinked): contain no information that could reasonably be used by anyone to identify individuals. Any coded link to individuals must be irreversibly broken.

Personal Data: information that directly identifies or can easily be used to identify individuals. Common identifiers are: name, address, phone number, full date of birth, NHS number, full postcode.

Document History

version (date)	amendments
v1 (Jul 2009)	
v2 (Feb 2010)	clarifications: 1.1.3; 1.1.4; 2.1.2; 2.1.5; 2.1.7; 3.1.2; 3.6; 3.8; Annex A procedure: 2.2; 5.3; 8.2.1
v3 (Jul 2010)	clarifications: 1.1.5; 2.1.5; 2.2.3; 3.1.1; 4.1.1; 6.1.1; 7; document history; committee information procedure: 8.2.3; 6.1.2
V4 (Jun 2011)	clarifications: 2.3.1; 3.3.1 amendments: 3.1.1; 3.2.1; committee information
V5 (Nov 2012)	clarifications: 1.1.8 procedure: Annex B amendments: committee information
V6 (Oct 2014)	major re-write of entire guidelines
V7 (Nov 2014)	clarifications 3.iv ; 3.vi
V8 (Jan 2015)	addition: 4 Metadata Submission

V9 (Mar 2016)	Addition: Summary statistics data sharing
V10 (Jan 2017)	Update data sharing timelines
V11 (April 2017)	Remove differentiation between resource and research and change data sharing timelines to provide ENA-release with a longer time frame. (Section 5)
V12 (June 2017)	Change RNA seq release timings (Section 5).
V13 (July 2018)	Update contact details (Liz Easthope to Hayley Clissold).