Australian Microbiome Initiative Data Policy v1.1

1. Introduction

The Australian Microbiome Initiative[1] is generating a resource consisting of reference genomic datasets of microbial community composition with extensive contextual metadata, to support national research efforts in defining the primary drivers of environment productivity, and to aid monitoring the health of our environment and conservation.

The consortium reserves the right to conduct 'global analyses' across these amplicons, metagenomes, and metatranscriptomes reference datasets and publish the results in the scientific literature. However, in accordance with the Bermuda[2] and Fort Lauderdale[3] agreements and the more recent Toronto Statement[4], which provide guidelines for scientific data sharing, the consortium is committed to ensuring that data produced in this effort are shared at appropriate times and with as few restrictions as possible, to advance scientific discovery and maximise the value to the community from this Australian Government National Collaborative Research Infrastructure Strategy (NCRIS)-funded dataset.

This policy describes the data associated with the consortium, roles and responsibilities of various consortium members and data users, release schedules and communications/publications expectations.

2. Reference Dataset Description and overall data/information flow

The reference datasets to be produced by the consortium will cover three areas:

1. An amplicon sequencing resource to generate a comprehensive phylogenetic framework of microbial communities across Australian environments
2. Metagenomics resources to generate a microbial genomic baseline across a broad range of Australian environments
3. Metatranscriptomic resources to generate a dynamic microbial genomic baseline across a broad range of Australian environments

Consortium members will determine the experimental design for each of the study areas above. DNA and/or RNA will be extracted by consortium members and genomic data will be produced from several Bioplatforms network data generation facilities.

Table 1: Facilities generating sequence data

Omics Type	Technology	Facility
Marker gene surveys	Illumina sequencing	Ramaciotti Centre for Genomics, Sydney[5]
Genomics		Australian Genome Research Facility (AGRF), Melbourne[6]
Transcriptomics		

Following production, raw data will be uploaded to a password-secured central data repository held at Amazon Web Services (AWS), and managed by the Queensland Cyber Infrastructure Foundation (QCIF, University of Queensland, Brisbane)[7] on behalf of Bioplatforms Australia. To enable recovery in case of disaster, all data in the AWS repository will be mirrored at a second site in Brisbane that is managed by QCIF.
Metadata associated with each file and files names will be made publicly available via a web portal and associated Application Programming Interface (API), which is managed by QCIF for Bioplatforms[8]. These will include metadata

[1] https://www.australianmicrobiome.com
[2] https://wellcome.ac.uk/funding/managing-grant/statement-genome-data-release
[3] https://www.genome.gov/pages/research/wellcomereport0303.pdf
[4] https://www.nature.com/articles/461168a.epdf
[5] http://www.ramaciotti.unsw.edu.au/
[6] http://www.agrf.org.au/
[7] https://www.qcif.edu.au/
[8] https://data.bioplatforms.com/organization/about/australian-microbiome

relating to the origin of each sample analysed and methods used for the extraction of DNA / RNA, preparation of sequencing libraries and the generation of sequence data. Access to the actual data files via the web portal and API will be restricted to authorised users and will require authentication through password use.

The data will be licensed for use under a Creative Common Attribution License (CC BY 4.0[9]) with the appropriate acknowledgement as defined in our Communications policy[10].

Sensitive data or metadata (such as GPS coordinates of rare and threatened species) will be handled using the approach applied by the Sensitive Data Service developed by the Atlas of Living Australia[11]

The data generated in this project will be made available under open-access conditions to the international research community, through a variety of relevant established international data repositories such as the European Nucleotide Archive (ENA)[12] (See also Section 4 - Data sharing schedule).

3. Roles and Responsibilities

The membership of the groups detailed below is indicated in Table 2.

3.1 Data Sponsor

Bioplatforms Australia, as the Data Sponsor, undertakes the overall duties of ownership, and is responsible for the following tasks (in consultation with):

- Defining the purpose of the data items;
- Defining access arrangements;
- Authorising any Data Users;
- Appointing a Data Custodian for copies of the data stored at various sites/on various systems.

3.2 Data Producers

Two broad types of data will be produced: raw and processed. Raw and processed data will be produced from the facilities listed in Section 2.

Producers of both raw and processed data are responsible for:

- Assigning a Data Custodian for copies of the data stored locally;
- Data generation and temporary storage;
- Ensuring data use is compliant with this policy;
- Quality assurance.

3.3 Data Infrastructure Providers

Data infrastructure providers provide data storage and/or compute infrastructure for the raw or processed data, and are responsible for:

- Assigning a Data Custodian for copies of the data stored locally.

3.4 Data Custodians

The Data Custodian undertakes the day-to-day management of each item of data stored at various sites/on various systems, and is responsible for:

- Data storage and disposal on that system;
- Ensuring data use is compliant with this and other policies/agreements;
- Providing access to Data Users that have been authorised by the Data Sponsor;
- Ensuring that any Data User who is given access to the data is aware of any data use policies (including this Policy) and their responsibilities.

3.5 Data Users

Data users include all end-users of the raw or processed data generated by the consortium. These comprise consortium researchers, any collaborators, training dataset users and any other approved members of the international research

[9] https://creativecommons.org/licenses/by/4.0/
[10] https://www.australianmicrobiome.com/initiative-activities/comms-policy/
[11] https://www.ala.org.au/faq/data-sensitivity/
[12] http://www.ebi.ac.uk/ena

community.

The Data User is any party who has been granted access, by a Data Custodian, to any item of data. They are responsible for:

- Requesting authorisation from the Data Sponsor;
- Requesting access from the Data Custodian;
- Using and safeguarding information according to the conditions stipulated by the Data Sponsor and/or Custodian - including observing any relevant ethics approvals, legislation, data use policies (including this Policy and other relevant data use policies imposed by the Data Owner) and their responsibilities.

Table 2: Group membership and details of their roles within the consortium

Consortium member	Someone who has a contributed meaningfully to the science and/or management of the initiative, such as through active involvement in project development, working groups and panels, or contribution of samples
Data Sponsor	Bioplatforms Australia
Research Champions	Steering committee members in consultation with working groups where appropriate
Data Producers (raw)	Ramaciotti Centre for Genomics, Sydney Australian Genome Research Facility (AGRF), Melbourne
Data Producers (processed)	Australian Microbiome initiative Analytical team
Data Infrastructure Providers	Queensland Cyber Infrastructure Foundation (QCIF), Brisbane
Data Custodians	All groups above are required to appoint a designated Data Custodian to ensure data assets generated throughout this project are managed according to the requirements of this policy

4. Data Sharing Schedule

4.1 Data Sharing Schedule

Various data types will be made available throughout the multistep process of generating, processing, assembling, annotating and dispersing the reference datasets.

The data will be made openly available with no mediated access phase immediately following deposition of data into QCIF data repository, and from resources including International Data Repositories. Data sharing and collaborative interactions are encouraged to advance scientific discovery and maximise the value to the community from this Australian Government (NCRIS)-funded dataset.

4.2 Data and metadata Retention/Persistence for items held in the Bioplatforms Data Repository

As noted in section 4.1 (Data Sharing schedule), it is the objective that all high-quality data[13] generated in this initiative, will be made publicly available. The preferred method for public release will be through deposition in an appropriate discipline repository (e.g. an ELIXIR Core Data Resource[14] or ELIXIR Deposition Database[15] - all of which are intended for the long-term preservation of biological data for a global audience).

4.2.1 Retention: Regardless of whether data was submitted to an appropriate discipline repository or not, Bioplatforms will ensure that all data and metadata submitted as part of this initiative to the Bioplatforms Data Repository will be retained for the lifetime of the repository. This is currently defined by the operational horizon of

[13] Note that some data (e.g. from pilot studies or data that fails QC) will not be submitted to such discipline repositories
[14] https://www.elixir-europe.org/platforms/data/core-data-resources
[15] https://www.elixir-europe.org/platforms/data/elixir-deposition-databases

Bioplatforms, which is currently the next 5 years at least.

4.2.2. Functional preservation: Bioplatforms makes no promises of usability and understandability of deposited objects over time.

4.2.3. Authenticity: All data files are stored along with a MD5 checksum of the file content. This may be used for assessing the integrity of data items stored.

4.2.4. Succession plans: In case of closure of the Bioplatforms Data repository, best efforts will be made to integrate all content into suitable alternative repositories.

5. Communications expectations

All communications (scientific or general publications and presentations) that arise from the consortium's work will appropriately acknowledge the input of all relevant contributions. The expectations are detailed in the Consortium Communications Policy[16].

[16] https://www.australianmicrobiome.com/initiative-activities/comms-policy/