SCHEDULE 1:

Genomics for Australian Plants Initiative Data Policy v1

1. Introduction

The Bioplatforms Australia (Bioplatforms)-sponsored Genomics for Australian Plants Initiative[1] is generating a resource consisting of whole genomes, target capture and reduced representation (ddRAD) reference datasets for Australian native plants, to support national research efforts in Australian plants genomics, evolution and conservation.

The Consortium reserves the right to conduct 'global analyses' across these whole genomes, target capture and reduced representation (ddRAD) 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, Bioplatforms are 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 maximize 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. Plant reference whole genomes (WGS).
2. Target capture to generate a comprehensive phylogenetic framework for Australian plants.
3. Reference 'reduced representation' and target capture genomic datasets for threatened species requiring conservation management.

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: The following facilities will be generating sequence data:

Facility
Ramaciotti Centre for Genomics, Sydney[5]
Australian Genome Research Facility (AGRF), Melbourne[6]
ACRF Biomolecular Resource Facility (BRF), Canberra[7]

[1] http://www.bioplatforms.com/australian-plants/

[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://brf.anu.edu.au/

Following production, raw data will be uploaded to a password-secured central data repository held at Amazon Web Services (AWS), and managed by the Centre for Comparative Genomics (CCG, Murdoch University, Perth)[8] on behalf of Bioplatforms.

To enable recovery in case of disaster, all data in the CCG-managed repository will be mirrored at a second site in Brisbane that is managed by the Queensland Cyber Infrastructure Foundation (QCIF)[9]. Data from the QCIF mirror of the central repository will be made available for direct download via a password protected web interface[10] to authorised users (see Section 3).

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 CCG for Bioplatforms[11]. These will include metadata 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)

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[12]

If determined necessary by Bioplatforms, in consultation with various research champions, copies of the intermediate and analysed data may also be stored elsewhere. As noted above, when this option is executed, access to any copies of the data and metadata must be controlled under identical conditions as required for the primary copies.

Raw data will be shared with Consortium researchers and bioinformaticians.

Where appropriate Intermediate and/or analysed data will be uploaded by bioinformaticians to the secure central data repository in Perth and mirrored in Brisbane as described above. Data downloads and API access will be provided under the same set-up as for the raw data.

Ultimately, 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)[13] (See also Section 4 - Data sharing schedule).

3. Roles and Responsibilities

3.1 Data Initiators

Research champions, listed in Section 5, will assess the data request in consultation with Bioplatforms Australia and are responsible for:

- Outlining the scope of work and agreeing upon the analysis in consultation with Bioplatforms facilities;
- Provide the metadata information relevant for each piece of work.
- Consultation with Bioplatforms for the tasks outlined in Section 3.2.

[8] https://ccg.murdoch.edu.au/

[9] https://www.qcif.edu.au/

[10] https://downloads-qcif.bioplatforms.com/bpa/project/

[11] https://data.bioplatforms.com/organization/about/project

[12] https://www.ala.org.au/faq/data-sensitivity/

[13] http://www.ebi.ac.uk/ena

3.2 Data Sponsor

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

- 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.

3.3 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.

3.4 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.

3.5 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.

3.6 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 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.

4. Data Sharing Schedule and Data Persistence

4.1 Data Sharing Schedule

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

Broadly, this will fall into two phases: a "mediated-access" phase, where access to the data will be limited to members of the Consortium and other authorised parties; and an "open-access" phase where the data will be made openly available from resources including International Data Repositories.

During the "mediated-access" phase, the process for gaining authorisation to access the data is to email data.access@bioplatforms.com with name, affiliation, specific data for which access is being requested and a brief outline of the intended data use. This information will be assessed by the Data Sponsor, Bioplatforms Australia, and the appropriate Consortium research champion(s). If approved, Bioplatforms as the Data Sponsor will inform an appropriate Data Custodian to provide access. Data sharing and collaborative interactions are encouraged to advance scientific discovery and maximize the value to the community from this Australian Government (NCRIS)-funded dataset.

The "open-access" phase is set at 12 months from deposition of data into CCG data repository to allow the members of the Consortium and other authorised parties to progress analysis and publications, but not hold up the release of the data to be made openly available from resources including International Data Repositories.

Table 2: Data Release Timescales:

Data	**Schedule for release of data to authorised users during the "Mediated-access" phase**	**Schedule for public release of data - resulting in the "Open-access" phase**
All Data sets	Immediately following deposition of data into CCG data repository	12 months from deposition of data into CCG data repository

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[14] 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[15] or ELIXIR Deposition Database[16] - 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 Bioplatforms, which is currently the next X 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

[14] Note that some data (e.g. from pilot studies or data that fails QC) will not be submitted to such discipline repositories

[15] https://www.elixir-europe.org/platforms/data/core-data-resources

[16] https://www.elixir-europe.org/platforms/data/elixir-deposition-databases

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.

6. Consortium Group Members/Roles

Table 3: Group membership and details of their roles within the Consortium

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 ACRF Biomolecular Resource Facility, ANU, Canberra
Data Producers (processed)	
Data Infrastructure Providers	Centre for Comparative Genomics (CCG), Perth 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