Australian Avian Genomics Initiative Data and Collaboration Policy v1.0

1. Introduction

The Bioplatforms Australia sponsored 'Australian Avian Genomics Initiative' is generating a foundational biomolecular data asset to advance our understanding and conservation of Australia's unique native birds.

Partners of the Australian Avian Genomics Initiative (Consortium Members[1]) reserve the right to conduct 'global analyses' across these whole genomes, phylogenomics, population genetics reference datasets and publish the results in the scientific literature. Bioplatforms are committed to ensuring that data produced in this effort are shared at appropriate times and with as few restrictions as possible (with full consideration of any sensitive information – see section *3.4. Consideration for sensitive data* for more information), 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 Australian Avian Genomics Consortium, the roles and responsibilities of consortium members and data users, as well as release schedules and communications/publications expectations.

[1] A consortium member is 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.

2. Roles and Responsibilities

Table 1: Roles and Responsibilities of the Australian Avian Genomics Initiative collaborators and data users

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.
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 committee, or sample contribution.
Project Leader	The principal investigator and primary contact for an approved data project under the Australian Avian Genomics Initiative. Project Leaders are responsible for completing the scope of work agreed upon in their project plan.
Project Partner/s	The listed partners contributing to the creation and/or use of data on an approved data project under the Australian Avian Genomics Initiative. Project Partner/s are responsible for assisting the Project Leader in achieving the activities and outcomes agreed upon in their project plan. Project Partner/s also include involved end-users of the data product.
Project Collaborators	Project Leader + Project Partners.
'Omics Facilities	Bioplatforms-supports genomics facilities involved in the generation of data for the Australian Avian Genomics Initiative. 'Omics Facilities are responsible for: • Collaborating and providing guidance on project design, sample requirements and quality control metrics • Undertaking sequencing/data generating services • Sending generated raw data to the Bioplatforms Data Portal team to ingest into the database
Bioplatforms Data Portal team	The Bioplatforms Data Portal team are responsible for ingesting data from the 'Omics Facilities in to the Bioplatforms Data Portal as per the conditions outlined in this agreement
Data User/s	Data Users include all end users of the raw or processed data generated by the Australian Avian Genomics Initiative, including Consortium Members, Project Leaders, Project Partners. Data User/s are responsible for ensuring their use of data generated under this agreement is in line with the conditions stipulated in this agreement, including attributing and citing data appropriately.

3. Dataset resource description and sharing

The reference datasets to be produced by the consortium will include, but is not limited to:

1. Reference whole genomes (WGS) and transcriptomics
2. 'Reduced representation' and target capture genomic datasets to generate a population and phylogenomic information

3.1. Dataset creation

Datasets will be created under agreed projects[2], and will include a Project Lead and Project Partners (collectively called the 'Project Collaborators'). Project Collaborators, with support from the Program Manager and 'Omics Facilities, will determine the experimental design suitable for the project aims. DNA (RNA, if required) will be extracted by the Project Collaborators and sent to the relevant 'Omics Facilities for sequencing.

Table 2: Example of facilities generating sequence data (including but not limited to):

Facility (Genomics & transcriptomics)	Facility (Proteomics & Metabolomics)
Ramaciotti Centre for Genomics, Sydney[3]	Australian Wine Research Institute (AWRI) (SA)
Australian Genome Research Facility (AGRF), Queensland[4]	Bio21 Institute, University of Melbourne (VIC)
ACRF Biomolecular Resource Facility (BRF), Canberra[5]	University of South Australia
Genomics WA, Perth[6]	Proteomics International, UWA (WA)

3.2. Dataset management and storage

Following production by one of the sequencing facilities, raw data will be uploaded to a password-secured central data repository managed by Bioplatforms Australia via the Queensland Cyber Infrastructure Foundation (QCIF, University of Queensland, Brisbane)[7]. This data is made discoverable and accessible through the Bioplatforms Australia Data Portal interface (https://data.bioplatforms.com/). This database is held at Amazon Web Services (AWS) Sydney location and mirrored at a second site at QRIS-Cloud Brisbane to enable recovery in case of disaster.

Metadata associated with each file and files names will be made publicly available (except where information considered to be sensitive – see Section 3.4) via the web interface and associated Application Programming Interface (API). These will include metadata relating to each sample analysed and methods used for the extraction of material, preparation of sample libraries and the generation of 'omics data. Access to the data files via the web portal and API will be restricted to authorised users and will require authentication through password use. User support guides[8] are available to facilitate access and use of the Data Portal.

All data will be licensed for use under a Creative Commons Attribution License (CC BY 4.0[9]) with the appropriate acknowledgement as defined in Section 4 (Data Use Attribution and Initiative Communications).

3.3. Dataset sharing schedule

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

Broadly, this will fall into two phases:

1. a "mediated-access" phase, where access to the data will be limited to members of the Consortium and other authorised parties; and
2. an "open-access" phase where the data will be made openly available on the Bioplatforms Data Portal and other resources including International Data Repositories. Please note that full consideration will be given to sensitive data and information (please see section *3.4. Consideration for sensitive data* for more information)

[2] Projects will be established through an open call for project plans based on a set of priorities. Projects are usually focused on generating data for a particular species or genus. Project plans include sampling strategies, sequencing strategy and final data product aim (e.g. creating a reference genome for Species X).

[3] http://www.ramaciotti.unsw.edu.au/

[4] http://www.agrf.org.au/

[5] https://brf.anu.edu.au/

[6] https://www.genomicswa.com.au/

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

[8] https://usersupport.bioplatforms.com

[9] https://creativecommons.org/licenses/by/4.0/

For the, the “mediated-access” phase will be active for up to 12 months from the creation of data, unless mutually agreed otherwise. The “mediated-access” phase enables primary access of the dataset to the Project Leader, and their partners, to provide first opportunity for use and publication. During the “mediated-access” phase, the process for gaining authorisation to access the data is to:

- Register for a Bioplatforms Data Portal account at https://data.bioplatforms.com/user/register
- When asked your reason for request, please include information on the project/s you are involved in, or interested in accessing, and your intended use of the data
- When asked which initiative/organisation you would like access to, check the Australian Avian Genomics Initiative box
- This will be sent to the Bioplatforms Program Manager to approve and/or mediate access/collaborations with the Project Leader

During the “open-access” phase, the process for gaining access to the data is to:

- Register for a Bioplatforms Data Portal account at https://data.bioplatforms.com/user/register
- When asked your reason for request, please include information on the project/s you are interested in accessing, and your intended use of the data
- When asked which initiative/organisation you would like access to, check the Australian Avian Genomics Initiative box
- This will be sent to the Bioplatforms Program Manager to approve and/or will be automatically approved by the system. You will have access to download any datasets that are outside of the “mediated-access” phase.

Table 3: Data Release Timelines:

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 datasets	Immediately following ingest of raw data from sequencing facilities in to the Bioplatforms Data Portal	12 months from ingest of raw data in to the Bioplatforms Data Portal

Datasets that are derived from analysing the raw data created through this project are considered to be analysed data products. Examples of these include, but are not limited to, reference genome assemblies, genome annotations, population genetics metrics, metabolite profiles/compounds, protein profiles/compounds. Analysed data products are created by Consortium members and can be uploaded to the Bioplatforms Data Portal, where data downloads and API access will be provided under the same set up as for the raw data. However, it is encouraged that analysed data products be published in discipline recognised international repositories (e.g. assembled genomes on NCBI) to enhance discoverability and access.

3.4. Consideration for sensitive data

It is generally agreed that the molecular data generated by this project is not considered to be sensitive in its own right. However, if during the course of generating this data the Project Collaborators and/or Consortium Members decide that the data, if published, would cause detriment to organisational or jurisdictional operations then the data will be held privately until mutually agreed otherwise. Likewise, metadata that is agreed to be sensitive (for example, latitude and longitude coordinates) will not be stored or published on the data portal. This metadata will be handled by the Project Leader, and requests for access will be deferred to the Project Leader to fulfil.

3.5. Data and metadata retention

It is the objective that all high-quality data[10] 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

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

Resource[11] or ELIXIR Deposition Database[12] - all of which are intended for the long-term preservation of biological data for a global audience). Access to any copies of the raw data or metadata must be controlled under identical conditions as required for the primary copies.

3.5.1 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 Portal will be retained for the lifetime of the repository. This is currently defined by the operational horizon of Bioplatforms.

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

3.5.3. 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.

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

4. Data Use Attribution and Initiative Communications

The Australian Avian Genomics Initiative has a strong interest in seeing that the data and capability produced by the project is actively communicated to and used by the scientific and wider communities. All users of the data and capability produced through the project are expected to exhibit professional courtesy and utilise the data with the highest scientific integrity. When a substantial proportion of the data has been provided through the research activities of others, Bioplatforms strongly encourages investigators to discuss details and expectations of the use of the data, as well as matters of authorship and acknowledgement on publications, with listed Project Leaders prior to the start of the study.

4.1. General request for all Communications

All communications (scientific or general publications and presentations) that arise from the Australian Avian Genomics Initiative leaders work will appropriately acknowledge the input of all relevant contributors. The publications should specify the collaborative nature of the project, and authorship is expected to include all those contributing significantly to the work (see sections below).

Acknowledging the Australian Avian Genomics Initiative consortium
Add the following text to the acknowledgements for general consortium recognition:

"We would like to acknowledge the contribution of the Australian Avian Genomics Initiative in the generation of data used in this publication. The Initiative is supported by funding from Bioplatforms Australia (enabled by NCRIS)."

If relevant, also credit other organisations involved in collection of the particular dataset you are using (as listed in 'project_lead' and 'project_collaborators' in the metadata record).

Citing an Australian Avian Genomics Initiative dataset
Add the following text to your citation list:

Australian Avian Genomics Initiative, [year-of-data-download], [full dataset title], [dataset-access-URL], accessed [date-of-access].

4.2. Use of data by investigators within or outside of the Consortium

Users of the Australian Avian Genomics Initiative data, whether members of the Consortium or not, should be aware of the publication status of the data they use and treat them accordingly. For example, all investigators including data from

[11] https://www.elixir-europe.org/platforms/data/core-data-resources
[12] https://www.elixir-europe.org/platforms/data/elixir-deposition-databases

other Consortium members should discuss use with the Project Partners before using unpublished data in their individual publications (see the guidelines below).

Investigators outside of the Consortium are free to use data generated by the initiative, but are asked to follow the guidelines developed at the Ft. Lauderdale meeting[13] and the Toronto Statement guidelines[14]. Specifically, data users should cite the source of the data and should acknowledge the Project Partners and funders of the Australian Avian Genomics Initiative. In addition, the data users are asked to recognise the interests of the Project Partners in publishing reports on the generation and analysis of their data, as described previously. Datasets from the Australian Avian Genomics Initiative may be released to public databases as pre-publication data and remain unpublished until the datasets appear in peer-reviewed publications. Investigators who are not a part of the consortium, who have performed an in-depth analysis of the data and are interested in publishing a report before the Project Partners have done so, should discuss their results with the Project Partners and are encouraged to establish collaborations.

The following guidelines may be used for acknowledgement and co-authorship, but specific details should be worked out between parties beforehand.

As a guideline, in addition to the acknowledgement of the Australian Avian Genomics Initiative:

- Small and complementing dataset used: Acknowledgement of the Project Partners is optional
- Significant but disconnected data used (from many different Project Partners): Consortium and Project Partners should be contacted for discussion, and receive an Acknowledgement
- Significant portion of data used – integral to the publication: co-authorship to be offered [Project Partners can naturally decline]

The Australian Avian Genomics Initiative Manager and/or Scientific Lead can also facilitate communication with Project Partners.

4.3. Reporting of communications

Copies of communications are to be sent to the General Manager Scientific programs (srichmond@bioplatforms.com) and the Program Manager (smazard@bioplatforms.com) for collection and circulation to the Consortium as appropriate.

For further information, please contact

Program Manager – Sophie Mazard (smazard@bioplatforms.com)
General Manager Scientific programs – Sarah Richmond (srichmond@bioplatforms.com)

[13] http://www.sanger.ac.uk/legal/assets/fortlauderdalereport.pdf

[14] https://www.nature.com/articles/461168a.epdf