Australian Genomics Policy on Data Access and Sharing for Secondary Use

The purpose of this policy is to generate a framework around reasonable access to the data collected through clinical flagship projects in the Australian Genomic Health Alliance (Australian Genomics) for the purpose of secondary research use. This policy is designed to reflect the opinions of flagships investigators, and respect the wishes of flagships participants: data collected will strictly be used in accordance with a participant’s consent for collection, use and data sharing. Some participants will dynamically and granularly update their consent preferences through an online platform (CTRL). The expressed or withdrawn consent is taken into account during the data access process.

Australian Genomics data access policies and agreements comply with any relevant Australian laws (federal, state and territory); Commonwealth laws, including relevant Acts (such as the Privacy Act (1988) and its Australian Privacy Principles (APPs); the Health record Act (2001) and the National statement on Ethical Conduct in Human Research (2018)). This policy also aims to align with the guidelines and policies set out by the Global Alliance for Genomic Health (GA4GH) with the Framework for responsible sharing of genomic and health-related data. See Appendix 1 for a summary of the principles and core elements.

Scope

This data access policy will explore reasonable access for researchers and clinicians both internal and external to the Australian Genomics collaboration (as defined below). The focus of this policy is data resulting from participants’ genomic sequencing as well as data captured through their clinical care (such as pathology results, reports) or research participation (such as survey responses).

Definitions and roles descriptions used in this policy:

Data Access Committee: The Data Access Committee is objectively and systematically reviewing data access requests from Data Requestors.
For Australian Genomics: comprised of a minimum three available parties of the following list: Australian genomic manager, Data Steward, one of the Program 2 lead, and/or an invited clinical member of the relevant flagship(s) to each access request.

Data Author: Lead investigators of clinical flagship projects enrolling participants into the study.

Data Custodian: The Data Custodian is responsible for the safe custody, aggregation and storage of data.
For Australian Genomics: University of Melbourne and Murdoch Children’s Research Institute.

Data Donor: The research participant whose data have been collected, held, used and potentially shared.
Data Owner: From the National Statistical Service: “The owner of intellectual property and materials (including data) produced by publicly funded research in Australia is usually the organisation administering the public research funds.” Note, that philosophically, Australian Genomics considers our research participants, the Data Donors, have a moral – if not legal – interest in the data generated from the samples they donate to research.

For Australian Genomics: Murdoch Children’s Research Institute.

Data Steward: The role of a Data Steward is to manage, oversee and utilize the organisation’s data governance processes to ensure fitness of data elements - both the content and metadata.

For Australian Genomics: Manager and Program 2 Data Coordinator.

Data User or Requestor: researcher or clinician seeking access to data.

Data Sharing: from the GA4GH Privacy and Security policy 4.3: “For the purpose of this Policy, “Data Sharing” includes, but is not limited to, the use, viewing, transfer, linkage, or exchange of Data between the Data Donor and another party, or between the Data Steward and a third party, either openly or under specified access conditions. Data Sharing may occur without having the data move from one place to another.”

De-identified Data: data with all its related Identifiers and other indices removed, such that there is a very small risk of re-identification of any data that are made accessible to researchers.

External Access: data access request(s) from an individual affiliated with an institute that is not a partner of Australian Genomics (see 'Internal Access').

Genomic Sequencing Data: Binary alignment Map (BAM); Variant Calling Format (VCF) and FASTQ text file format storing biological sequence and corresponding quality scores.

Human Research Ethics Committees (HREC): review all research proposals involving human participants to ensure that they are meet ethical standards and guidelines.

Identifier or Identifying Information: Information where the identity of an individual is apparent or can reasonably be ascertained by the holder of the information. Information that may directly, or indirectly, lead to identifying individuals from whom the samples and associated information are collected as a link (or multiple links) exists between the participant’s personal Identifiers and the data.

Internal Access: data access request(s) from an individual affiliated with an Australian Genomics partner institute, i.e. a participating institution as a signatory of the Australian Genomics Multi-Institute Agreement or Secondary Research Collaboration Agreement.

Metadata: is data about data (including but not limited to: quality metrics; methodological information; laboratory information; and possible coded Identifiers and clinical information).

Moratorium: is a formal request by an author to restrict access to documents or data for a specified period of time.
Type of data stored:

Within the scope of Australian Genomics, data is stored in, or accessible from, different locations depending on its type. Attached in Appendix 2 is the list of data types and tools created, or under development, including contacts for each tool. A brief summary is listed here:

- **REDCap study Database** - The REDCap study database stores identifiable data, including but not limited to: personal information; demographics; clinical (phenotype) and genotype information; pathology reports and results; study survey responses; study logistics and tracking information. REDCap is hosted on a secure server at Murdoch Children’s Research Institute. This set of data is only made available according to tiered permission levels, determined by user role and geographic responsibility. Outputs of De-identified Data can be requested as per the contacts listed.

- **Standardised, computer-readable clinical descriptions (FHIRCap)**: The clinical information in REDCap is translated into standardized phenotype ontology codes to allow computerized search, discovery and coding.

- **Genomic Data Repository (GDR)** - The raw genomics sequencing data (BAM/VCF/FASTQ files) as well as a minimal subset of Metadata is stored in the GDR, managed by University of Melbourne.

- **Variant Atlas** - Aggregated variant data (VCF) of each clinical flagship cohort can be visualised and filtered by key clinical features via the tool hosted at Garvan Institute.

- **Shariant** - Clinically curated variant classifications, the evidence underpinning the classification, and information of the providence of the classification (submitting laboratory and/or curator) is shared via the online platform Shariant, hosted by SAPathology.

In support of data re-analysis and sharing, the data and any associated Metadata are to be kept accurate, verifiable, unbiased and annotated with providence and date. Data integrity check shall be performed on every genomic sequencing file (FASTQ/BAM/VCF) stored on the GDR as a two-step process. Firstly, a corruption check to ensure the file is complete and not compromised. Secondly, a data quality control generating a summary report and visualisation on the depth of read and other quality associated metrics, as well as a comparison of the quality of each file with that of its own cohort.

The use and sharing of data are to be limited to what is relevant and necessary to accomplish the pre-defined research purpose; minimise privacy risks for participants; and be in line with participant consent.

**Timeline for access to genomic data:**

The primary purpose of the data repository is to facilitate ethical access and sharing of genomic data, while securely storing sequence genomic as well as a subset of related health and meta-data in support of future research re-analyses. This is crucial to maximise research knowledge, clinical understanding and patient benefit.

In recognition of the role of flagship clinicians and researchers as data Authors and lead investigators, a Moratorium on sharing of the data will be put in place. This Moratorium will be 6
months for Internal and 12 months for External Access to the data, starting at the end of the flagship
recruitment. After this Moratorium period, data access requests will be accepted for consideration
as per the processes described below.

Any publication resulting from analyses of the data, or part thereof, collected and stored by
Australian Genomics shall acknowledge the Australian Genomics Health Alliance (NHMRC
GNT1113531) and the relevant Data Author(s). It should also cite the REDCap database if using any
data that has been stored during its lifetime in REDCap.

Data Access Process: Background

The Australian Genomics access identification process follows the registered access policy model as
set out by GA4GH\(^\text{viii ix}\) and should be limited to bona fide researchers and clinicians with approved
HREC relevant for the research purpose. Following is an extract from “Registered Access: authorizing
data access” by S. Dyke et al:

“Our model of registered access in the GA4GH context comprises a three-stage “Triple-A registration”
process (Authentication, Attestation, and Authorization [9]), which aims to ensure both user
identification and agreement to a standard set of general responsibilities while considerably
simplifying the data access application process. Through the identification and authentication process,
the individual provides “proof” that an asserted identity is their own. The attestation process
establishes that the potential Data User meets the requirements expected by the consent agreements
and ethical approval of datasets in question and includes agreement to comply with the terms of data
use required of registered users. Finally, authorization is the overall process by which users are granted
access to data and permission to perform specific actions.”

Extract from “Global Alliance for Genomic and Health: Privacy and Security Policy”:

“Requests by Data Users to Data Stewards for access to Data should demonstrate, at a minimum: (1)
legitimate intended uses that are in the public interest (i.e., securing an objective commonly valued by
society) and with regard to established human rights; (2) assurances that Data are being accessed only
by authorized individuals, e.g., accredited persons accessing Data that will be held and used only in
safe environments; (3) a legitimate and specified time period of access; and (4) secure disposal or
return to the Data Steward of the Data after use and outside of any required retention period.”

The data access request could thus, eventually, be automated for simple and straight forward
requests. For more involved access demands, the Data Access Committee shall be reviewing the
request.

Once the bona fide status of the requestor has been confirmed, (s-)he would have access to the data
tools and data sets inter-crossing his/her research purpose and participants consents via the use of a
Library card type model'. Following is an extract from “Simplifying research access to genomics and
health data with Library Cards” by M. Cabili et al, describing the Library card type model:

“The Library Card model consists of four elements:

1. A standard real-world process for verifying a researcher’s identity, and relating it to an
online identity
2. A standard model for electronically representing a researcher’s identity and access-related information

3. A standard model for a DAC to electronically represent their policy for who is allowed to access a dataset

4. A standard protocol for servers to authenticate and authorize a user who is requesting access to a dataset

Internal data access – for a requestor involved with the Australian Genomics program or affiliated at a participating institution (as defined above), the Authentication process is pre-approved as the bona fide status of that researcher is known. Further, the terms of research collaboration have been defined and agreed under the Multi Institute Agreement, or other Collaboration Agreement. Should the proposed research application of the requestor be defined under Australian Genomics Ethics (MH2016/224), Attestation has been established – so Approval shall be granted. Should the internal data access request be for research use beyond the scope of Australian Genomics Ethics, the proposal shall be reviewed for HREC approval and alignment with the permissions associated with the data to establish Attestation, and Approval granted.

External data access - For requestors seeking External Access, an Authorisation process must be conducted to establish bona fide researcher status, before review of the proposed research use for Attestation, confirmation of HREC for the study, and Approval sought from the Data Access Committee.

For both, internal or External Access requests, data access is tiered by the sensitivity of the data (re-identifiability), permissions associated with the data:

- **Tier One** – Individual (unit) level, identifiable data – this is identifiable clinical and genomic information that is available only to the participant’s treating team, and Australian Genomics Data Stewards. These data will not be available for research access/sharing.

- **Tier Two** – Unit level, coded, De-identified Data – when genomic and clinical information is identified only with the Australian Genomics study Identifier, and has all personal information removed, it is referred to as ‘coded’ or ‘de-identified’. While there is a slim chance of re-identification of participants with this data – so it is still considered sensitive – the Data Access Agreements stipulate that any attempt to re-identify individuals is expressly forbidden. These data will be made available to bona fide researchers who demonstrate HREC permission for use of data of this tier, for a purpose in line with data permissions.

- **Tier Three** – Aggregate, anonymous data – these are the data sets that are not identified with personal information nor code, and are made available as a cohort. Given the chance of re-identification is low-to-nil, these datasets are ‘low sensitivity’.
Figure 1, below, is showing the spectrum of identifiability as per the explained tiered data sensitivity.

Datasets Metadata will remain discoverable, in order for a Data Requestor to be able to search and find relevant data sets prior to sending a request access to the Data Steward.

*Figure 1: Spectrum of identifiability as per the tiered data sensitivity. Re-created from Understanding Patient data*

**Audit to be performed on data usage:**

Audit shall be performed on a regular basis on the data usage in order to ensure access is granted to relevant parties and used appropriately. Audits would also allow for feedback and potential improvements on the usefulness of the repository and systems in place. Audit can be performed by means of approved requestors or their institution providing an annual update and a list of resulting publications and outcomes. Timely reports will ensure maintenance of access to required data. In the case of overdue reporting, access will be suspended by the Data Steward until the audit report is completed.
Data Breach process:

Should you suspect or are made aware of a data breach is or has occurred, please immediately notify: security@australiangenomics.org.au.

In this alert email please add any information that you are aware of regarding the data breach such as, but not limited to, what type of data is affected; which individual or cohort of individual is affected; the time when the breach first occurred; the type of breach (see Australian Genomics Data Breach Policy); etc. The data breach response team, composed of the manager, Data Steward, as well as relevant members of the data Custodian and data Owner institutions, will then follow the relevant steps to contain, stop and, where possible, revert the data breach as mapped in the Australian Genomics Data Breach Policy. The data breach response team will then contact you for any further details if required.
Appendix 1:

The foundational principles and core elements for responsible sharing of genomic and health-related data as highlighted in the GA4GH guidelines are as follows and founded on Article 27 of the 1948 Universal Declaration of Human Rights:

**Principles:**

- Respect Individuals, Families and Communities
- Advance Research and Scientific Knowledge
- Promote Health, Wellbeing and the fair Distribution of Benefits
- Foster trust, Integrity and Reciprocity

**Core elements:**

- Transparency
- Accountability
- Engagement
- Data quality and security
- Privacy, data protection and confidentiality
- Risk-benefit analysis
- Recognition and attribution
- Sustainability
- Education and training
- Accessibility and dissemination
Appendix 2: Australian Genomics Data Management

REDCap
REDCap is the Australian Genomics study database. The clinical, demographic and survey data of all our participants is stored and managed in REDCap, which is hosted at MCRI.
Contact: Keri Pereira keri.pereira@vcgs.org.au
Stefanie Elbracht-Leong stef.elbrachtleong@mcri.edu.au

Standardised, computer-readable clinical descriptions (FHIRCap)
The clinical information in REDCap is translated into standardized codes (phenotype ontologies: HPO/SNOMED) to allow computerized search, discovery and coding.
Contact: Alejandro Metke Alejandro.Metke@csiro.au
David Hansen David.Hansen@csiro.au

Genomic Data Repository (GDR)
The cloud-based GDR stores Australian Genomics’ Genomic Sequencing Data (BAM/VCF/FASTQ), and is hosted at University of Melbourne. A subset of a participant’s clinical data is linked to their genomic data, to facilitate meaningful analysis and sharing.
Contact: Lavinia Gordon gordonl@unimelb.edu.au
Oliver Hofmann oliver.hofmann@unimelb.edu.au

Variant Atlas
Variant Atlas is an interactive Genotype-Phenotype data platform, hosted at Garvan Institute. With this tool, researchers can visualise aggregated Flagship variant data, and filter by key clinical features to describe the cohort.
Contact: Warren Kaplan w.kaplan@garvan.org.au

Shariant
Shariant is an online platform for laboratories to share curated variant classifications. Hosted at SAPathology, Shariant builds upon international variant databases like ClinVar to display not only the variant information – but also the submitting lab, and the evidence underpinning the classification.
Contact: Emma Tudini Emma.tudini@qimrberghofer.edu.au
Amanda Spurdle Amanda.Spurdle@qimrberghofer.edu.au

Data access Agreements and Policies
Our participants provide consent to both national and international Data Sharing, for the benefit of healthcare. Clinicians and researchers can request access to Australian Genomics datasets, and this will be granted according to the level of data sensitivity; the specific consent of the participant; and the researcher’s HREC, where applicable. Australian Genomics ascribe to GA4GH policies and standards.
Contact: Stefanie Elbracht-Leong stef.elbrachtleong@mcri.edu.au

For general Australian Genomics Data Management enquiries, contact Marie-Jo Brion, Program Two Manager: m.brion@garvan.org.au

australiangenomics.org.au
## Policy Revision History

<table>
<thead>
<tr>
<th>Policy Version</th>
<th>Date effective</th>
<th>Summary of Revision</th>
</tr>
</thead>
<tbody>
<tr>
<td>V 1.0</td>
<td>Ratified by NSC</td>
<td>Original Document</td>
</tr>
<tr>
<td></td>
<td>November 2018</td>
<td></td>
</tr>
</tbody>
</table>
References and bibliography:


vi HRECs definition from NHMRC: https://nhmrc.gov.au/research-policy/ethics/human-research-ethics-committees


xi Understanding Patient data website http://understandingpatientdata.org.uk/what-does-anonymised-mean