National Genomics Infrastructure Briefing

Australian Genomics Implementation Progression Plan

June 2021

EXECUTIVE SUMMARY

Purpose

This report summarises key implementation elements of the Blueprint for a National Approach To Genomic Information Management (NAGIM) and other key genomic reports, and presents the Australian Genomics strategy for developing national implementation recommendations.

Summary of current infrastructure recommendations

Recommendations for national genomic data management were issued in the **GHFM Operational Plan (2018),** by the Information and Communication Technology (ICT) subcommittee to the Genomics Health Futures Mission (GHFM); and in the **NAGIM Blueprint (2020),** led by Queensland Genomics, based on extensive national consultations.

The NAGIM and GHFM ICT recommendations on infrastructure indicate that:

- It is **unlikely a single fully centralised system for genomic data storage and analysis** can meet the needs of both research and clinical users nationally;
- A data model based on **federation across multiple repositories** using standardised interfaces will be necessary, especially for clinical data; however, a **single unified repository for research data** may be tractable and would be preferable;
- There is growing preference for **cloud-based models** for international genomic data solutions, rather than models based on centralised high-performance computing (HPC) ;
- **Decisions on potential platforms should be based on evaluating working prototypes managing data across multiple jurisdictions.**

Developing a **national proof of concept for data sharing** across clinical and research settings was also identified as a key Implementation Action Item of the Department of Health's National Genomics Policy Framework and Implementation Plan (National Framework, 2018).

Australian Genomics has now been tasked by the Federal Government to develop implementation recommendations for a national approach to genomic data management.

Proposed approach

Australian Genomics will develop **a series of phased recommendations** that will be based on:

- Outcomes from clinical and research stakeholder engagement;
- Infrastructure prototyping of components that meet NAGIM recommendations;
- International evaluation and best practice

National infrastructure stakeholders will be invited to participate in prototype construction, definition of technical specifications, and infrastructure evaluation. This will be designed as an open process to facilitate communication and collaboration between groups involved in developing prototypes, with the goal of **identifying the best combination of components that can serve as the basis for long-term national infrastructure.**

Clinical and research stakeholders will be engaged in parallel to identify and progress additional core elements of the NAGIM ecosystem, including priorities for clinical data and data governance.

Delivery of Recommendations

Phase	Outcome	Components
1	Summary briefing (this document)	● NAGIM implementation elements mapped ● Proposed strategy for recommendations
2	Initial recommendations	● Research prototyping specifications defined ● National research and clinical stakeholder engagement
3	Prototype-informed recommendations	● Functional research prototypes built and evaluated ● Additional priorities identified for clinical and governance
4	Final recommendations for implementation	● Framework for national scale up ● Evaluation of clinical solutions ● International evaluations and best practice

Timeline

Phase	Outcome	Delivery Timeframe
1	Summary briefing and proposed approach	2021 June
2	Technical specifications	2021 July
3	Prototyping	2021 December
4	Evaluation	2022 January
5	High-level recommendations	2022 February
6	Comprehensive recommendations	2022 May

1. INTRODUCTION

Background: the 'National Framework'

The **National Health Genomics Policy Framework** published by the Federal Department of Health and the Australian Health Ministers Advisory Council (National Framework, 2018-2021) set the direction for a nationally coordinated approach to genomics in Australia. The aim of the National Framework was to avoid duplication of effort and leverage current activities, to drive improvements in health outcomes for Australians and provide a pathway to personalised health care.

'Data: the responsible collection, storage and management of genomic data' - was one of five strategic priorities in the National Framework. The following **'**national implementation action items' have been set for Data, under the Framework's Implementation Plan:

Action 19: Develop a national genomic **data governance framework** that provides for appropriate decision-making for governments and aligns with international frameworks.

Action 20:
A: Adopt international best practice standards on **cybersecurity and privacy standards** for genomic data systems and data sharing, across all levels of the health system, including consideration of vulnerable populations.
B: Consider the national adoption of appropriate **international standards** on (but not limited to) phenotypes, disease classification systems, and pathogenic variants.

Action 21:
A: Leverage opportunities for integration of individual genomic information with **electronic health records** (including, but not limited to, My Health Record) in ways that maintain public trust
B: Explore opportunities to capture and integrate **population genomic information** to inform health care decisions, research and policies.

Action 22: Through consultation and engagement, develop **information resources** tailored to the general population and vulnerable groups, in the community on the implications and benefits of genomic data sharing to build community trust in the delivery of health care and for secondary purposes such as research.

Action 23: Build on existing work to develop a **national proof of concept for data sharing** across IT systems in different health care and research settings (such as pathology laboratories, hospitals, registries and research institutions).

Subsequent to the National Framework, recommendations were issued by the GHFM ICT subcommittee, and the National Approach to Genomic Information Management (NAGIM) Blueprint. The National Framework also informed the data management strategy of the Australian Genomics pilot data projects that ran from 2016 to 2021.

Australian Genomics

The national recommendations from the GHFM and NAGIM will be addressed in a final briefing that integrates information and learnings from the Australian Genomics capability reports, data pilot projects and international best practice.

The Australian Genomics Capabilities Reports (2020)

Domestic Report. Responses from 17 Australian organisations managing genomic data infrastructures, including university or medical research institutes, diagnostic testing laboratories, translational research centres or programs, and data service providers, both research and clinical.

International Report. Responses from 17 large-scale genomic initiatives internationally, representing North and South America, Europe, Africa and Australasia. Infrastructures included large-scale national precision medicine initiatives, research cohorts, service-based platforms for data storage and data analysis, and variant databases.

Broadly, these reports indicated:

- Transitioning to **federated infrastructures** is a key next theme for international initiatives;
- A high proportion of domestic (83%) and international (54%) infrastructures currently use either **cloud or hybrid** cloud/non-cloud infrastructures;
- Most of the international initiatives surveyed (81%) currently support **external data sharing,** while Australian infrastructures are not typically engaging in external data sharing (31%), citing governance challenges among key limitations;
- Most of the international initiatives surveyed are adopting one or more forms of **standardised terminologie**s, while domestically, few infrastructures surveyed were collecting or storing clinical information in standardised terms.

The Australian Genomics Data Program Pilot Work (2016-2021)

Piloted data infrastructure, tools, and policies

The Australian Genomics data program supported 18 clinical research studies, recruiting participants and collecting clinical information from 32 clinical sites across every state and territory in Australia, and genomic sequencing data from 6 laboratories in 4 states. This supported primary clinical activities, secondary research projects and broad data sharing for further ethically-approved projects.

This national data pilot program established data access and governance policies, agreements and processes, dynamic and standardised national research and clinical consent, clinical data tools and databases, and genomic data repositories. These include the **Australian Genomics research data repositories, clinical information sharing tools (Shariant, PanelApp), dynamic consent portal (CTRL) and the Acute Care genomics end-to-end data management cycle**. (*Further details provided in the Appendix)*

The current phase of the Australian Genomics data projects are on track for completion in July 2021 and the infrastructure, tools and findings are now able to inform development of a national prototype for genomic information management in Australia.

International Best Practice

Australian Genomics is also contributing to international standards and tool development, including those of the **Global Alliance for Genomics and Health** *(Appendix, A4)***.** These include standards and tools for data sharing, encryption, repository querying, and remote data analysis and will be crucial for informing a national genomics infrastructures strategy that is aligned to international best practice.

2. GENOMICS HEALTH FUTURES MISSION (GHFM) RECOMMENDATIONS

Summary

The Executive Advisory Committee to the Genomics Health Futures Mission (GHFM) developed its Operational Plan (2018, not publicly released)[1] with recommendations on data provided by its ICT subcommittee. These GHFM ICT recommendations emphasised the need for standards-based approaches, interoperability, international standards, and improved cybersecurity such as "model to the data" approaches. An ICT survey for the GHFM identified cloud capabilities as a key priority in the genomics landscape. The GHFM ICT recommendations further described three key users, (patients, clinicians, researchers) whose requirements will need to be met by the future ICT infrastructure.

Vision for Genomics in Australia

In considering ICT requirements, the GHFM Operational Plan notes its vision for genomic data sharing in Australia as requiring a full data steward role with:

- Provision of a **single access point** for use and reuse of (potentially federated) genomic and phenotypic data
- Connecting to other **domestic and international datasets**
- Establishment of **standard protocols for collecting, storing and accessing**; security, privacy, and consumer data rights in accordance with Australian law and research ethics obligations.

Interoperability

As with the NAGIM Blueprint, the GHFM Operational Plan emphasises that ICT for genomics needs to consider Australia's **National Digital Health Strategy and a national interoperability framework** – co-ordinated data services, seamless access and control by citizens, establishing national digital infrastructure to support digital health initiatives, such as MHR. The Operational Plan indicates that the NDHS will be part of the ecosystem for any future genomics capability and leveraging national infrastructure will need to be considered.

The Operational Plan suggests that the GHFM objectives will likely be achieved with nationally accessible data that allows **flexible access**. This could be through dedicated physical infrastructure or cloud.

Standards

The GHFM OP for ICT also identifies standards and regulatory compliance as important, for confidence and speed of adoption, public trust, data portability, national/international ecosystem interoperability, efficiency, and maximising impact. Designing compatibility with international standards, with full compliance with Australian legislations and regulations, compliance with international regulations, and interoperability with other systems and international collaboration.

The GHFM OP noted its commitment to national interoperability standards under the National Digital Health Strategy, and that future ICT solutions for the GHFM need to interface with interoperability standards and frameworks. Further, it expects future proposed ICT solutions to detail how they will be compliant or enable:

- Industry and **international standards for genomic data storage** – particularly those of Global Alliance for Genomics and Health (GA4GH)
- Current Australian and **international standards for clinical and medical terminology**, including SNOMED, ICD11, HL7, and Australian Medicines terminology.

[1] Used with permission from the Australian Medical Research Advisory Board chair

Key User Requirements

The GHFM Operational Plan for ICT identified three key user types whose needs and requirements will require consideration: **patients, clinicians and researchers**.

***Table 1**. Key national infrastructure user requirements*

Infrastructure Users	User Requirements*
Patients	• accurate and timely testing • clinical genomic data stored in accordance with regulatory requirements • research test data stored in accordance with research protocols • consent for data use • privacy, security, information about data usage
Clinicians	• clinical genomic testing information reported to aid decision making • data maintained and stored in accord with regulatory requirements • data reused only where consent obtained and clinically indicated
Researchers	• genotype and phenotype data that can be analysed, stored and shared • data accessible in an environment with sufficient compute that can support multiple analysis tools and techniques • genotype and phenotype data generated through multiple research projects (including clinical trials) and clinical activity made available for research • research can be executed collaboratively with multiple national and international researchers • data captured from multiple sources (including from new studies or samples, international and national research, clinical datasets) • data quality, interoperability, and standards compliance for optimum reuse • outcome: perform research using data across a range of applications – rare disease, oncology, genomic variation etc

*summarised selected extractions

Cybersecurity

The Operational Plan indicates it expects emerging technologies for data security to be explored (including homomorphic encryption, secure multiparty computation, zero knowledge proofs, and secure enclaves) which enable **"sharing without access**" and "**model to data"** paradigms, which are changing the traditional ways in managing security and privacy risks for biomedical data sharing.

Priorities in the Australian genomics landscape

A survey conducted for the GHFM and reported in the Operational Plan, from current clinical and research genomics stakeholders in Australia, identified potential gaps and underlying priorities for change in genomics technology.

The GHFM survey reports that most of the current ICT environments for genomics were on premises, with less than 30% using cloud-based storage. However, the majority of respondents identified that there scope for change and improvement in general within ICT for genomics. A similar survey (conducted by Australian Genomics in 2020) found that 83% of the responding Australian genomic data infrastructures have either cloud or hybrid cloud/non-cloud infrastructure. This could reflect an increased uptake of cloud services, from 2018 to 2020, or a different set of responders in the more recent survey.

Among the top priority areas identified by respondents of the GHFM ICT survey were access to cloud-based computing, cloud based storage, data sharing capacity, analytics and data curation (Table 2).

Table 2. *GHFM Operational Plan ICT Priority Areas from survey respondents*

Priority	Area	Definition
1	Storage, access and sharing capacity	Storage availability, technology to support data sharing, and to control access to shared data
2	Analysis and informatics	Availability of specific applications and capabilities to execute analysis on genomic data
3	Data curation	Applications and standard data-sets to support data integration and management
4	Cloud computing resources	Cloud compute capacity for genomics data
5	ICT networking and data transfer	Enhanced WAN/LAN capacity
6	Cloud storage	Cloud storage capacity for genomics data
7	Sequencing devices	Capacity/additional sequencing devices
8	Local storage	Local/on premises storage for genomics data
9	On premises computing resources	Local/on premises computing (e.g. HPC) for genomics data

3. NATIONAL APPROACH TO GENOMICS INFORMATION MANAGEMENT (NAGIM)

Summary

Established following national consultations, the Blueprint for a National Approach to Genomics Information Management (NAGIM; Oct 2020) delivered a proposal for an **integrated genomics ecosystem** across healthcare and research organisations, to support genomic research and genomic medicine, simultaneously.

From evaluating the current jurisdictional, operational and technical landscape in Australia, the NAGIM Blueprint concluded that a **standards-based approach, using a federated or hybrid model,** is likely to be the most appropriate strategy for a national approach to genomic information management in Australia.

The NAGIM Blueprint also reported an increasing preference, in clinical and research settings, for progressing **cloud-based solutions**, and identified the following factors as crucial for a future national genomics infrastructure:

- **Interoperability** across systems, with standards-compliant application programming interfaces (APIs) to support seamless access to information, within and across healthcare services and research organisations;
- Adoption of **international standards** and alignment with large-scale genomic initiatives, to enable international data sharing.

High-level Requirements

The NAGIM Blueprint identified several high-level requirements in its "Considerations for designing a framework" (*Chapter 4*).

These include:

1. **Interoperability** of systems between i) research and healthcare systems nationally, and ii) between Australian and international systems;
2. Federated approaches to **genomic medicine** data repositories using standards-based development to support the different jurisdictions that have different data management solutions, legislation and regulation;
3. **Access** to genomic, clinical and phenotype data from healthcare, for Australian[2] researchers
4. Nationally coordinated approach to **genomic research** capabilities, with multiple standards-based repositories, to address the inability for all research repositories to be able to be combined;
5. Address specific needs of **Aboriginal and Torres Strait Islander** communities and initiatives;
6. Improvements in **privacy, consent and security**

[2] '*Australian*' specified in this NAGIM high-level requirements section, however, *international* data sharing is consistently emphasised throughout the Blueprint.

Proposed Architecture for Genomic Research and Genomic Medicine in Australia (*Chapter 5)*

The Blueprint outlines a proposed logical architecture for a future national ecosystem supporting: i) genomic medicine, ii) genomic research and iii) integrations between the two.

The current architecture in Australia

The current architecture is described in the Blueprint for:

- Australian genomic medicine systems: including information systems for operating health services (Electronic Health Record (EHR) and laboratory information management systems (LIMS), clinical sequencing and bioinformatic analysis capabilities, and for exchanging clinical genomics knowledge;
- Australian genomic research systems: including information systems for managing research data (Research Data Management Systems (RDMS), Data Access Management systems (DAMS), research sequencing and bioinformatic analysis capabilities, and for exchanging genomic information;
- Both frameworks: the Blueprint notes there is a combination of local system-focused repositories (EHRs and LIMS, for genomic medicine; RDMS and DAMS, for genomic research), data staging systems on local and cloud infrastructure, and core genomic data stores in a variety of databases, repositories and formats.

Considering these architectures and based on their consultations, the NAGIM project identified:

- **Critical to national adoption**, the proposed future architecture will require standards-based interfaces (APIs) to other systems; and data orchestration between repositories and systems;
- Importance of data flow connections to external data repositories, that are both providers of critical information for genomic activities and recipients of data from our health service and research organisations;
- Increasing preference for cloud-based genomic data stores, amongst clinicians and researchers in Australia and internationally, who were consulted as part of the NAGIM project.

The NAGIM Blueprint also identified 'patterns of interactions' that would support a mature national genomics ecosystem:

- Point-to-point requests for data (standard data custodian process);
- Synchronisation of data sets (agreed standards for storage, transport, access; federated national genome archives);
- Remote querying (data requestor defines query; data provider executes and returns results; supports FAIR data);
- Federated queries (remote query for multiple datasets);
- Self-describing repositories (capabilities with an interface that allows high-level queries by others).

Increasing Genomic Data Interoperability

The Blueprint noted that the current architecture in Australia is largely bespoke 'point-to-point' data flow, with inconsistent standards use. Potential next stages are described that would ensure progress to a mature genomic data ecosystem, by incrementally establishing:

i) standards-based genomic data processes: adding standards-based processes to existing systems, across organisations; developing consistency to allow aligned solutions to emerge.
ii) standards-based integration: nationally-agreed standards for federated queries across multiple data providers, nationally supported identity management, national approach to consent.
iii) standards-based interoperability-enabled systems: the 'highest level of maturity', progressing from integration to an interoperability-enabled standards-based system.

The NAGIM Blueprint identifies the need for **flexibility** in working towards a mature genomic data ecosystem, noting that not all jurisdictions or research organisations will have the same capabilities or priorities, and a mix of technologies and capabilities will likely remain over the short to mid-term, as those organisations with capability and capacity lead development. Standards-based APIs will allow us to add these organisations to the ecosystem as their capabilities improve.

Determining the correct approach

The NAGIM Blueprint notes that when considering the architectural approach for Australia, not everything can, or should, be centralised. But, where appropriate, the Blueprint recommends a central authority establish the aspects of the national genomic information network that need centralising.

The NAGIM Blueprint also acknowledges that growing capabilities of cloud technologies are changing perspectives about what 'national' or 'centralised' systems should look like, and urges the future governance group and national participants to be open to new approaches to federation.

The Blueprint encourages consideration of the following, for a federated interoperable system:

- Jurisdictional differences in data management requirements, policy, funding, priorities;
- Central management of some essential functions through a federated model, to lower barriers to uptake by less mature or 'resource poor'[3] jurisdictions;
- Federation would enable scale and cost levelling;
- Centralised, distributed, large or small compute capabilities - either part of or outside the system - are all compatible with a federated model.

The Blueprint argues that **a hybrid solution employing advanced technology and best practice approaches could leverage both centralised and highly decentralised models.**

[3] As termed in NAGIM Blueprint

Genomic Data Governance Framework (*Chapter 6*)
The NAGIM Blueprint asserts that a national approach to genomic information management will require a strong governance framework that can be applied with consistency, with an operational model that can support the **diverse requirements of both the clinical and research sectors**.

The Blueprint notes a future national approach will need to address **significant complexities around data sovereignty**, for national and international data sharing and data storage, when considering jurisdictional data, and data from Aboriginal and Torres Strait Islander peoples; as well as complexities with data ownership, permissions, IP and rights, that are **unique to human genomic data** generated from healthcare and research.

The Blueprint defines 5 governance operational models (Decentralised, Network, Centralised, Federated, Hybrid)[4] and concludes **that the complexities of the Australian healthcare and research sectors suggest that a federated or hybrid model are most appropriate.**

Standards and interoperability (*Chapter 7*)
The NAGIM Blueprint defines interoperability as "..the ability of different information systems, devices and applications to access, exchange, integrate and cooperatively use data in a coordinated manner, within and across organisation, regional and national boundaries to provide timely and seamless portability of information and optimise the health of individuals and populations globally".

The Blueprint emphasises the importance of interoperability for Australian genomic data systems, further noting interoperability as one of the seven strategic priorities for the Australian health sector, by the National Digital Health Strategy, and the progression of the Australian Digital Health Agency's interoperability program and it's expected potential impact for genomic data standardisation.

The NAGIM Blueprint has concluded that to support interoperability in the health sector - **architectures, application interfaces (APIs) and standards are required** to enable data to be accessed and shared appropriately and securely, within all applicable settings and with the relevant stakeholders. These include HL7 standards for integration and exchange of electronic health information, Observational Health Data Sciences and Informatics (OHDSI), and metadata standards.

In particular, the Blueprint specifically highlights the international standards and tools of **GA4GH**, as core standards considered important to achieving interoperability. These include APIs such as for authenticating researchers, issuing approvals for data access, discovering and querying data, extracting data, and running workflows in different computing environments.

[4] NAGIM Definitions: **Decentralised operating model:** Data management responsibilities are distributed across multiple functions with no single owner. This provides the simplest structure, but governance and decision-making are more difficult. **Network operating model:** More formalised than a decentralised model, a network model introduces defined relationships and accountabilities. The difficulty is in maintaining the defined relationships and expectations. **Centralised operating model:** The most formal and mature model but requires substantial organisational change to achieve and the separation of data management from the operational 'coal face' can lead to a lack of focus on the strategic outcomes. **Federated operating model:** A federated model provides a centralised strategy with decentralised execution. A centralised coordination process is required, and this can introduce complexity through the need to balance operational independence against the needs of the whole. **Hybrid operating model:** In a hybrid model, data management is coordinated through a centre of excellence working with more decentralised operating areas, supported with more tactical working groups.

NAGIM Proposed Roadmap for Implementing the Blueprint

The NAGIM Blueprint outlines three phases ('Horizons'), to achieving a future national approach to genomic information management, as an ecosystem that would support both genomic medicine and genomic research (*pp. 59-63*). The Horizons, or phases are defined around core themes of Governance, Medical Genomics, Genomic Research, and *Infrastructure.*

The proposed NAGIM '*Infrastructure*' activities are extracted below.

***Table 3**. Extract from the NAGIM Blueprint 'roadmap for implementation' - Infrastructure*

NAGIM Horizon: Infrastructure	Infrastructural activities required to support the delivery of the medical genomics and genomic research
Horizon 1: Leverage and Plan	• Undertake **implementation studies** of the leading genomics systems in use across Australia to map against the logical model and establish baseline and learnings for future implementations. Such studies should examine existing research partnerships (ideally cross-jurisdictional) as well as existing and emerging jurisdictional solutions. A study of clinical/research partnerships would be beneficial. • Develop a **standards-based, interoperable approach to cloud adoption** to support storage and retrieval of genomic data in both medical and research domains. • Work with international groups (such as GA4GH) to agree **standards for self-describing repositories** that can identify their content and capabilities. • **Trial the establishment of a shared, cloud-based repository for genomic research** data across at least two jurisdictions to establish baseline and learnings to inform future implementations. • Establish **standards for federated query** across genomic data repositories. • Work with international groups to agree **standards for international research data sharing**
Horizon 2: Build on Foundations	• Adopt **national interoperability for cloud** infrastructure • **Trial federated query standards** across repositories to support a national genomic information network operational framework. • Expand a **shared, cloud-based repository for genomic research** data across at least two jurisdictions to establish baseline and learnings to inform future implementations. • Work with international groups to operationalise **international research data sharing**
Horizon 3: Transition and Operate	• Continue roll out and standardisation of **national interoperability for cloud** infrastructure. • Leverage **federated query standards** across repositories to support a national genomic information network operational framework. • Expand a **shared, cloud-based repository for genomic research** data across all jurisdictions to complete the national genomic information network operational framework. • Monitor and leverage **international research data sharing**.

The NAGIM Horizon activities for *Infrastructure* show a clear focus on:

i) Supporting federated querying
ii) Standards-based processes
iii) Interoperability across systems
iv) Alignment and interoperability with international data sharing initiatives
v) Cloud-based solutions
vi) Basing decisions on national infrastructure on working pilot repositories

Further activities, under the NAGIM proposed roadmap, for additional areas of Governance, Medical Genomics and Genomics Research are listed below. These highlight the importance of progressing national data governance, clinical data sharing, phenotype capture, national consent mechanisms and national data sharing agreements, as part of an operational national genomics ecosystem.

***Table 4**. Extract from the NAGIM Blueprint 'roadmap for implementation'*

Area	NAGIM Horizon 1 Proposed Roadmap Activities
Governance	• Establish or leverage a **national governance group** comprising clinicians, researchers, policy makers, funders, consumers and Aboriginal and Torres Strait Islander people to coordinate activities over the three horizons. The governance group should be informed by focused working parties and be inclusive of industry players acting in partnerships. • Developing a robust **data governance framework** that ensures that relevant protections are in place to protect the genomic information of individuals and groups should be a priority first action of the national governance group. • Consideration should be given to whether a national or jurisdictional **Data Custodian/Steward** is required to provide oversight of how data is managed, accessed and shared. • **Confirm or amend the roadmap** elements of this national approach to genomic information management. • **Identify an organisation/group** with the capabilities to operate a national genomic information network or build a federated structure for all jurisdictions to participate equally. • **Establish a national consumer engagement group** to ensure that genomic data activities meet community expectations for addressing risks and benefits. This group should include representation of Aboriginal and Torres Strait Islander people and other groups with specific needs (such as Culturally and Linguistically Diverse (CALD) communities). • Agree/adopt national standards for **genomic data storage formats, genomic data exchange methods, computable consent and cybersecurity policies, guides and standards** informed by existing national and international standards. • Agree on an **interoperability capability model** that allows for organisational self-assessment in support of planning and funding decisions. • Agree on **national data retention policies** for all classes of genomic data that consider both clinical, diagnostic service and research requirements.
Medical Genomics	• Promote **collaboration and share learnings** between the jurisdictions undertaking activities, those planning such activities and other interested parties. • Establish a cross-jurisdictional working group to **standardise access to familial and pedigree data** for clinical purposes. • Establish **national agreements for genomic data sharing for clinical purposes**, leveraging existing clinical data sharing agreements working with private and public providers. • Establish an **agreed approach to capture or mapping of phenotype data** within clinical systems to support genomic diagnosis, predictions and research. • Support **ongoing operation and expansion of variant curation repositories and tools** (e.g. Shariant) to support genomic medicine.
Genomics Research	• Establish **national agreements for genomic data sharing for research**, leveraging existing data sharing agreements. • Establish a **national research consent mechanism** for genomic data utilising strong credentialing for participants with dynamic approaches to ongoing engagement • Continue **trials of research data sharing** with leading clinical groups, leveraging existing genomic programs, to establish baselines and learnings for later implementations. • Establish **national arrangements** to consider Australia's access to and use of global genomics data assets, our dependencies and role on the world stage.

4. PROPOSED APPROACH

These materials suggest consensus on several key points relevant to implementation strategy:

- It is unlikely that a single fully centralised system for genomic data storage and analysis can serve the needs of both research and clinical users nationally;
- A data model based on **federation across multiple repositories** using standardised interfaces will likely be necessary, especially for clinical data; however, a **single unified repository for research data** may be tractable and would be highly preferable;
- There is a growing preference for **cloud-based models** for international genomic data solutions, rather than models based on centralised HPC;
- **Decisions on potential platforms should be made based on the evaluation of working prototypes managing data across multiple jurisdictions.**

Australian Genomics will develop **a series of phased recommendations** that will be based on:

- Infrastructure prototyping of components that meet NAGIM recommendations;
- Outcomes from clinical and research stakeholder engagement;
- International evaluation and best practice

National infrastructure stakeholders will be invited to participate in prototype construction, definition of technical specifications, and infrastructure evaluation. This will be designed as an open process to facilitate communication and collaboration between groups involved in developing prototypes, with the goal of **identifying the best combination of components that can serve as the basis for long-term national infrastructure.**

Clinical and research stakeholders will be engaged in parallel to identify and progress additional core elements of the NAGIM ecosystem, including priorities for clinical data and data governance.

Strategy

Approach	Components
Stakeholder Engagement	● Engage key stakeholders and infrastructure implementers ● Define the intended NAGIM implementation ecosystem ● Identify what can be centralised ● Determine priorities for clinical and governance
Infrastructure Prototyping and Evaluation	● Define required components ● Define critical specifications and evaluation process ● Compare systems and technologies: ○ Computational infrastructure providers ○ Data warehousing platforms ○ User authentication and access control systems ○ Scalable platforms for data processing and analysis ○ User interfaces for core analyses ○ Other components identified by stakeholders ● Assess the respective role of HPC and cloud computing platforms ● Assess capability to support federation across states
Phased Recommendations for National Implementation	● Initial recommendations strategy ● Prototype-informed recommendations ● Research and clinical approaches

Delivery of Recommendations

Phase	Outcome	Components
1	Summary briefing (this document)	● NAGIM implementation elements mapped ● Proposed strategy for recommendations
2	Initial recommendations	● Research prototyping specifications defined ● National research and clinical stakeholder engagement
3	Prototype-informed recommendations	● Functional research prototypes built and evaluated ● Additional priorities identified for clinical and governance
4	Final recommendations for implementation	● Framework for national scale up ● Evaluation of clinical solutions ● International evaluations and best practice

Timeline

Phase	Outcome	Activities	Timeframe
1	Summary briefing and proposed approach (this document)	● NAGIM implementation elements mapped ● Proposed approach developed ● Review by Australian Genomics Data Network and National Steering Committee	2021 June
2	Technical specifications	● International advisory committee (IAC) identified ● Core prototype specifications and evaluation framework defined ● Stakeholder feedback on the prototype specifications and evaluation framework received	2021 July
3	Prototyping	● Open call for participation in prototyping (including circulation of the specifications and evaluation framework) ● Prototyping teams assemble, register their participation, and begin development work ● Australian Genomics coordinates monthly progress meetings with the prototyping teams	2021 December
4	Evaluation	● The IAC conducts a technical evaluation of the prototypes ● The prototyping team conduct a technical peer review evaluation of the prototypes ● Suitability of pilots/components for clinical sequencing providers and clinical warehouses assessed ● Additional components of the NAGIM Roadmap evaluated for Governance, Medical Genomics and Genomics Research activities ● Australian Genomics consolidates all evaluations to arrive at a preferred model for implementation	2022 January
5	Initial Report	● Australian Genomics develops an initial report outlining high-level recommendations for approaches to national implementation	2022 February
6	Comprehensive Recommendations	● Comprehensive national recommendations for research and clinical solutions and evaluations, international best practice and national scale-up ● Draft report is circulated to the IAC, prototyping teams, and clinical and research stakeholders for consultation ● Australian Genomics delivers the final comprehensive report to Government	2022 May
7	Implementation	● *TBD*	*TBD*

5. APPENDICES

A1: Australian Genomics Data Pilot Program 2016-2021

This Program has been developing data standards and processes to capture and use clinical and genomic data from across Australian clinics and laboratories, to support the activities of Australian Genomics clinical Flagship studies.

Key products

Australian Genomics Database

The clinical, demographic and survey data of all our participants is stored and managed in our database. Study data are collected and managed using REDCap electronic data capture tools hosted at Murdoch Children's Research Institute. Our database integrates information collected from over 32 clinical recruitment sites across Australia.

Genomic Data Repository (GDR)

The Australian Genomics GDR is a cloud-based storage system, hosted at The University of Melbourne. The GDR stores genomic data from the Australian Genomics clinical Flagship studies and balances security, data integrity and access for researchers.

Data Access Agreements and Policies

Our participants provided consent to both national and international data sharing, for the benefit of healthcare. From this, a Data Governance Framework has been defined, which is guiding access principles and policies, agreements and terms of research access. Clinicians and researchers can request access to Australian Genomics datasets, which will be granted according to the level of data sensitivity; each participants' consent; and the researcher's Human Research Ethics Committee.

CTRL dynamic consent & DUO standard

Individualised consent is being sought through a trial implementation of our CTRL dynamic consent platform. Participants' consent choices about future research use of their data are being captured and integrated into the REDCap database, using a Global Alliance for Genomics and Health data use ontology (DUO) standard. This means each participant's consent preferences are captured in real-time and applied to cohorts for data sharing. Standardising data use conditions in this way eliminates ambiguity in data use permissions and streamlines complex data sharing processes.

Standardised Computer-Readable Clinical Data Tools

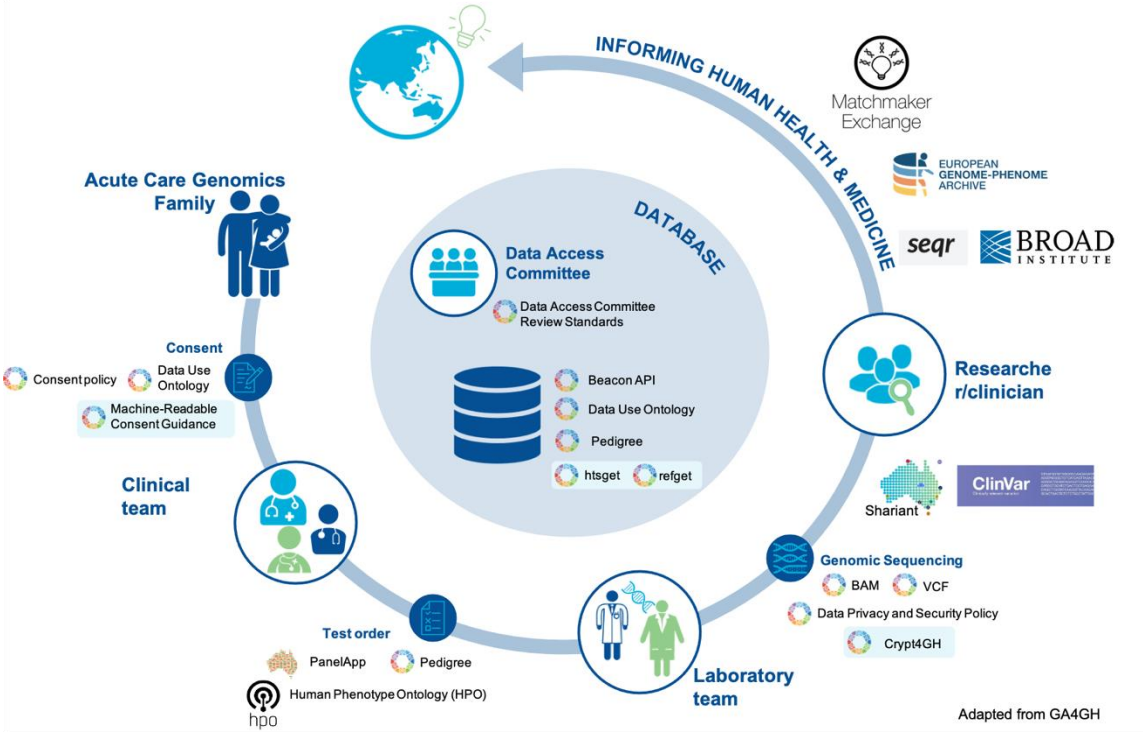
Our tools enable detailed clinical information from Flagship patients to be collected and translated into standardised codes (phenotype ontologies: HPO/SNOMED) and clinical data formats that are compatible with eHealth systems and diagnostic tools. This allows computerised search, discovery and coding, to power genomic research and clinical activities.

'Variant Atlas' Genotype-Phenotype Platform Variant Atlas is an interactive genotype-phenotype data platform, developed by the Garvan Institute. With this tool, researchers and clinicians can visualise aggregated Flagship data, query variants across the Flagships, and filter by key clinical features to explore and describe the different patient populations.

'Shariant' Platform for Interpreting Variants Shariant is a centralised platform for Australian pathology laboratories to share interpreted variant classifications and detailed evidence in real-time. Labs are automatically notified of classification differences with other groups and can connect to international databases.

A2: Australian Genomics Exemplar - Acute Care End-to-End Data Management Tools

The Acute Care Genomics Flagship provides ultra-rapid genomic testing to critically ill infants and children with rare conditions from around Australia. The Flagship data capture and management systems **use the Australian Genomics data infrastructure,** incorporating **Global Alliance for Genomics and Health (GA4GH) standards** where possible and **have been optimised for speed** as well as for **data sharing** to facilitate secondary data use.



<u>The Acute Care 'end-to-end' process using Australian Genomics data tools:</u>
Phenotypic and clinical data capture uses a **streamlined electronic genomic test order** which incorporates the Human Phenotype Ontology (HPO), the Human Ancestry Ontology, and an **electronic pedigree drawing tool**, aligned with the GA4GH Pedigree standard. The order links to **PanelApp** Australia, an open repository of curated virtual gene panels, promoting evidence-based diagnostic analysis. **Consent is captured electronically**, and consent clauses are mapped to the GA4GH Data Use Ontology.

These electronic clinical workflows are not only efficient, but also enable the use of telehealth, which has been important in facilitating study enrolment from sites without clinical geneticists, and during periods of COVID-19 restrictions. **Data is collected in machine-readable formats** where possible, enabling direct integration with laboratory systems.

Once diagnostic reporting is complete, variants are deposited into **Shariant**, and from there will be submitted to international repositories such as ClinVar, contributing to international efforts in promoting evidence-based practice. Novel gene candidates are deposited to the MatchMaker Exchange, and this has already resulted in at least three novel gene discoveries to date.

Genomic data is deposited into the **Genomic Data Repository**, and has also been deposited into the European Genome-Phenome Archive, enabling further research use. Genomic and phenotypic data

from the study is also currently being deposited into *seqr*, which will enable further research reanalysis nationally.

A3: Australian Genomics Data Pilot Program Details

Australian Genomics established a range of pilot projects, for the collection, storage, access and re-use of genomic and clinical phenotype data. Australian Genomics adopted a hybrid method in its data program, with federated collection of clinical and research genomic data into centralised storage and management.

Genomic Data Repository

Australian Genomics piloted a genomic data repository (GDR), using a cloud-based system supported by AWS. A number of platforms were evaluated as part of the pilot work, including Arvados, Human Cell Atlas, Terra and Gen3. The GDR has been ingesting genomic data from six separate laboratories across five states, for 12 clinical research studies with processes to validate and aggregate the data, and prepare the cohort datasets into an appropriate format for data sharing and long-term archiving.

Standardisation of data intake has proven to be a key challenge but is vital for data harmonisation and sharing efforts that are now driving the development of Gen3 and Terra as implementation pilots for the cancer and rare disease communities, respectively. Australian Genomics is paying particular attention to the interoperability of pilot implementations, reflecting future requirements to interact with a diversity of international partner organisations and architectures.

Data Sharing Tools

Australian Genomics has been exploring data sharing tools, and data access models such as 'open-access' (public), 'registered access' (streamlined), and 'controlled-access' (committee-approved and ethics-reviewed). The tools specifically developed by Australian Genomics, under these data access models, include:

- **Shariant (clinical):** an online platform for diagnostic laboratories to share curated variant classifications.
- **PanelApp (clinical):** a publicly available knowledge base for virtual gene panels related to human disorders to be created, stored and queried. It allows genes to be reviewed by experts, to support standardisation of gene panels, and consensus on gene-disease associations.
- **Variant Atlas (research):** a Genotype-Phenotype data platform that allows researchers to visualise aggregated Flagship variant data, and filter by key phenotypic features.
- **GDR and long-term archives (research):** genomic data warehousing for researchers to access genomic datasets for ethically approved research.

The AG data projects confirmed the feasibility of adopting different data access models to streamline low-risk data sharing, including databases for clinical variant classifications (Shariant), and disease-specific variant frequency data (Variant Atlas), for clinical and research purposes; as well as feasibility of a central data custodian model for 'controlled access' management of national genomic studies.

Data Governance Resources

Australian Genomics developed a suite of data access and governance tools and processes, including:

- **Agreements:** Laboratory Data Transfer, Data Sharing for Secondary Data Use
- **Policies:** for Secondary Data Use, Data Breaches, and Secondary Research Findings
- **Data management resources:** Data Management Plan, Online Data Request Forms, Data Governance instruments
- **Consent resources and tools:** Online dynamic consent platform (CTRL), a national clinical consent form, streamlined consent for time-critical scenarios (Acute Care), e-consent processes, and a genomics research consent template.

The Australian Genomics data projects indicate that data curation and governance processes, to support genomic and clinical data sharing, are highly resource-intensive and reliant on extensive manual processes, suggesting significant challenges for national scalability.

Specifically, the AG national data projects indicate challenges can be expected around: constrained resources of data submitters; complexities of data and project ownership (cohort authors, data contributors, sequencing laboratories) the execution of agreements supporting data use (data transfer agreements, multi-institute agreements, data sharing and secondary use agreements); and collating complete clinical and genomic datasets, with the required associated metadata into an appropriate format for broad on-sharing; **central data coordination was critical to driving these activities.** Collectively, the AG national data pilot projects confirm the need to establish streamlined and scalable processes to operate data access and governance processes at a national level for genomic data.

Standardised clinical data tooling and resources

To achieve interoperability of clinical data that is relevant to genomics, both the **clinical terminology** and the **information model** need to be standardised.

Information model: the structure of the information captured for individual patients.

- AG selected the HL7 FHIR standard that supports interoperability. AG produced a FHIR implementation guide (IG) to represent patient phenotype and ensure consistent querying across multiple data sets.

Clinical terminology: shared knowledge about health that can be used to populate certain fields in the information model

- AG developed software that enables different clinical terminology systems to be used in FHIR (FHIR-OWL software). This includes terminologies such as SNOMED for healthcare systems, HPO for rare disease, MONDO for variant curation. This allows flexibility in supporting the different terminology systems that are in use across organisations, a requirement identified in the AG pilot projects and international initiatives.

Majority of the clinical information received in the AG data pilot projects, from the national clinical sites, were not available in standardised clinical terms. Retrospective mapping was required to convert clinical information, captured in REDCap study database, into standardised terms. The significant complexities of retrospective mapping large-scale datasets highlight the critical need to implement clinical terminology systems at the initial point of data collection, for processes to be sustainable and scalable to a national approach.

To address scalability of standardisation of clinical data, several open source tools were developed:

- **A terminology app for REDCap:** enables initial data collection in REDCap study databases to be captured in clinical terminologies using autocomplete functions. (REDCap Ontoserver Plugin).
- **Software for standardised data formats:** enables information in REDCap to be transformed into international FHIR format, using a domain-specific rules language. (REDMatch)

Standardised clinical data capture templates for use in REDCap, are now being developed by AG, together with the corresponding mapping rules, to facilitate efficient establishment of a full, standardised clinical data collection system.

The Australian Genomics clinical data pilot projects indicate that for standardisation and interoperability to be effectively integrated in Australian genomic infrastructures, **a set of standards should be selected and tested in a next-stage pilot**. This will require stakeholders to be actively engaged in the development of these standards, to identify and address gaps.

A4: International Best Practice - The Global Alliance for Genomics and Health Ecosystem