Queensland Genomics

Queensland Government

# BLUEPRINT

## FOR A NATIONAL APPROACH TO GENOMIC INFORMATION MANAGEMENT

—

OCTOBER 2020

**Acknowledgement**

Queensland Health proudly acknowledges Aboriginal and Torres Strait Islander people as the First Peoples and Traditional Custodians of the land and waters on which we meet, live, learn and work. We pay our respects to the Ancestors of this country, Elders, knowledge holders and leaders – past, present and emerging. We give our gratitude to the many people who generously contributed to the development of this Blueprint. We are particularly grateful to the Aboriginal and Torres Strait Islander people who contributed their experience and cultural authority.

We also acknowledge those who participated in workshops and meetings, provided review feedback or otherwise gave of their time to support the development of this work – it would not have been possible without their expertise and support.

This project was funded by the Australian Health Ministers' Advisory Council, and this report would not be possible without the leadership, support and guidance of the Commonwealth Department of Health's Project Reference Group for Human Genomics.

**Blueprint for a National Approach to Genomic Information Management**

This document is maintained in electronic form and is uncontrolled in printed form. It is the responsibility of the user to verify that this copy is the latest revision.

This Blueprint was prepared under the auspices of the Australian Health Ministers' Advisory Council and prepared by Queensland Health.

Copies can be obtained from Queensland Health, Office of the Chief Clinical Information Officer at DDGandCMO_PreventionDivision@health.qld.gov.au Enquiries about the content of the Blueprint should be directed to the Queensland Health Chief Clinical Information Officer at the above address.

# Table of Contents

# Executive summary

The *National Health Genomics Policy Framework* (NHGPF) [1] established five strategic priorities to support the integration of genomics into health care for Australians:

- **Person-centred approach:** Delivering high quality care for people through a person-centred approach to integrating genomics into healthcare
- **Workforce:** Building a skilled workforce literate in genomics
- **Financing:** Ensuring sustainable and strategic investment in cost-effective genomics
- **Services:** Maximising quality, safety and clinical utility of genomics in health care
- **Data:** Responsible collection, storage, use and management of genomic data

Each of these priorities are complex areas in their own right. However, addressing the subject of ‘data’ (or information in the broader sense) can be challenging. In one sense, it can be a straightforward discussion relating to the nature and structure of data to be collected, stored and used. But the importance of health data, and in particular genomic data, means that issues of ethics, privacy, confidentiality, security and more need to be overlaid on these simpler discussions. Moreover, the nature of genomic data itself is rapidly evolving, and so even the simpler discussions of content and structure are changing. Our notions of value are shifting to acknowledge the important role data plays beyond its collection at the point of care to its subsequent ethical and privacy-sensitive use helping other patients and populations (by definition a secondary use).

The work covered in this document applies a contemporary architectural approach, building a bridge between strategy and policy positions to the decisions and the choices solution implementers make over time. These decisions cover both the technologies applied to match requirements, as well as the mechanisms required to consistently describe the moving parts of discreet solutions. It must also consider how they interact within the system (integration) and within a broader eco-system of discreet and interoperable systems, with essential national infrastructure supporting data sharing.

The rapid increase of applicability of genomics medicine to clinical care, prognostics and prevention is well acknowledged, as is the seismic shift in the origin of new genomic sequencing of humans moving from the research context to the healthcare delivery context. The work undertaken to develop this Blueprint fundamentally builds on the concept of a learning health system. One “in which science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience” [10]. As such the scope of the work, whilst founded in the public acute health setting, has sought to both acknowledge and outline advances made in the management of genomics information in the research setting, as well as the important role of translational research in advancing new knowledge into clinical practice and policy.

Effective, empowering data governance and complete lifecycle information management are critical building blocks to guide implementation and advances in our approaches to manage genomics information. Most of the international frameworks available for data governance (and specifically data sharing) are focused on research uses and less on clinical reuse (which should not to be confused with clinical research). The needs of researchers, clinicians, policy makers and individuals may not align and must be balanced. Therefore, this work sets out to encompass these issues, considering matters of legislation and regulation, of ethics, privacy and security and hence consent and consumer choice. Each of these topics warrants detailed investigation, and for many, they connect with their own priority area under the NHGPF and works undertaken by state, territory and Commonwealth health agencies.

This work however is about 'data' and while the issues listed provide context for the data and are important, the *Blueprint for a National Approach to Genomic Information Management* (the *NAGIM Blueprint*) attempts to provide a semantic expression for each of these matters as associated with the data. In this light, the *NAGIM Blueprint* does not 'solve' the consent issue, for example. It does however outline approaches and the essential requirement to manage consent as it relates to a set of data about someone, managed in a repository alongside other information about other people, where the notion of sharing that information, its provenance and agreement on its use is essentially connected to the data.

This *NAGIM Blueprint* attempts to address these complexities and provide a framework for implementers that recognises the ongoing evolution in the field. To achieve this, the *NAGIM Blueprint* adopts a principles-led approach that defines six broad domains of interest:

- **Consumers and communities**: This domain explores attitudes and approaches needed to gain and maintain the trust of the broader community supporting their meaningful participation and involvement, and hence shared benefit from genomics.
- **Aboriginal and Torres Strait Islander peoples**: This domain addresses the specific needs of Aboriginal and Torres Strait Islander communities to ensure that genomics benefits these communities without repeating mistakes of the past.
- **Genomic research**: This domain covers the needs and responsibilities of the research community as they relate to genomic discovery, as well as the management of sharing information.
- **Translational genomics**: This domain explores the translation of genomic discovery to clinical care to advance our understanding of the cycle of research into practice, and practice informing research. It is a critical area bridging the interests of healthcare delivery and research.
- **Genomic medicine**: This domain area covers the ongoing 'mainstreaming' of genomics in clinical care and the significant impact genomics will have on the way healthcare can be delivered.
- **Data management**: This domain covers general principles required to ensure that data is managed appropriately, with effective governance and applicable standards across all domains of interest.

The principles' purpose is to provide practical guidance on considerations for system implementers. The implications derived and associated with principles are deliberately non-prescriptive and purposefully not specific to individual implementations and technology standards. They provide a set of 'guardrails' within which implementers can operate and evolve their respective systems for managing genomic information. Furthermore, the principles are contextualised by reference to other national and international frameworks where possible.

Based on these principles, the types of data under management and the additional factors to consider, a logical architecture is proposed that describes the types of functionalities and data flows that need to be contemplated in both clinical, translational and research settings. The models provide a common vernacular to describe systems and to allow qualitative and quantitative comparison of implemented solutions. For example, how many genomes, the conditions in which the data was generated and assessed, related clinical impact, and how that information might be appropriately shared.

Using these models, a roadmap is proposed that outlines how the current state environments in Australia may be transitioned over time to a fully interoperable ecosystem that supports genomic information management in a range of settings. Indicative activities are discussed that describe the incremental steps needed to move towards a learning healthcare system. This document does not seek to direct, through policy or funding, the evolution of an ecosystem in which genomics information can be shared appropriately. Rather, the *NAGIM Blueprint* simply acknowledge that the desire and will to share for collective value is present, and hence an ecosystem of genomic data repositories will emerge and this will occur with greater certainty for outcomes in terms of value, privacy and context with planning – and a bridge between strategy and implementation.

This principles-based approach also acknowledges that genomics is a rapidly evolving discipline. Technologies today may be replaced by new technologies in the coming years. Even the nature of the data that is produced can and will change over time. When discussing 'genomic data' or 'genomic information', it is therefore important to have a common language or set of shared definitions to describe this data. The *NAGIM Blueprint* serves to establish a shared methodology to describing what we mean by genomic information and management approaches. Commonly, this is done using a tiered and related set of definitions referred to as a 'classification framework'. To avoid confusion with the concept of variant classification, the *NAGIM Blueprint* defines a 'genomic data categorisation framework' that proposes a structure for grouping similar data types under a set of defined descriptors. This categorisation framework therefore supports the process of defining characteristics such as data retention periods to similar types of data.

Any approach to genomic data must recognise that while still an evolving discipline, genomics is not a 'green field' environment, and the approach to be taken must consider existing influences. The *NAGIM Blueprint* therefore explores factors that must be considered, including:

- The similarities and differences between research and clinical practice and the nature of the bioinformatics analysis systems used in both areas. While there are many similarities, especially at the technology level, application of these technologies is influenced by regulation and accreditation, and the complex issue of consent and local and prescribed data retention policies.
- The impact the mainstreaming of genomics into clinical care may have on the availability of high quality genomic and clinical data to support research (subject to consent).
- The need for repositories of data to support federated requests for the managed data. This is leading to a demand for self-describing repositories that are interoperable, nationally and internationally. There will not be a single system, therefore a system of systems geared by interoperability is essential, most likely with appropriate application of regulation to enforce standards for sharing.
- A need for a valued-based approach to provision of benefits to varying communities, including consumers and Aboriginal and Torres Strait Islander people.

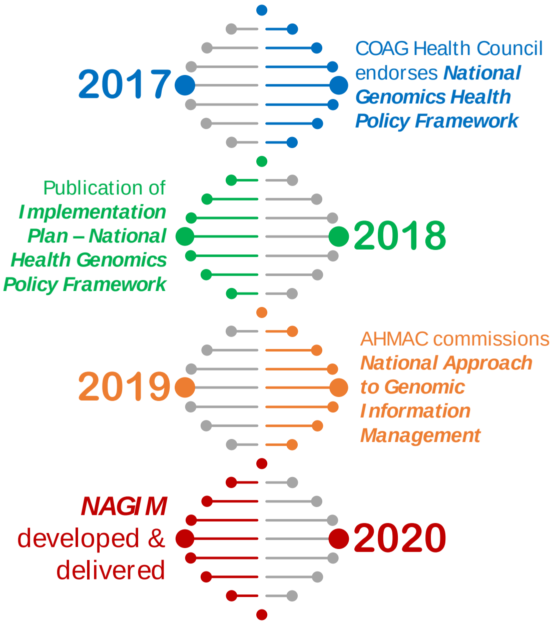
Cumulatively, this leads to a set of high-level requirements that describe desirable traits for any discreet solution for genomic information management, and specifically solutions that intend to work in an interoperable, standards based eco-system. A system where ethical and privacy-sensitive, context-based sharing is encouraged to advance our understanding of genomics-based medicine and its application to improve health outcomes for people and their communities.

This evolution towards an ecosystem underpinned by appropriate sharing will require a national genomic data governance framework to address both clinical, translational and research data governance. The elements of a data governance framework for genomic information management are described in the *NAGIM Blueprint*, with reference to national and international comparators, including:

- genomic data lifecycle management
- the data aspects of consent
- data sovereignty from a jurisdictional and Aboriginal and Torres Strait Islander perspective
- the issues of data ownership and commercialisation
- privacy and security
- data sharing, in both a research and clinical setting
- data quality, provenance and metadata
- data retention to meet accreditation and research requirements
- governance structures required to support all the above.

Finally, the *NAGIM Blueprint* addresses the current ecosystem of standards that may be applied to genomic data and information to support an interoperable ecosystem.

# 1 Introduction


2017
COAG Health Council endorses *National Genomics Health Policy Framework*
Publication of *Implementation Plan – National Health Genomics Policy Framework*
2018
2019
AHMAC commissions *National Approach to Genomic Information Management*
*NAGIM* developed & delivered
2020

Genomics is having an increasing role in healthcare in Australia, and our research community is working with others globally in the discovery of additional opportunities to apply genomic knowledge to healthcare and the prediction and prevention of disease. Likewise, the increasing application of genomics into everyday care is driving our health system to invest in genomics capabilities to understand clinical utility, sustainability and associated policy implications.

With this in mind, Health Ministers agreed Australia's first *National Health Genomics Policy Framework* (NHGPF) [1] in 2017. This framework provides a collaborative and coordinated approach at all levels of government and across stakeholders to align efforts to integrate genomics into the national health system.

The NHGPF identified five strategic priorities to support the integration of genomics into health care for Australians:

- **Person-centred approach:** Delivering high quality care for people through a person-centred approach to integrating genomics into healthcare
- **Workforce:** Building a skilled workforce literate in genomics
- **Financing:** Ensuring sustainable and strategic investment in cost-effective genomics
- **Services:** Maximising quality, safety and clinical utility of genomics in health care
- **Data:** Responsible collection, storage, use and management of genomic data

Following significant detailed and broad consultation, the *Implementation Plan—National Health Genomics Policy Framework* [2] was agreed by Health Minsters in 2018. Under Strategic Priority 5 (Data) of the *Implementation Plan*, it was noted that a key priority is to develop a digital health framework that can capture genomics information, so it ensures that Australia's digital health foundations support the advancement of genomics.

In line with the *Implementation Plan*, the National Approach to Genomic Information Management (NAGIM) project was sponsored by the Project Reference Group on Health Genomics, as an AHMAC cost-share funded (2019-20) project.

This *Blueprint for a National Approach to Genomic Information Management* (the *NAGIM Blueprint*) aims to establish a future state for national genomics information management in Australia to harmonise investments in, and linkage between, clinical delivery systems and research infrastructure.

## 1.1 Objective and scope

The objective for the *NAGIM Blueprint* is to provide guidance to those activities identified in Strategic Priority Area 5 (Data) from the NHGPF and the *Implementation Plan*. This will be achieved by:

- building on existing works being undertaken at a state/territory, national and international level
- establishing a blueprint and prime recommendations to address genomic data use
- informing and guiding future investment through a set of principles.

This Blueprint for genomics information management in Australia should inform and guide future investment for genomic medicine and research and facilitate sharing of experiences and approaches across Australia.

## 1.2 Structure of this document

This document contains these sections:

1. this introduction
2. a set of principles to guide future implementations
3. a genomic data categorisation framework to provide a consistent language for describing genomic data, including those elements that genomics work relies on
4. a discussion of considerations that must be made in the Australian context
5. options for logical architectures for genomic information
6. a framework for data governance of genomic data
7. a discussion about standards required to support interoperability.

An appendix is provided containing background material that may inform readers but is not necessary reading. This will cover a description of the general workflows within genomic medicine and genomic research, and the data used and produced by these processes.

## 1.3 Who should read this document?

This document will interest a broad audience, outlined below, representing a range of skills and understanding of genomics. This document attempts to convey these concepts in a fashion respectful of, and accessible to, this broad audience:

- **Clinicians, including pathologists, genetic counsellors and clinical network leaders**, who may need to better understand the genomic data needs of the research community and the existing national and international efforts in addressing these data requirements. Note that many clinicians are also researchers.
- **Policy makers, strategists and funders**, who need to gain an understanding of the specific nature of genomic data in both clinical and research settings to better plan for the management and utilisation of genomic data.
- **Health system administrators and operators of clinical genetics/genomics and diagnostics services**, who need to plan for adoption of genomics and the consequent impact on system data requirements and health service sustainability.
- **Information management professionals, digital health implementers and system integrators**, who need to understand the genomic data requirements for integration of systems and the management of genomic and related data.
- **Researchers, including bioinformaticians and medical scientists**, who will leverage the value of genomic data generated through clinical settings to make discoveries that will improve healthcare delivery. Note that many researchers are also clinicians.
- **Other diagnostic staff** (who may not be clinicians or researchers) including sequencing technicians, bioinformaticians, medical scientists and curators, who will support the delivery and operation of systems providing or manipulating genomic information.

- **Industry bodies and commercial organisations** engaged in planning, preparing and delivering genomic data services in Australia to clinicians working in the public and private health system, or to researchers and research funders.

Note that for readers less familiar with genomics, the Glossary in Appendix B includes information on a wide range of genomic topics. Appendix A provides a more detailed overview of the workflows and processes involved in genomics and the data created and used by these processes.

## 1.4 How does this document relate to the NHGPF?

The NAGIM project supports delivery of elements in both the NHGPF and the *Implementation Plan*. This can be seen in the following table.

| Document providing direction on activities | How this project supports this work |
|---|---|
| ***National Health Genomic Policy Framework*** | |
| **5.1** Establish a national genomic data governance framework that aligns with international frameworks. | Refer to the *NAGIM Blueprint* Section 6 (Genomic data governance framework) which highlights genomic governance requirements and proposes approaches to addressing them. |
| **5.1.1** Explore infrastructure options for national genomic data collection, storage and sharing. | Refer to the *NAGIM Blueprint* Sections 4 (Considerations for an Australian framework) and 5 (Proposed logical architecture). |
| **5.1.2** Strengthen public trust of data systems and mechanisms so that people are empowered to engage with genomic interventions in the health system. | Refer to the *NAGIM Blueprint* Section 2 (Principles for genomic information management) which addresses these issues. |
| **5.2** Promote culturally safe and appropriate genomic and phenotypic data collection and sharing that reflects the ethnic diversity within the Australian population, including for Aboriginal and Torres Strait Islander peoples. | Refer to the *NAGIM Blueprint* Section 2 (Principles for genomic information management) which addresses these issues. |
| **5.3** Develop nationally agreed standards for data collection, safe storage, data sharing, custodianship, analysis, reporting and privacy requirements. | Refer to the *NAGIM Blueprint* Section 7 (Standards and interoperability) which addresses these areas. |
| **5.4** Promote public awareness of the contribution of all research activities, including those funded through private industry, to advancing the application of genomic knowledge to health care. | This is beyond the project scope as defined by the Project Reference Group. |
| **5.5** Support sector engagement with international genomic alliances to promote shared access to data for research and global harmonisation of data where appropriate. | Elements of this form part of the Roadmap in the *NAGIM Blueprint* Section 5.6. |
| ***National Health Genomics Policy Framework – Implementation Plan*** | |
| **ACTION 19:**<br>Develop a national genomic data governance framework that provides for appropriate decision-making for governments and aligns with international frameworks. | Refer to the *NAGIM Blueprint* Section 6 (Genomic data governance framework) which highlights genomic governance requirements and proposes approaches to addressing them. |

| Document providing direction on activities | How this project supports this work |
|---|---|
| **ACTION 20:**<br>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.<br>B: Consider the national adoption of appropriate international standards on (but not limited to) phenotypes, disease classification systems and pathogenic variants. | Refer to the *NAGIM Blueprint* Section 7 (Standards and interoperability) which addresses these areas. |
| **ACTION 21:**<br>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 and improve engagement.<br>B: Explore opportunities to capture and integrate population genomic information to inform health care decisions, research and policies. | Refer to the *NAGIM Blueprint* Section 3.1 (Genomic Data Categorisation Framework) which addresses these issues. |
| **ACTION 22:**<br>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. | This is beyond the project scope as defined by the Project Reference Group. |
| **ACTION 23:**<br>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). | Elements of this form part of the Roadmap in the *NAGIM Blueprint* Section 5.6. |
| ***National Health Genomics Policy Framework – Implementation Plan*** | |
| **Priority Area 18 – Data and Digital Health**<br>Establishing plans to embed national health genomics data standards and agree a national approach to sharing data. | The *NAGIM Blueprint* lays out a set of principles and a framework for a common language to support sharing data, and guidance on governance issues and a roadmap for broader adoption. |
| **Priority Area 2 – Aboriginal and Torres Strait Islander Health**<br>Establish a national approach to optimise the clinical usefulness of a reference genome for Aboriginal and Torres Strait Islander peoples. | The *NAGIM Blueprint* addresses the particular needs of the Aboriginal and Torres Strait Islander peoples and provides principles that will guide how genomics can address these needs and benefit this community. |

## 1.5 A note on terminology used in this document

### 1.5.1 Genomics versus genetics

Consistent with the NHGPF, the term ‘genomics’ is used throughout these documents to refer to both the study of single genes (genetics) and the study of an individual’s entire genetic makeup (genome) and how it interacts with environmental or non-genetic factors.

While genetic testing for clinical purposes is already embedded in the health system, the term genomics is used for brevity and to acknowledge the cross-over of issues between genetics and genomics, other than where it is necessary to differentiate between them.

The terms genomics and/or genomic knowledge are used in this document and refer to the data, information and learnings derived through genomic research. It also refers to the technologies used for testing, analysing and furthering the discovery of genomic knowledge [3].

### 1.5.2 Genomic domains

Throughout these documents, these terms are used to reference three key areas where genomics is used:

- **Genomic medicine:** The application of genomics to healthcare services in a clinical setting (and sometimes called clinical genomics). This includes genetic counselling, clinical genetics, diagnostic and screening testing using genomic technologies and the clinical application of genomics.
- **Genomics research:** The study of genomics to discover new or refined information about how genomics influences or affects human health.
- **Translational genomics:** The translation of genomic research into healthcare delivery. This includes clinical trials and translational research. This term is used specifically to address aspects related to translational activities. However, many aspects of genomics research also apply to translational genomics.

NOTE: The term ‘genomic medicine’ has been used in preference to the term ‘clinical genomics’ because feedback from the community suggested that ‘clinical genomics’ can be confused with ‘clinical genetics’. The latter is a specific discipline in medicine and is therefore often associated with clinical genetics services. However, genomic medicine (also called genomic testing) reaches much farther than the discipline of clinical genetics and genomic testing can come from many disciplines in medicine. Genomic tests referrals do not necessarily go through clinical genetic services, especially in specialty areas such as neurology, nephrology, acute, oncology and pharmacy.

### 1.5.3 Germline and somatic genomics

This document refers to the application of genomics as defined by the NHGPF, encompassing both germline (heritable) and somatic (non-heritable) genomics, in diagnostic, predictive or therapeutic applications. This includes both germline genomics such as rare diseases, somatic genomics such as cancer, prenatal screening and other forms of genomics.

### 1.5.4 Genomics and other ‘omics

Unlike genetic testing, where the deoxyribonucleic acid (DNA) sequence of a single gene is checked for changes, genomics is the investigation of many genes at one time. Scientific and medical understanding of other ‘omics fields, including proteomics and metabolomics is moving forward quickly, and while these new areas are not the focus of this work, the project has remained mindful of the relation and emergence of these areas of science and their application to genomics discovery and medicine.

### 1.5.5 Data versus information

The terms data and information are often used interchangeably. However, according to Ackoff [4], data is considered the raw symbolic content that has no meaning beyond its existence, whereas information is data that has been processed and given meaning through connections to other data and information. This is reflected in genomics where raw sequence data comprising the four nucleotides can be transformed through analysis into genomic information.

This project includes ‘information’ in its title, but variously refers to data governance and information management in a variety of contexts. The intent of the project is to encapsulate the management of genomic data and information as a cohesive whole. This includes the generation and management of the raw genomic data, associated metadata and clinical information to support genomic interpretation, and the governance and processes to administer that data.

# 2 Principles for genomic information management

*“Principles are general rules and guidelines, intended to be enduring and seldom amended, that inform and support the way in which an organisation sets about fulfilling its mission.”* [5]

This Blueprint is based on a set of principles, rather than more detailed types of guidance, as principles remain more stable over the long term. Principles guide implementation without being prescriptive.

This document has been developed within the context of the principles that underpin the NHGPF. The NHGPF principles guide national decision-making in relation to genomic information management.

The principles identified in this section are already consistent with the National Health and Medical Research Council (NHMRC) *National Statement* [6] and Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) Guidelines [7] with which researchers and clinicians must comply. However, it is important to restate them in this new context, particularly with the consideration of the CARE Principles [8] (which are discussed in more detail in Section 2.9.4).

While these principles are already applied to research and clinical care, genomics introduces new challenges in managing large datasets that remain linked to individuals, that may persist across generations and be dynamically curated. This contrasts with the collection of data for a specific research or clinical application with a defined purpose and a prescribed period of use. These principles attempt to address the implications presented by the challenges and opportunities of genomics.

## 2.1 Structure of principles

The proposed principles conform to a consistent format [5] that includes:

- An **identifier** for reference throughout the *NAGIM Blueprint*
- A **name** that represents the principle and is easily remembered
- A **statement** which succinctly and unambiguously communicates the fundamental rule
- A **rationale** which expands on the statement and explains the logic behind the principle and the benefits from following it
- A set of **implications** which will describe consequent matters that fall out of the principle. Most of these will support /guide later decisions about implementations using the principles.

While principles can cover many aspects of an organisation, the principles of this Blueprint will focus on genomic data. The implications guide implementers on what factors should be considered regarding data (or things that affect data) when designing or deploying systems that manage genomic data.

Five criteria distinguish a good set of principles [5]:

- **Understandable**: The underlying concepts must be easily understood by individuals throughout an organisation or sector. The intention of the principle must be clear and unambiguous, so violations, whether intentional or not, are minimised.
- **Robust**: Robust principles inform good decisions about designs and plans and support the creation of enforceable policies and standards. Each principle should be sufficiently definitive and precise to support consistent decision-making in complex, potentially controversial situations.

- **Complete**: All important principles governing the management of information and technology are defined. The principles cover every anticipated situation.
- **Consistent**: Strict adherence to one principle may require a loose interpretation of another principle. The set of principles must be expressed so it allows a balance of interpretations. Principles should not be contradictory to where adhering to one principle would violate the spirit of another. Every word in a principle statement should be carefully chosen to allow consistent yet flexible interpretation.
- **Stable**: Principles should be enduring, yet sufficiently flexible to accommodate adaptation.

For the NAGIM Principles, non-prescriptive language has been used (‘should’ not ‘shall’ or ‘must’). While these principles describe a desirable approach to genomic information management, without a compliance scheme, using prescriptive language is without merit. This work has been limited to engagement and time available and requires further consultation before more prescriptive language being used.

## 2.2 A framework for the NAGIM Principles

The scope of this Blueprint covers a wide range of aspects of genomic data and creating a framework to structure the NAGIM Principles allows them to be placed into logical groups (domains) that are more easily applied.

The NHMRC published a set of principles for the translation of ‘omics’ in 2015 [9], and while these were focused on the translation of research into healthcare, a framework was defined for the principles. Figure 1 shows a set of domains used in this document, inspired by the NHMRC principles framework. This ‘virtuous circle’ demonstrates the interconnectedness and flow of benefits within the healthcare and research communities and is supportive of a learning health system [10].

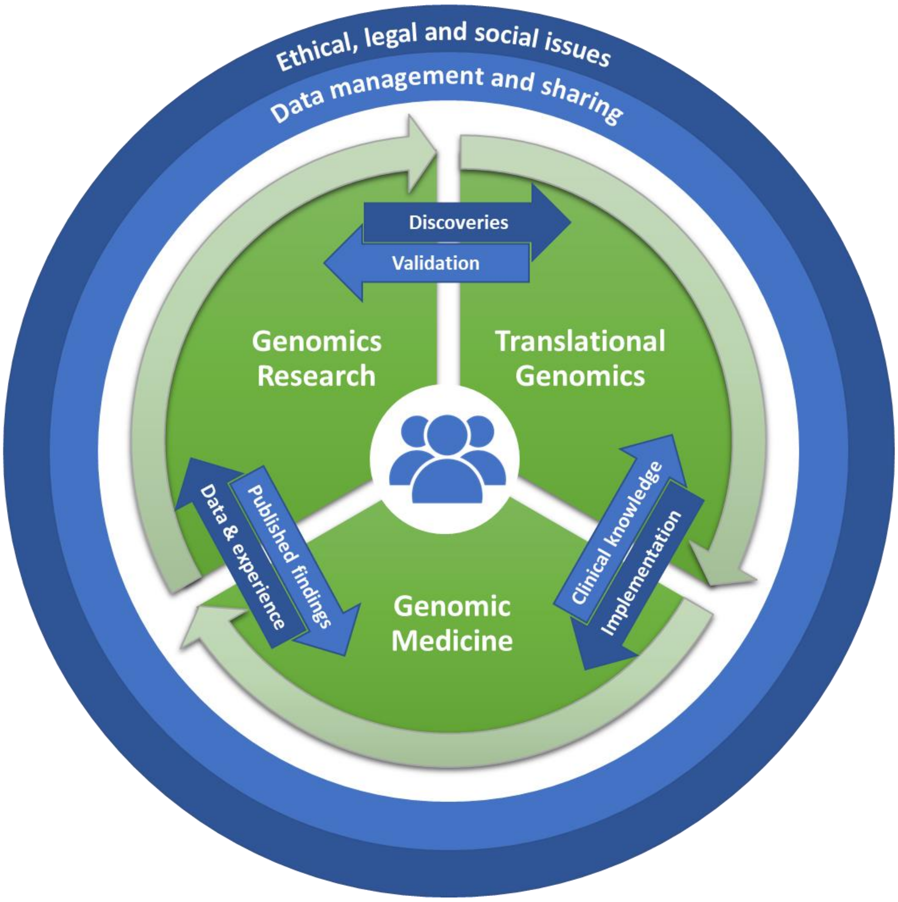


*Figure 1: Domains of interest within this Blueprint*

Elements of this framework include:

- **Genomics research** can use clinical data to support discovery but has traditionally relied on purpose specific collections. Research would benefit through access to richer, well curated genomic data and information resulting from clinical practice. Important aspects in genomic research include access to data, the value of shared infrastructure and the relationship with other realms of scientific endeavour (e.g. proteomics).
- **Translational genomics** transforms discovery into clinical practice. Translational research is impactful, supports the prioritisation of research activities and informs clinical practice.
- **Genomic medicine** leverages research discoveries and genomic knowledge to provide quality care. As a clinical discipline it is driven by the needs of accreditation, clinical attestation and clinical outcomes.
- **Data management** employs data governance to support data sharing between the above elements. It includes aspects common across all the three genomic areas above.
- Ethical, legal and social principles that frame all the above, including how we work with **Consumers** and specifically **Aboriginal and Torres Strait Islander peoples**.

Each principle must primarily serve the genomics domain where it resides but must also enable delivery of outcomes from principles in related domains. For example, the data management domain is a stand-alone set of principles that apply in context to genomics research, translation and medicine and through the overlap provide context and hence specific implications.

Together, these principles build a trust relationship between the clinical and research communities and the broader community at large (including Aboriginal and Torres Strait Islander people).

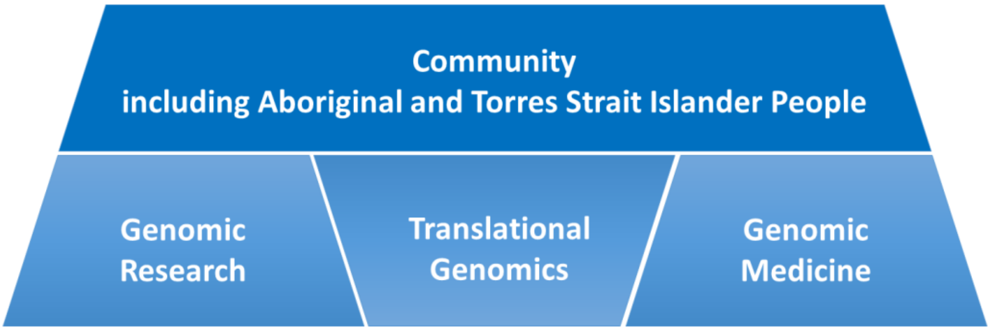


*Figure 2: Principles building a trust relationship*

## 2.3 Principles applicable to consumers and communities

Regardless of using the data, genomic data inherently interests consumers, carers and communities. Strong principles protecting these interests are critical to gaining the trust and social licence to use genomic data. This follows the person-centred approach recommended by the NHGPF [1].

| Principle | Statement, rationale & implications |
|---|---|
| **CN01: Person-centred focus** | **People must be the focus and beneficiary of advances in genomics and should be considered partners in this work.**<br>**Rationale:**<br>Ultimately people are the beneficiaries of genomics research and care. Genomics is capable of significant benefits for the wellbeing of consumers but can also create harm if not applied appropriately.<br>**Implications:**<br>• Researchers should understand the expectations of the individuals who provide genomic data, as expressed by the context of the consent provided.<br>• Researchers should consider both the benefits and potential harm that can result from research, especially when dealing with data related to specific communities at risk of vulnerability.<br>• Clinicians, funders and policy makers should collaborate, so genomics sequence results are used to benefit patients.<br>• A collective, informed conversation is essential to understand the implications of genomics on human health and society.<br>• Researchers should develop a ‘value statement’ that explains how people can benefits as partners in research.<br>• Data management planning needs to consider an individual’s cultural and religious beliefs on retention and destruction of genomic samples. |
| **CN02: Trust** | **Gaining and retaining trust of individuals and the community is fundamental to the practice of genomic medicine and research.**<br>**Rationale:**<br>Genomic and other health data is uniquely personal to consumers, and clinicians and researchers rely on an individual’s trust in systems to allow them access to this data. Gaining that trust and maintaining it is of critical importance to the ability of clinicians and researchers to access and use genomic data.<br>**Implications:**<br>• Free, prior and informed consent should be gained to ensure individuals trust the healthcare system with their genomic data.<br>• Strong governance models should support communities of interest and groups with specific needs.<br>• Transparency in how genomic data is managed and used to assure individuals of the strength of the governance put in place. This transparency should cover all aspects of genomic data. |

| Principle | Statement, rationale & implications |
|---|---|
| **CN03: Informed consent** | **Granting free, prior and informed consent is a foundation for all care and research.**<br>**Rationale:**<br>Free, prior and informed consent is a precondition for testing, treatment and research. This consent and the reason for its granting are core data elements that need to be managed as part of the overall genomic data management approach.<br>**Implications:**<br>• Systems should support the digital recording of consent and its context.<br>• These systems should support a variety of consent mechanisms, including family and community consent concepts.<br>• Systems should be able to cope with changes to consent, including the withdrawal of that consent.<br>• Consent as it applies to testing and treatment is different to and does not depend upon consent for research.<br>• Consent for research data sharing needs to consider the broad range of research scenarios in order to record flexible but appropriate consent settings.<br>• Consent for 'secondary uses' such as public benefit (such as population health analysis) or commercial interests need to be considered and a social licence for such use established. |
| **CN04: Right to access** | **Consumers may request access to their genomic and clinical data.**<br>**Rationale:**<br>As genomic data is the most personal of data, consumers should be able to request access to identifiable data about them to support their ongoing healthcare and that of their family. This is supported by Australia's privacy laws [11].<br>**Implications:**<br>• Systems should ensure they provide a mechanism for consumers to request access to their data.<br>• Systems should establish mechanisms for dealing with requests for data access.<br>• Consideration should be made for requests for data by family members.<br>• Data should be accessible to enable the best care, and to ensure that consumers benefit from the use and reuse of their data. |

| Principle | Statement, rationale & implications |
|---|---|
| **CN05: Use of data/portability** | **Consumer genomic data should be accessible and leveraged in multiple care settings.**<br>**Rationale:**<br>Consumers may seek care within a variety of healthcare settings. Allowing them to use existing genomic data in alternative settings avoids the need for additional testing and the costs to the system associated with such duplication.<br>**Implications:**<br>• Laboratories should be able to share data between health care settings/services where there is clinical consent from a consumer.<br>• Allowing this form of clinical data sharing will avoid the need for additional testing and costs to the system associated with unnecessary duplication.<br>• Systems should consider how to exchange genomic data between healthcare settings.<br>• Agreed standards should support interoperability.<br>• There may be opportunities to leverage the existing national digital health systems (such as the My Health Record system) to support consumer and clinician access to genomic data across care settings. |
| **CN06: Equity of access** | **All consumers have the right to equitable access to genomics-based care.**<br>**Rationale:**<br>The provision of care informed by genomic technologies should be available to all consumers equitably, regardless of location, race or socioeconomic background.<br>**Implications:**<br>• Reference data for specific groups, including Aboriginal and Torres Strait Islander peoples, should be available to allow these groups to benefit from genomic technologies.<br>• A national approach should consider how smaller organisations can benefit from the advances in genomics infrastructure and systems. |
| **CN07: Benefit from use** | **Consumers should benefit, either individually or collectively, from the use of their genomic information.**<br>**Rationale:**<br>Commensurate with the risks associated with the provision of their genomic data to research, consumers expect to obtain benefits from such research. These benefits may accrue to them individually or may benefit consumers collectively within communities of interest or more generally. Such benefits may not necessarily be financial.<br>**Implications:**<br>• Researchers should understand the expectations of the individuals that provide genomic data.<br>• Researchers should plan for and implement mechanisms for providing these benefits, including communication of their findings to individuals and communities. |

## 2.4 Principles applicable to Aboriginal and Torres Strait Islander genomics

While the principles for consumers and communities outlined in the previous section also apply to Aboriginal and Torres Strait Islander peoples, additional considerations are also required. The past experiences of Aboriginal and Torres Strait Islander people with scientific research, especially genomic research, has not always been positive [12]. Internationally, Indigenous communities, including Aboriginal and Torres Strait Islander communities, have suffered harm associated with lack of community engagement, lack of informed consent for secondary research, and negative representation in publications [13].

The ethical and cultural needs of both individuals and communities must be understood if the benefits and value of genomics is to support improvements in health and wellbeing for Aboriginal and Torres Strait Islander peoples. This requires genuine partnerships to be developed [14].

It should be noted that principle *CN03: Informed consent* addresses the important issue of consent. This is particularly important in the context of Aboriginal and Torres Strait Islander people and communities, and this is reflected in the inclusion of specific clauses calling for rights to free, prior and informed consent in the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP) [15].

The Global Indigenous Data Alliance (GIDA) have developed the CARE Principles [8] to address specific concerns of Indigenous populations internationally. The CARE Principles are described more fully in Section 2.9.4 below.

| Principle | Statement, rationale & implications |
|---|---|
| **IG01: Collective and individual benefit** | **Aboriginal and Torres Strait Islander peoples should derive collective and/or individual benefit from the use of their genomic data.**<br>**Rationale:**<br>Data ecosystems should be designed and function in ways that enable Aboriginal and Torres Strait Islander peoples to derive benefit from the data, through inclusion in the use of their data, improved governance and citizen engagement and equitable sharing of benefits derived.<br>**Implications:**<br>• Researchers and clinicians should understand the expectations of the individuals that provide genomic data. This requires engagement with the providing communities.<br>• Researchers and clinicians should plan for and implement mechanisms for providing these benefits back to providing individuals and/or communities.<br>• Reference data for specific groups, including Aboriginal and Torres Strait Islander peoples, should be available to support these groups to benefit from genomic technologies. |

| Principle | Statement, rationale & implications |
|---|---|
| **IG02: Authority to control** | **Aboriginal and Torres Strait Islander peoples have the authority to control the use of their genomic data for research purposes.**<br>**Rationale:**<br>The rights and interests of Aboriginal and Torres Strait Islander people in their data should be recognised and their authority to control such data be empowered, through recognition of those rights, use of their data in self-governance and the right to develop cultural governance protocols for their data.<br>**Implications:**<br>• Aboriginal and Torres Strait Islander peoples should have a majority representation in groups governing their data to support cultural security.<br>• Aboriginal and Torres Strait Islander peoples should have representation in consumer groups providing advice on the use of their data.<br>• Data repositories should have the granularity to allow representation by different groups or communities.<br>• Aboriginal and Torres Strait Islander peoples should define the protocols that define how data is used and by whom.<br>• Systems should ensure that they provide a mechanism for consumers to access their data.<br>• Systems should establish mechanisms for dealing with requests for data access. |
| **IG03: Responsibility** | **Researchers working with Aboriginal and Torres Strait Islander genomics should manage genomic data consistent with the wishes of Aboriginal and Torres Strait Islander peoples.**<br>**Rationale:**<br>Those working with research data about Aboriginal and Torres Strait Islander peoples should share how those data are used to support this community's self-determination and collective benefit. Accountability requires meaningful and openly available evidence of these efforts and the benefits accruing to Aboriginal and Torres Strait Islander peoples.<br>**Implications:**<br>• Researchers and clinicians should engage and understand the wishes of Aboriginal and Torres Strait Islander individuals and communities before working with genomic data so they can comply with those wishes.<br>• Systems should be transparent in how data is used and provide the ability to evidence these uses. |

| Principle | Statement, rationale & implications |
|---|---|
| **IG04: Ethics** | **Researchers should engage Aboriginal and Torres Strait Islander individuals and communities appropriately to ensure ethical standards are understood and maintained.**<br>**Rationale:**<br>Aboriginal and Torres Strait Islander people's rights and wellbeing should be the primary focus from the start and at all stages of the data life cycle and across the data ecosystem, to minimise harm and maximise benefits.<br>**Implications:**<br>• Researchers and clinicians should engage and understand the wishes of Aboriginal and Torres Strait Islander individuals and communities before working with genomic data so they can comply with those wishes.<br>• Researchers and clinicians should maintain engagement with the community to ensure that the outcomes of research follow the wishes of Aboriginal and Torres Strait Islander communities.<br>• Systems should support the digital recording of consent and its context.<br>• These systems should support a variety of consent mechanisms, including community consent concepts. |

While critical when considering managing the data for Aboriginal and Torres Strait Islander peoples, the CARE Principles provide guidance that could be applied to any group in society, and there is commonality between these principles and those within the other domains.

## 2.5 Principles applicable to genomic research

As genomics research scales to take advantage of larger datasets available through increased genomic testing in clinical practice, these principles are likely to drive the consideration of genomic data in a research setting.

| Principle | Statement, rationale & implications |
|---|---|
| **GR01: Rationalised repositories** | **The number of repositories storing genomic data shall be as many as needed but as few as possible to minimise the cost of duplication.**<br>**Rationale:**<br>There is a non-trivial cost of managing multiple repositories. By limiting the number of repositories as much as necessary, we can maximise the value of limited research funding and provide greater equity for access supporting scientific endeavour. However, even if there was only a single research genomics repository in Australia, it would be desirable to share data with other international repositories, so if we accept there will always be more than one, we need to make any repositories interoperable.<br>**Implications:**<br>• Repositories need to be designed with multiple uses in mind at the beginning.<br>• Repositories need to be extensible so they can contain genomic data from a variety of sources.<br>• Repository managers need the operational mandate and technical capacity/capability to appropriately share data with other genomics users to minimise unnecessary duplication.<br>• To support interoperability, repositories need to be standards-based.<br>• To maximise utility, repositories need to describe their contents, use and other factors so systems can interrogate and understand. Among a set of key drivers, provenance of data is essential within a data governance, lifecycle and reuse context.<br>• Repositories must comply with jurisdictional and national requirements and standards for interoperability, as outlined in *DM02: Multiple repository environment*. |
| **GR02: Ethical data and provenance** | **Data collected shall be collected under NHMRC guidelines.**<br>**Rationale:**<br>All research data should be collected ethically and with provenance, consistent with the guidelines from the National Health & Medical Research Council [6]. Due to the nature of genomic data, this is especially important, and the provenance extends to all aspects of how genomic data is processed and interpreted.<br>**Implications:**<br>• System must allow recording all elements of the data analysis process to allow for control of quality and audit.<br>• Systems that support access of external information sources need to ensure that a record of the source, when it was accessed, and the version is added to provenance data.<br>• Repositories shall be created that apply the FAIR and CARE Principles to maximise the value of the data and the repository.<br>• Use of data shall be made under the consent given for its use.<br>• NHMRC guidelines seek to remain relevant to contemporary expectations of data creators, users and importantly of public sentiment.<br>• The implications of *GM06: Genomic data is clinical data* share some of these requirements. |

| Principle | Statement, rationale & implications |
|---|---|
| **GR03: Reuse with permission** | **To maximise the value of genomic data, repositories should be designed with reuse and necessary permissions in mind.**<br>**Rationale:**<br>When designing data storage approaches, consideration should include not only the primary use of the data, but possibilities of reuse later and by others. This requires repositories to provide metadata information about their structure and content, and to be open and interoperable. Such systems should also implement permission-based access controls.<br>**Implications:**<br>• Repositories should store data using standards-based formats to promote use and reuse.<br>• Repositories should publish data dictionaries in a standardised format.<br>• Repositories should record the context of use with permission information so future use can be validated.<br>• Service definitions and standards are available to be employed by repository operators.<br>• Linkages between genomics data and other data sets should be considered where feasible. Data linkage has the potential to increase the range of uses for, and value of, genomics data.<br>• This principle relies on the concepts of consent for research data sharing noted in *CN03: Informed consent*. |
| **GR04: Plan for change and scalability** | **Changes to the nature and volumes of data available for genomic research should be understood and planned for.**<br>**Rationale:**<br>As genomic medicine becomes mainstream, the potential growth in genomics data from clinical sources is predicted to rapidly increase. This should be factored into the design of data repositories and processes to ensure that we can maximise the value of these repositories for research.<br>**Implications:**<br>• Healthcare services should analyse the likely changes to their data retention and data sharing plans for genomic data.<br>• Research and clinical groups should plan for how the additional data types and capacity can be leveraged to maximise value.<br>• Research organisations should plan for changes to their systems to support access to and use of the expanded capacity.<br>• As the number of connections to a few non-duplicated repositories increases, the demand on some services is likely also to increase.<br>• The architecture of repositories should support load balancing, region-based mirrors (if needed), and dynamic scalability.<br>• To share the value of repositories, repository managers should to budget for increased costs or develop methods of implementing cost recovery strategies.<br>• As Australian data repositories increase, the likely interest from international groups may increase demand on the Australian repositories. |

## 2.6 Principles applicable to translational genomics

Translational research holds a distinct place between basic and applied research and clinical services. These principles are likely to drive the consideration of genomic data as research is validated and applied in the clinical setting.

| Principle | Statement, rationale & implications |
|---|---|
| **TG01: Efficacy, utility and effectiveness** | **Translational research creates a focus on implementation factors supporting the effective application of genomic research into clinical practice.**<br>**Rationale:**<br>The healthcare system needs genomic research to be translated into clinical care to improve the effectiveness and sustainability of clinical care. This needs to include data that demonstrates the efficacy, utility and effectiveness of genomics in a clinical care setting.<br>**Implications:**<br>• Research institutions and healthcare systems should continue support for the translation of genomics into clinical settings to allow for improvements in efficacy, utility and cost effectiveness.<br>• Translational genomics should support the case for sustainability by collecting information that demonstrates the cost effectiveness of genomics and value-based healthcare.<br>• Clinical data should be available for support both pure research and translational genomics.<br>• Research and clinical participants should agree on data standardisation to ensure the incorporation of research data into usable clinical information (systems) within a clinical context. |
| **TG02: Leverage research flexibility** | **Research organisations could leverage their more flexible environments to develop learnings for healthcare.**<br>**Rationale:**<br>Research environments have flexibility to adapt and change more quickly than clinical systems. This flexibility should be leveraged to develop learnings for later application in clinical systems.<br>Genomic data needs to be stored using secure, privacy-aware approaches. Research environments can explore and test new approaches and technologies, helping to maintain trust in systems to a level that allows clinical data to be shared with research.<br>**Implications:**<br>• Leverage technologies that use security-first technologies to build secure system.<br>• Storage and compute capabilities should be demonstrated as safe, secure and low risk. The risks, vulnerabilities and regulatory aspects may differ between cloud-hosted services and local capabilities.<br>• Research institutions and healthcare systems should look to how continuous improvements in research environments can be managed safely in clinical settings, and compliant with regulation. |

| Principle | Statement, rationale & implications |
|---|---|
| **TG03: Emerging and validated knowledge** | **Rapidly emerging and validated knowledge from genomics research and translation supports genomics medicine.**<br>**Rationale:**<br>Interpretation of genomic tests and their application to clinical care relies on information shared or published by genomic researchers and others (including diagnostic laboratories). This is a rapidly evolving field, and access to and sharing of genomic data will support this.<br>**Implications:**<br>• Access to national and international databases is fundamental to supporting genomic medicine.<br>• Contribution of variant classifications and supporting rationale and detailed information to a national repository is essential for harmonisation of variant interpretations nationally. It is preferable this is a real-time contribution accessed at the time of variant curation, to limit conflicting interpretations between laboratories.<br>• Contribution of variant classifications and associated summary evidence to international repositories is an important contribution to a global pool of knowledge.<br>• Sharing clinical interpretation of genomic testing is critical for quality clinical outcomes.<br>• Agreement by laboratories to share variant classifications and interpretations and a system to exchange this information is critical.<br>• Efficient and granular sharing of phenotypic data, including Aboriginal and Torres Strait Islander status, is critical. |

## 2.7 Principles applicable to genomic medicine

As genomic medicine becomes mainstreamed in healthcare delivery, these principles should guide the use of genomic data in a clinical setting.

| Principle | Statement, rationale & implications |
|---|---|
| **GM01: Maximise clinical benefit** | **The management of genomics data will support clinical benefit in line with contemporary and developing practice.**<br>**Rationale:**<br>While translational research may be undertaken in a clinical setting, genomic testing of patients is undertaken to achieve a diagnosis, predict illness or to inform management. While there are cases where genomic data is re-examined later, the diagnostic process and the report (i.e. the clinical test) is generally the focus for clinical applications of genomics.<br>**Implications:**<br>• Data should be related to the patient in question and/or related parties.<br>• Data should be retained for as long as necessary (with consent) to provide quality care for the patient and for others. National Pathology Accreditation Advisory Council (NPAAC) requirements provide guidance on data retention.<br>• Systems should support moving from 'diagnose and treat' to 'predict and prevent' models of care.<br>• High quality genomic data is required to maximise clinical benefit. High quality genomic data extends beyond the statistical construct. Data can only be of high quality if associated meta data and provenance information are accessible so fitness for use and reuse can be reasonably determined.<br>• Evolving applications of genomics (such as pharmacogenomics and individualised treatment options) will continue to be developed, and systems need to be flexible to address the needs of these new applications. |

| Principle | Statement, rationale & implications |
|---|---|
| **GM02: Support for data sharing** | **Clinical sharing of genomic data within a healthcare setting should support patient care.**<br>**Rationale:**<br>Like other healthcare data, data related to genomics needs to be shared by clinicians within a health service to support patient care. Governance of data sharing should be implemented and controlled through standards, technology and products for the health service involved.<br>**Implications:**<br>• Access controls for genomics data should comply with local business rules, regulations and legislation.<br>• Access to all data (but especially genomic data) should be logged to support access audits.<br>• Within that context, genomic information should be available for all clinicians who require access to it to support provision of care.<br>• Genomic data should be available only to services accredited to manage, analyse and interpret genomic data, such as laboratories.<br>• Data exchange and access requires adoption of agreed standards to support interoperability and transparency.<br>• Genomics Information should be managed consistently with emerging standards and approaches employed in the Digital Health ecosystem.<br>• This principle relies on the concepts of consent for data sharing noted in *CN03: Informed consent*. |
| **GM03: Sharing across boundaries** | **Data sharing between healthcare settings should support patient care.**<br>**Rationale:**<br>Unlike traditional healthcare services, genetic health services frequently record information about families, not just the subject of care. As populations becomes increasingly mobile, this may require the exchange of data between health services and/or jurisdictions and their traditional regulatory and jurisdictional boundaries.<br>**Implications:**<br>• Standardisation of data should support sharing or querying across systems and applications. This may include inter-jurisdictional queries (as opposed to transferring data sets).<br>• Existing healthcare legislation, regulation and policy may need to be updated to reflect the need for sharing familial and other healthcare data across jurisdictional boundaries.<br>• Existing EHRs are patient-centric and episodic, and a way of recording both longitudinal and familial data consistently may need to be developed or employed through sophisticated systems integrated. |

| Principle | Statement, rationale & implications |
|---|---|
| **GM04: Sustainable genomics** | **Genomic data management should be cost-effective and sustainable.**<br>**Rationale:**<br>Undertaking genomic testing and retaining the data should support improved cost effectiveness and sustainability of the healthcare system. Data custodians should balance the cost of maintaining their data assets and the value of those assets.<br>**Implications:**<br>• Systems should monitor both costs and cost savings to establish evidence for using genomic testing.<br>• Custodians should engage with other actors to understand the value and the full (and potentially extensible) lifecycle of data.<br>• Data custodians should measure the usage of the data they manage as a proxy to assessing its value, and the impact of any usage.<br>• The benefits of clinical and research use of clinically generated genomic data need to be understood as an offset of any costs incurred.<br>• Data custodians should consider direct feedback mechanisms to assess the value of the data assets held by researchers.<br>• Data must be of high quality to ensure that the value of the asset is maximised.<br>• Data sets for groups with specific needs should be identified to allow for reporting on equity of access. |
| **GM05: Frequently updated knowledge** | **In a rapidly changing field of knowledge the outcomes may change with the application of updated information.**<br>**Rationale:**<br>Genomics is an evolving science and is based upon information that is rapidly changing. Interpretations and decisions made at a point in time may change as greater knowledge and understanding evolves.<br>**Implications:**<br>• Accredited laboratories should be supported to review and revise diagnoses made on changing genomic information/knowledge.<br>• Changes to diagnoses based on new genomic information/knowledge should be understood as learned improvements and not past mistakes.<br>• Mechanisms to identify and assess changing genomic information/knowledge need to be established.<br>• Consent mechanisms need to consider reprocessing of data. |

| Principle | Statement, rationale & implications |
|---|---|
| **GM06: Genomic data is clinical data** | **An individual’s genomic data should be treated no differently to other forms of clinical data.**<br>**Rationale:**<br>Regulation already protects the privacy and confidentiality of all health records (medical, pathology etc.), including genomic data within those records. Well established policies and procedures exist within organisations to comply with the relevant regulatory environment. Exceptions for genomic data may increase the risk of lack of compliance.<br>**Implications:**<br>• Patients should be informed of when and why data may be accessed and used.<br>• Patients should be informed of the implications of their data to other family members.<br>• An individual’s privacy should be protected when their data is used to assist the clinical care of other family members.<br>• The collection of genomic data must meet the same standards for ethical and professional collection as other clinical data. This is similar to the requirements of *GR02: Ethical data and provenance*. |

## 2.8 Principles applicable to data management

Across the three domains, there are shared considerations when looking at the way data is managed, governed and shared.

| Principle | Statement, rationale & implications |
|---|---|
| **DM01: Clinical vs research data** | **The differences between clinical and research data use should be understood and documented.**<br>**Rationale:**<br>While the underlying data may be identical in form, the way genomic data is used in a clinical setting is different to its application in a research setting. It is important to maintain visibility of these distinctions when addressing issues such as consent, privacy and security.<br>**Implications:**<br>• Jurisdictions and research communities should document the nature of and use of data to support data sharing agreements.<br>• Where possible these uses should be harmonised to support national implementation efforts.<br>• Implementations should recognise and support a variety of use cases and data sources.<br>• There are legislative, regulatory and policy requirements for both clinical care and research operating at national and jurisdictional levels which need to be considered when working with genomic data. |
| **DM02: Multiple repository environment** | **Multiple repositories should be supported by interoperability.**<br>**Rationale:**<br>There are jurisdictional, institutional and practical constraints on the way data is stored within healthcare and research. Genomic information management will be delivered via multiple, interoperable data repositories, but the number and type of these repositories should be managed. This means that a national approach should be flexible in how it works with existing and planned or future systems.<br>**Implications:**<br>• Interoperability between systems will be supported by the selection and adoption of standards to allow the efficient exchange of data.<br>• Access to clinical information held in different locations will be supported by standards-based access and compute capabilities.<br>• Interoperability permits the coexistence of a diverse set of repositories, of varying maturity and sophistication. |

| Principle | Statement, rationale & implications |
|---|---|
| **DM03: Genomic data retention** | **Minimum laboratory data retention rules may not support research needs for genomics.**<br>**Rationale:**<br>Regulation and accreditation define minimum data retention periods focused on clinical and medicolegal considerations, which vary depending upon the data involved. Retention of data beyond these periods is common but comes with a cost that the health system incurs.<br>**Implications:**<br>• The costs of genomic data management within healthcare systems should be understood and budgeted.<br>• The benefits of genomic research in clinical settings should be understood as an offset of any costs incurred.<br>• Longer term storage of data may require establishment of infrastructure to support clinical and research genomic objectives.<br>• Understanding the value of data and the cost of retention may be best served through an ecosystem approach, rather than through the view from a single creator or user of the data. |
| **DM04: National and international collaboration** | **Genomics knowledge and use is an international endeavour and national and international collaboration will be required to maximise benefits to healthcare.**<br>**Rationale:**<br>Organisations across Australia and internationally are working to support the application of genomics in healthcare and research settings. This includes efforts in standardisation, addressing ethical, legal and social issues and developing better processes for consent and data sharing. Australia should assess, evaluate and leverage the learnings from these projects, and continue to collaborate nationally and internationally.<br>**Implications:**<br>• National and cross-jurisdictional collaboration should maximise the value of genomic data to Australia and Australians.<br>• International collaboration is important to leverage the value of investments made by international partners to Australian organisations.<br>• Investments in genomic collaborations should be leveraged to provide support and equity for genomics in healthcare across Australia. |

| Principle | Statement, rationale & implications |
|---|---|
| **DM05: Strong governance models** | **Strong governance models should support communities of interest and specific groups.**<br>**Rationale:**<br>Genomic data is considered intensely personal by the community. While strong governance is important for all individuals and communities, there are groups within the population that can only benefit from genomics if strong and equitable governance is put in place to support protection and the flow of benefits to those communities.<br>**Implications:**<br>• National governance models should be agreed to provide consistency of how genomic data is managed.<br>• Governance models should protect the rights and privacy of specific groups within our community, not just the majority.<br>• Groups with specific needs should be sufficiently represented in the governance groups managing their data.<br>• The governance models should incorporate strong clinical governance considerations.<br>• Data protections should be incorporated to address the risks for family and community members of inappropriate use of genomic data. |
| **DM06: Contemporary use and contribution** | **Australia should contribute to international repositories and should use contemporary genomic references.**<br>**Rationale:**<br>Genomics relies on reference genome resources and curated interpretation of variants to this reference. Australian genomic organisations should enrich reference genomes datasets and contribute to this and curated interpretations.<br>**Implications:**<br>• Investments should be made to support currency of reference genomes to reflect improvements and extensions of this resource.<br>• Australian laboratories should contribute to the ongoing improvements in the reference genomes to reflect our diverse population, including Aboriginal and Torres Strait Islander peoples.<br>• Australian laboratories should contribute to sharing clinically validated variant interpretations used by researchers and clinicians.<br>• This principle relies on the concepts of consent for data sharing noted in *CN03: Informed consent*. |

| Principle | Statement, rationale & implications |
|---|---|
| **DM07: Pragmatism** | **Pragmatism and open and informed discourse will support a balance between the needs of clinicians, researchers and consumers.**<br>**Rationale:**<br>Balancing the needs of consumers, clinicians and researchers will identify areas of contention. Pragmaticism supports a balance of the needs against the outcomes.<br>**Implications:**<br>• When designing data management systems, a broader view with multiple stakeholders may be required.<br>• A collaborative approach to design that includes all stakeholders will be necessary.<br>• Innovation and changes to existing behaviours may be needed to allow all participants to obtain value from genomics.<br>• There is a cost to retain data which needs to be balanced against the benefits of retaining that data.<br>• Not all participants have the same levels of maturity or are changing at the same speed, meaning that solutions must cope with a variety of scenarios concurrently. |
| **DM08: Data quality** | **The quality of all data contributed to the genomics ecosystem is critical to maximising the potential of the whole ecosystem.**<br>**Rationale:**<br>Genomic data and the processes used to provide it are complex and require high levels of assurance to ensure that results and interpretation of those results are accurate and insightful.<br>**Implications:**<br>• Systems, including information technology, should contribute and support data quality rather than impede it.<br>• Workforce development is required to support participants to use the emerging techniques and technology.<br>• Jurisdictions and research institutions should work with accreditation bodies to ensure that all parties develop approaches for maximising quality of data and interpretation.<br>• Recognise that quality goes beyond the laboratory processes currently the focus of accreditation processes.<br>• Standards for measuring data quality are necessary to support improvements.<br>• Data quality extends beyond the statistical notion of quality of the genomic sequence data. Data quality includes metadata and provenance information that allows for determination of fitness for use and reuse of the data.<br>• Systems must support capture, storage and access to accurate, contextual and current metadata and provenance information. |

## 2.9 Other principles and related frameworks

Besides the principles described above, several sets of principles broadly apply across the domains and should be part of any broader implementation.

### 2.9.1 Australian Privacy Principles

The Office of the Australian Information Commissioner (OAIC) is the independent national regulator for privacy and freedom of information. They promote and uphold rights to access government-held information and the protection of personal information [11].

OAIC manages the Australian Privacy Principles (APP) [16] which provide the privacy protection framework for personal information in Australia, in accordance with the Privacy Act 1988. The 13 APPs govern standards, rights and obligations around:

- the collection, use and disclosure of personal information
- an organisation or agency's governance and accountability
- integrity and correction of personal information
- the rights of individuals to access their personal information.

The Privacy Act is a federal law which does not cover local, state or territory government agencies, except the Norfolk Island administration. Most Australian states and territories have equivalent legislation which covers their public sector agencies. Some state authorities and instrumentalities are bound by the Privacy Act [17].

The Privacy Act provides extra protections around handling health information. However, state and territory public hospitals and health services are not covered by the Privacy Act but may be covered by state or territory legislation [18].

The European Union's General Data Protection Regulation (GDPR) [19] has similarities to the APPs, but also specific differences [20]. Specific implementations will need to consider whether they need to comply with the GDPR, particularly where data sharing with Europe is proposed.

### 2.9.2 National data sharing principles

The Office of the National Data Commissioner (ONDC) [21] is responsible for streamlining how public sector data is used and shared to:

- promote greater use of public sector data
- drive innovation and economic benefits from greater use of public sector data
- build trust with the Australian community around government's use of data.

The ONDC is developing a data sharing framework for public sector data. The new legislative framework will help overcome barriers which prevent efficient use and reuse of public sector data, while maintaining the strong security and privacy protections that the community expects. While not finalised during writing, this framework and associated legislation will need to be considered by any national genomics information management arrangements.

ONDC has published a set of five principles related to data sharing [22]. This is based on the Five Safes [23], [24], a framework for helping make decisions about making effective use of data, which is confidential or sensitive, as the ONDC principles are as follows:

- Projects: Data is shared for an appropriate purpose that delivers a public benefit
- People: The user has the appropriate authority to access the data
- Settings: The environment in which the data is shared minimises the risk of unauthorised use or disclosure
- Data: Appropriate and proportionate protections are applied to the data

- Output: The output from the data sharing arrangement is appropriately safeguarded before any further sharing or release.

While these principles relate to sharing public sector data, they also provide good general guidance for any data sharing activity.

### 2.9.3 FAIR principles for researchers

The FAIR principles [25], [26] where defined in 2016 and are designed to:

- support knowledge discovery and innovation
- support data and knowledge integration
- promote sharing and reuse of data
- be applied across multiple disciplines
- help data and metadata to be 'machine-readable'
- to support new discoveries through the harvest and analysis of multiple datasets and outputs.

The principles are:

- **Findable**: This includes assigning a persistent identifier, having rich metadata to describe the data and making sure it is findable through local or international search portals.
- **Accessible**: This may include making data open (where possible) using standardised protocols. Sensitive data may not be made open due to privacy concerns, national security or commercial interests. Data that is not open, should provide clarity and transparency around the conditions governing access and reuse.
- **Interoperable**: This involves using languages, formats and vocabularies in the data and metadata accepted by the community of practice. Metadata should reference and describe relationships to other data, metadata and information through using identifiers.
- **Reusable**: Reusable data needs a clear (machine-readable) licence that defines usage and provenance information on how the data was formed. Domain-specific data and metadata standards should give it rich contextual information that will support reuse.

FAIR data is supported by most academic institutions and the Australian Research Data Commons (ARDC) [27].

### 2.9.4 CARE Principles

The Global Indigenous Data Alliance (GIDA) have built on the FAIR principles to address specific concerns of Indigenous[1] populations internationally. The CARE Principles [8] are:

- **Collective benefit**: Data ecosystems shall be designed and function in ways that enable Indigenous peoples to derive benefit from the data, through inclusion in the use of their data, improved governance and citizen engagement and equitable sharing of benefits derived.
- **Authority to control**: The rights and interests of Indigenous peoples in their data must be recognised and their authority to control such data be empowered, through recognition of those rights, use of their data in self-governance and the right to develop cultural governance protocols for their data.
- **Responsibility**: Those working with Indigenous data have a responsibility to share how those data are used to support Indigenous peoples' self-determination and collective benefit. Accountability requires meaningful and openly available evidence of these efforts and the benefits accruing to Indigenous peoples.

[1] The term Indigenous is used here to refer to First Nations people internationally and as used by GIDA in their publications.

- **Ethics**: Indigenous peoples' rights and wellbeing should be the primary concern at all stages of the data life cycle and across the data ecosystem, in order to minimise harm and maximise benefits.

The CARE Principles are people and purpose-oriented, reflect the crucial role of data in advancing Indigenous innovation and self-determination, and complement the existing FAIR principles encouraging open and other data movements to consider both people and purpose in their advocacy and pursuits.

### 2.9.5 GA4GH framework for responsible sharing of genomic data

The Global Alliance for Genomics & Health (GA4GH) have published their own framework [28]. The framework guides the responsible sharing of human genomic and health-related data, including personal health data and other types of data that may have predictive power in relation to health. Its guidance incorporates the principles in both the FAIR and CARE models.

The GA4GH framework highlights and is guided by Article 27 of the 1948 Universal Declaration of Human Rights, and in particular the rights of privacy, non-discrimination and procedural fairness. In addition to technical standards around the management and sharing of genomic data, GA4GH has developed a suite of policies on specific issues such as ethical governance, consent, privacy and security.

### 2.9.6 World Economic Forum whitepaper

The World Economic Forum has released a whitepaper entitled *Genomic Data Policy Framework and Ethical Tensions* [29] which provides principles that address:

- Consent – Comprehension, openness, respectfulness, fitness for purpose and renotification
- Data privacy – Autonomy, confidentiality, non-maleficence, beneficence, transparency
- Data access – Restraint, consideration, responsibility, reliability, accountability, vigilance
- Benefit sharing – Justness, cooperation, clarity, dignity, inclusion

The whitepaper also addresses six ethical tensions:

- Balancing individual privacy and societal benefits
- Balancing open and restricted data access
- Balancing receiving benefits and altruistic donations
- Balancing community and researcher oversight
- Balancing inclusion and exclusion
- Balancing confidentiality and duty to inform

This work reflects many of the principle outlined in this Blueprint and other referenced documents, and leverages work undertaken by the GA4GH and others.

# 3 Types of data covered by this Blueprint

The term 'genomic data' is frequently used to refer to the raw data derived from sequencing instruments, the aligned genomic sequence data, a person's genome (in whole or in part) or individual DNA variations. However, when considering a national blueprint for clinical, translational and research applications of genomics, a broader definition is demanded.

## 3.1 Genomic data categorisation framework

A data categorisation framework[2] has been developed to allow the broader data required to support genomics to be identified, as shown in Figure 3. This is intended to cover genomic medicine, translational genomics and genomic research.

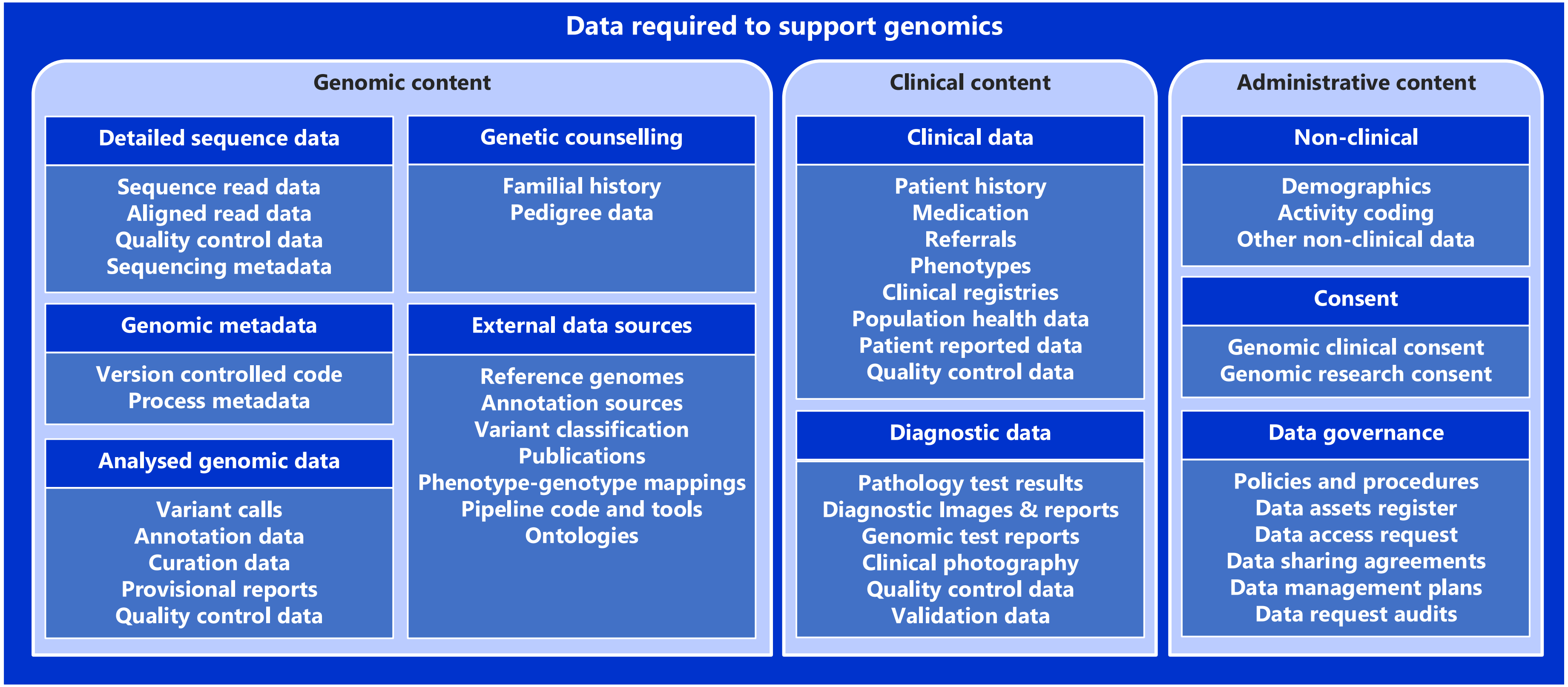
Categorisation of information against tiered 'domains' (specific subject areas) is a technique that provides a consistent and convenient method for logically grouping elements of an enterprise architecture. This allows the architecture to reflect the nature of the business being supported or the function of assets/services [30].

The categorisation framework identifies three broad groups of data classifications:

- **Genomic content:** This groups all data that may traditionally have been considered 'genomic data', including data from clinical areas and research. It also covers genomic data from direct-to-consumer sources.
- **Clinical content:** All genomic data activities require access to a range of clinical content to support decision-making, genomics interpretation and to support research.
- **Administrative content:** This group includes all supporting information required to manage data governance and to support the broader discipline of genomics.

These groups and their categories are described below. Note it is impossible to list all current and potential future classes of data, but this categorisation framework should assist to guide data categorisation and implementation activities.

[2] Such architectural frameworks are normally called "classification frameworks". However, the term "categorisation framework" has been selected to avoid confusion with the process of variant classification used in genomics.



Data required to support genomics

Genomic content

Detailed sequence data

Sequence read data
Aligned read data
Quality control data
Sequencing metadata

Genetic counselling

Familial history
Pedigree data

Genomic metadata

Version controlled code
Process metadata

External data sources

Reference genomes
Annotation sources
Variant classification
Publications
Phenotype-genotype mappings
Pipeline code and tools
Ontologies

Analysed genomic data

Variant calls
Annotation data
Curation data
Provisional reports
Quality control data

Clinical content

Clinical data

Patient history
Medication
Referrals
Phenotypes
Clinical registries
Population health data
Patient reported data
Quality control data

Diagnostic data

Pathology test results
Diagnostic Images & reports
Genomic test reports
Clinical photography
Quality control data
Validation data

Administrative content

Non-clinical

Demographics
Activity coding
Other non-clinical data

Consent

Genomic clinical consent
Genomic research consent

Data governance

Policies and procedures
Data assets register
Data access request
Data sharing agreements
Data management plans
Data request audits

*Figure 3: Data categorisation framework for data required to support genomics*

## 3.2 Genomic content

This grouping includes the data related to the process of genomics from genetic counselling to interpretation of genomic results. It is also the category into which other 'omics data would be added as these technologies develop.

### 3.2.1 Genetic services

The role of genetic services includes an education process that seeks to assist affected (and/or at risk) individuals and their families to understand the nature of the genetic disorder, management/surveillance options, the risk to family members, and the role, options, availability and possible outcomes of genetic testing.

Genetic services data is rarely held within traditional electronic medical record systems for two reasons: confidentiality issues associated with assumed family relationships and the patient-centric approach taken by most EHRs does not well support families or family relationships.

| Sub-classification | Description |
|---|---|
| Familial history | A comprehensive family history may be held within the genetic service and relies on (electronic) records of the treating clinician and information related by the patient. |
| Pedigree data | Pedigree information must support analysis of genomic data, especially for inherited disease. |

### 3.2.2 Detailed sequence data

This classification includes the 'traditional' concepts of DNA sequence data suggested by the term genomics.

| Sub-classification | Description |
|---|---|
| Sequence read data | The sequence read data is the data generated by the sequencing technology used. The most common file format is FASTQ. FASTQ format is a text-based format for storing both a biological sequence and its corresponding quality scores. |
| Aligned read data | Sequence alignment refers to aligning the sequence read data against a reference genome. Once the raw sequence data has been aligned, it is stored for later analysis. The most common file format is Binary Alignment Map (BAM) which is a compressed format. |
| Quality control data | Technical artefacts are introduced into genomic data by the sequencing process. The specific bias introduced by each technology must be analytically accounted for to correctly call genomic variants and discount artefacts. Quality control metrics are designed for this purpose [31]–[34].<br>Other quality control factors include but are not limited to specimen quality; read depth (ensuring adequate coverage to call variants); fragment insert size; and the quality of the sequence library. |
| Sequencing metadata | Metadata describe the what, where, how and when of the process from collection to sequence generation, plus contextual data such as environmental conditions or clinical observations [35]. Metadata can include data about how the data was generated, for example, library method, instrumentation, batching, alignment/build, pipelines. |

### 3.2.3 Genomic metadata

Bioinformatic analyses orchestrate files through transformations, called a pipeline or a workflow. Massively parallel sequencing (MPS) generates millions of short DNA sequences, which has increased the need for robust pipelines. MPS analyses involve steps such as sequence alignment (see Appendix A.8.1) and genomic annotation that are both time and computationally intensive.

| Sub-classification | Description |
|---|---|
| Version controlled code | Pipeline code may be written in one or more computer languages specific to the infrastructure that will run the pipeline, or may use a program (an orchestration or workflow engine) that executes pipeline commands written in an orchestration language used to define pipeline commands, which the program then translates to the appropriate vendor-specific commands.<br>Change management of workflow code is critical metadata to be recorded as part of the provenance of the result. Pipelines for clinical diagnostics are typically relatively stable, as accreditation requirements require changes to be rigorously tested. Pipelines in a research context are generally developed as part of the research process and may undergo significant changes during this process.<br>Components and approaches (workflow languages and tool definitions) may be shared, however, complete pipelines are less commonly shared between laboratories. |
| Process metadata | Throughout the steps of the pipeline process, metadata must be recorded to support the provenance of the resulting genomic data. The metadata can be significant data streams and should include the version and source of any tool or data source used. |

### 3.2.4 Analysed genomic data

Once the raw sequence data is aligned, the analysis processes produce several new data types on the journey to interpretation.

| Sub-classification | Description |
|---|---|
| Variant calls | The process of variant calls and copy number variations identifies changes between the sample and a reference source. Data is usually stored in a Variant Call Format (VCF) [36], which is the de facto standard format of a text file used in bioinformatics for storing gene sequence variants.<br>There are also other file formats used for different purposes for example Mutation Annotation Format (MAF) files commonly used for somatic single-nucleotide polymorphism (SNV) and indel data. Once data is annotated, tools may generate other outputs such as tab-separated values (TSV). |
| Annotation data | Variant-level information, relevant to variant classification. Variant annotation may describe the variant in Human Genome Variation Society (HGVS) format. Details relevant to disease phenotype (from test request etc.), may be added. Annotation could include designation of common low-risk variants, should these relate to risk profiling for patients. |
| Curation data | Evidence relevant for clinical interpretation of a variant regarding the genetic test request (and possibly for secondary/incidental/additional findings). This will include reference clinical information (collected as part of the test request, or elsewhere). |

| Sub-classification | Description |
|---|---|
| Provisional reporting | In genomic medicine, the classification of variants and clinical interpretation of the results is critical to the delivery of a diagnostic report. This report and its interpretation are important to:<br>• help to make/refine a diagnosis<br>• inform further testing, treatment plans and management strategies<br>• reveal patterns of inheritance and assess likelihood of genetic disease in relatives<br>• highlight need for specialist referral<br>• correct any family misconceptions.<br>In genomic research, the interpretation process will assess the results against the initial research hypothesis and prepare for later publication of the results. |
| Quality control data | Quality control data may include the number of variants, the proportion of variants in public databases such as dbSNP, the sequence change and context and the proportion of reads containing each variant. Variants should be visualised in applications such as the Interactive Genomics Viewer to check quality. |

### 3.2.5 External data sources

Genomic activities rely on many external data sources. These will evolve over time to reflect changes in technology and available data sources. Examples are used in the description to illustrate concepts only.

| Sub-classification | Description |
|---|---|
| Reference genomes | Reference genomes support alignment activities and variant calling processes. |
| Annotation sources | Information about variants including frequency in the population, known biological function of variants, sequence features and conservation and predicated impact on protein. |
| Variant classifications | Information shared by other laboratories supporting classification of variant pathogenicity are critical to genomic curation. This will include information on curated variants and evidence on how a variant was classified (e.g. American College of Medical Genetics (ACMG) guidelines) [37]. |
| Publications | Access to current published literature on genomic discoveries is critical to assessing variants. |
| Phenotype-genotype mappings | The mapping of phenotype against genotype support interpretation of genomic findings. |
| Pipeline code and tools | The software, tools and code required to undertake development of bioinformatic analysis pipelines are commonly shared between institutions. |
| Ontologies | Ontologies support semantic interoperability between systems by providing standardised controlled vocabularies for data storage. |

## 3.3 Clinical content

All genomic data activities require access to a range of clinical information to support decision-making, genomic interpretation and to support research. Within a clinical setting, much of this information will be available in the electronic medical records, patient administration systems and diagnostic management systems.

To support research, access to high quality, longitudinal clinical data on treatment decisions and clinical outcomes, linked to or stored with genomic data, can lead to clinical benefits through genomic discovery.

### 3.3.1 Clinical data

Clinical data is critical for understanding genomic data. It influences selection of genomic investigations and interpretation of the results.

Clinical systems can contain a wide range of data from family history, patient clinical phenotypes, medical records and diagnostic test requests and results. While digital health programs across Australia endeavour to promote the storage of this information in interoperable electronic formats, adoption varies widely.

| Sub-classification | Description |
|---|---|
| Patient history | A detailed patient history can be held within the (electronic) medical records of the treating clinician and is fundamental to establishing phenotype and providing context for interpretation of genomic testing. This includes observations of vital signs and other clinical and lifestyle indicators (such as smoking status), and coding of clinical data and records of care (such as International Classification of Disease (ICD) coding) [38]. |
| Medications | Medication data, within both a hospital and community setting, is a key element of the clinical narrative. Such data has considerable benefit to precision medicine and pharmacogenomics [39]. |
| Referrals | Referrals both in and out of a clinical setting provide basic patient data and clinical indicators, which may relate to genomics. |
| Phenotypes | Phenotype data is critical to genomics, with phenotype-genotype relationships important for test selection and interpretation. However, EHRs are largely focused on data collected for clinical care and funding purposes, not biomedical research.<br>A clinical phenotype repository holds identified clinical/phenotype data for patients. It may be separated from the genomic data and restricted to those people doing the final reporting process to address privacy requirements.<br>Types of phenotype data that may be stored include general biochemical, specialist biochemical genetic, imaging or histopathology. |
| Clinical registries | Clinical registries provide longitudinal data about patients within specific disease classifications, and are generally deemed credible, effective and feasible tools to measure variation and drive quality improvement at the national and jurisdictional health system levels [40]. When linked to genomic data, they can provide insight into patient outcomes, improve patient treatment decisions and provide important data to support clinical and translational research [41]. |
| Population health data | Population-based registers capture longitudinal data about entire populations and can be used to investigate information including specific outcomes (e.g. cancer diagnosis, death, survival etc.). Organisations such as the Australian Institute of Health and Welfare (AIHW) and jurisdictional groups record this type of data. |
| Patient reported data | Patient recorded data can include patient recorded outcome measures (PROMs) and patient recorded experience measures. They may include surveys, results from focus groups, patient diaries and observations. |
| Quality control data | Assessments of completeness and used of standardised coding can be taken to assess the quality of clinical data. |

### 3.3.2 Diagnostic data

Access to other diagnostic test results including pathology and imaging is important to the selection of gene panels for study and the interpretation process within genomics.

In clinical settings, the raising of a genomic test order is the start of the genomics diagnostic process. The quality of phenotype information in such orders has a demonstrable impact on genomic interpretation efficiency (and test turnaround time). Existing order entry systems rarely support such phenotype data entry or sharing.

| Sub-classification | Description |
|---|---|
| Pathology test results | Electronic medical records may support the entry of pathology test requests (including genomic testing) and a record of the results. In Australia, this varies widely between health services depending upon the combination of EHRs and Laboratory Information Management System (LIMS). |
| Diagnostic images and reports | The combination of genomics and modern imaging techniques is leading to new applications of imaging genetics in neuroscience [42] and radiogenomics in oncology [43], [44]. Availability of imaging reports and the related images are important in these and other applications of genomics. |
| Genomic test report | For genomic medicine, the most common output is the diagnostic report. This may be provided with atomic data to the LIMS but is commonly stored as PDF or other text attachment files.<br>Diagnostic reports are generally held for 100 years or more (or 10 years with somatic samples) [45]. |
| Clinical photography | Clinical photography, including facial imaging and 3D scanning, is an area of data growth, especially with genotyping technologies. The matching of phenotypes based on this technology and genotypes is accelerating diagnosis and enriching research, particularly in the rare disease space [46]. |
| Quality control data | Like genomics, all diagnostic systems undergo quality control processes. The nature of the data collected will be determine by the diagnostic process. |
| Validation data | Data developed and/or used specifically to support laboratory validation such as NATA accreditation, validation, verification and ongoing quality assurance activities. Production, storage and sharing of such data is invaluable to laboratories. They might be as simple as NA12787, an anonymised reference set that labs can share and benchmark to, or something sophisticated like a dilution series. |

## 3.4 Administrative content

The last top-level classification includes data categories required across both clinical and genomics content. They are ‘administrative’ because they are used to control, govern or describe the actual data.

### 3.4.1 Non-clinical data

A variety of non-clinical data associated with healthcare delivery may support genomics research and healthcare economics.

| Sub-classification | Description |
|---|---|
| Demographics | EHRs universally record basic demographic data including date of birth, sex, race/ethnicity and address, which are important considerations for both research and medical genomics. |
| Activity coding | Sometimes, it may be useful to access activity coding, which is the basis of funding for public health services. This may include test costs, laboratory name and setting in support of health economics analysis. |
| Other non-clinical | Other administrative hospital data may also support analysis of health economics research (e.g. death notifications; MBS/PBS records). |

### 3.4.2 Consent

Free, prior and informed consent underpins all clinical and research activities, however, recording such consent has traditionally been in paper forms with low levels of consistency. A national approach to consent is under consideration, however, these groups describe the information that may be included. This data will ideally support dynamic consent as it becomes available.

| Sub-classification | Description |
|---|---|
| Genomic testing consent | The record of patient consent for clinical genetic or genomic testing may include:<br>• patient autonomy<br>• test purpose and process<br>• potential outcomes from the result of testing<br>• structure and potential impacts of results (patient, family, insurance, etc.)<br>• storage of genomic data<br>• use of de-identified data in reference databases |
| Genomic research consent | In addition to the above, the record of participant consent for genomic research may include:<br>• Use of de-identified data in research, especially in support of public health initiatives<br>• Use of genomic data in research, and whether the patient may be re-identified to provide information to them. |

### 3.4.3 Data governance

Strong governance of clinical and genomic data is critical to gain and retain trust of consumers in the health and research sectors. This aligns with the data management principle *DM05: Strong governance models*.

| Sub-classification | Description |
|---|---|
| Policies and procedures | Foundational to good governance are documented policies and procedures, and related supporting documents. |
| Data asset register | Understanding the data assets held by an organisation is critical to soundly governing those assets and safely and securely sharing the data in line with consent provided. |
| Data access request | If consent has been granted to share data for research or clinical purposes, it is important to be able to record all requests to access the data. This allows data custodians to understand the reasons for data access and assess them against the authorised purposes. |
| Data sharing agreement | An inter-institutional or intra-institutional agreement to share data according to certain terms. Data sharing agreements identify the parameters which govern the collection, transmission, storage, security, analysis, reuse, archiving and destruction of data. This category can include templates for data sharing agreements and the agreements themselves. Also known as data transfer agreements [47]. |
| Data management plan | A Data Management Plan typically outlines what research data will be created during a research project and how it will be created, plans for sharing and preserving the data and any restrictions that may need to be applied [48]. |
| Data access audit | All access to data needs to be logged and audited to support traceability and monitoring of compliance against data access requests. |

## 3.5 Other data type considerations

Further to the groupings described above, the following should also be considered.

### 3.5.1 Personal identifiers

In a healthcare setting, the data classifications described in this section will include personal identifiers that link data to provide healthcare services.

The retention of personal identifiers within the context of research data sets allows them to link records to other relevant data sets with the required approvals, but requires researchers to 'exercise care in handling confidential or other sensitive information used in or arising from a research project' [49].

The *National Statement on Ethical Conduct in Human Research* (2007, updated 2018) [6] no longer uses terms such as 'identifiable', 'potentially identifiable', 're-identifiable', 'non-identifiable' or 'de-identified' as descriptive categories for data or information due to ambiguities in their meanings. Re-identification and de-identification are best understood as processes that change the character of information and are only used with this meaning.

Data61 and the ONDC have published *The De-identification decision-making framework* [50] that provides guidance on dealing with personal identifiers.

If personal identifiers are removed from or not able to be associated with research data sets, the ability to link with other data sets is reduced along with the usefulness of the data.

Reasonably identifiable data is subject to Commonwealth and state/territory privacy legislation and regulations, and the entity holding the data determines what legislation applies. Legislative requirements are an important consideration in the management of genomic data and potential inconsistencies/barriers/risks will need to be addressed to support a national approach to genomic information management.

### 3.5.2 Biobanking

While biobanks and the physical specimens they manage play a broader role than for genomics, the data associated with managed samples interests the genomic sector. Like genomic data, they may hold data associated with patient demographics, patient history and phenotypes. In some jurisdictions internationally, biobanks also hold genomic data.

A finding from Queensland Genomics patient engagement is that consumers do not differentiate between their genomic sequence data and their biological samples [51].

### 3.5.3 Other 'omics data

Besides the categories listed, there are other 'omics related data that should be part of a national approach as it develops:

- epigenetics
- metagenomics
- proteomics
- a range of biochemical, haematological, immunological and other assays

While outside the scope of this Blueprint, it is envisaged that the data categorisation framework could be extended to include categories to cover these evolving technologies.

# 4 Considerations for designing a framework

While international perspectives can assist in the design of a blueprint for the management of genomic data and information in Australia, there are specific drivers and constraints important to our nation. This section explores some of these issues to provide a background to the influences that have led to the solutions described in this document.

## 4.1 Differences between research and medical genomics

One of the most fundamental differences between the delivery of healthcare (using genomics or otherwise) and genomics research is the source of funding.

Healthcare delivery in Australia is funded predominantly by the Australian Government (41% in 2016-17) and the state and territory governments (27% in 2016-17) through the Medicare system and activity-based funding [51]. However, at the time of writing, only a few genomic tests are funded via Medicare, with states/territories funding the majority of more complex (and hence costly) genomic tests.

Research funding is largely delivery through competitive funding via the ARC, NHMRC, the Medical Research Future Fund (MRFF), universities and private bodies [52].

Variation in funding and constrained resources will influence how solutions for genomic information management can be implemented and funded. Genomic data infrastructure design should reflect intended data use, funding source and consider the associated limitations or constraints.

Besides funding, other differences between research and medical genomics include:

- the estimated data size by 2025
- the rationale for exome or genome sequencing
- the timelines required for sequencing and analysis
- the number of sequences required
- whether data is routinely shared for further research
- data sharing mechanisms
- the legal and regulatory frameworks required to support data sharing
- the languages used to prepare agreements and for discussion.

## 4.2 The changing nature of genomic data

Current genomic research is frequently reliant on comparatively small datasets, depends upon research funding to collate or access the data, and often utilises infrastructure established for a specific project.

Paradoxically, the increase in exome and genome sequencing worldwide is likely to generate almost 60 million genomes [53], of which about 80 per cent will be generated in clinical settings [54]. Availability of genomic data at these volumes could have a significant positive impact for genomic research, subject to appropriate consent and ethics approval.

A national approach for genomic data management in Australia would enable organisations to leverage this shift in data sources between genomic medicine and genomic research. Creating a ‘virtuous circle’ of data derived through genomic medicine supporting genomic researchers would deliver outcomes that result in better clinical decision support, health service improvements and better education and training opportunities.

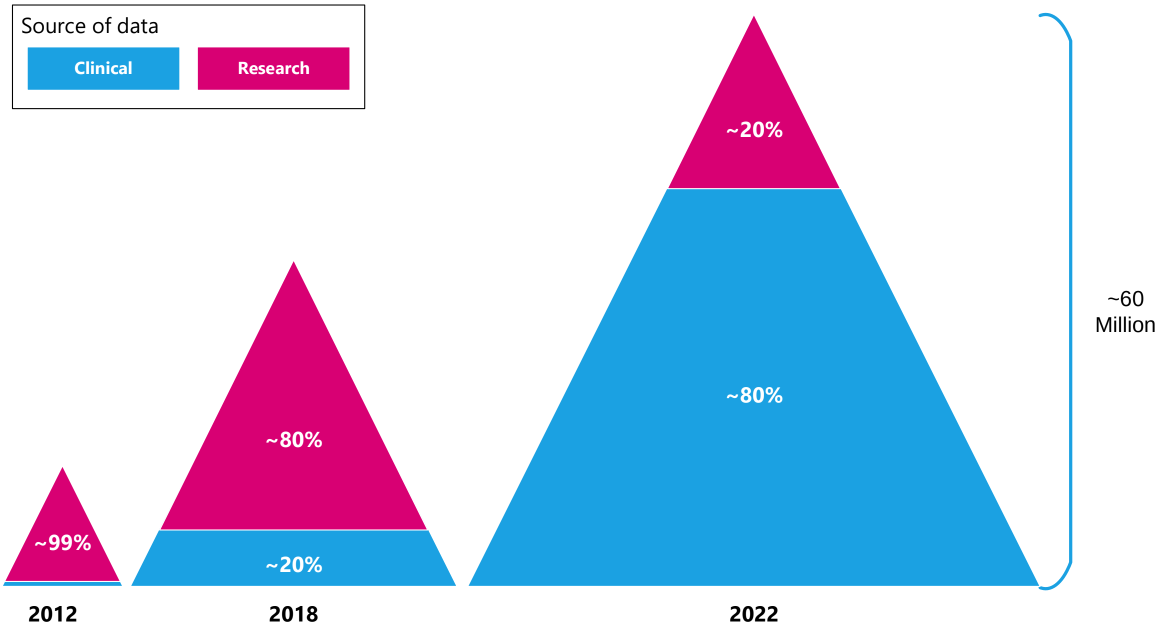


*Figure 4: Worldwide growth in genomics and the split between clinical and research sources*

This would support the delivery of a learning health care system which is:

> *...one in which science, informatics, incentives, and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the care process, patients and families active participants in all elements, and new knowledge captured as an integral by-product of the care experience* [55].

## 4.3 Bioinformatics analysis pipelines

Genomics capability for both research and clinical delivery rely on bioinformatic analysis. While the approach described in Appendix A is common to all pipelines, many variations exist between specific pipeline implementations, including:

- Pipelines for research are developed in a more agile manner whereas those used in NATA accredited diagnostic settings exhibit a slower rate of changes due to the regulatory aspects of their use.
- The key requirements for diagnostic pipelines can differ between laboratories, such as resources for the establishment and maintenance of internal systems versus utilising commercially provided pipelines.
- Pipelines may be focused on specific clinical/research areas of interest (e.g. neurological disease versus cancer diagnosis) meaning that a single pipeline may not apply to all uses.
- In the innovation space of genomic testing, there may be opportunities for cross laboratory sharing of pipelines, however, this would likely be affected by perceived competition and intellectual property considerations.
- The output of the pipeline needs to be tailored to the intended use by the curation or reporting process.

Sharing of pipeline capabilities may seem superficially attractive but is often limited by the highly diverse nature of the computational infrastructure, data policies, technical proficiencies and unique requirements of different laboratories. The overall effort to harmonise these elements can be greater than the actual effort of implementing and validating an analysis pipeline in-house. While sharing of concrete implementations of pipelines can be challenging, sharing at the conceptual and knowledge exchange level can be valuable.

Despite this, research pipelines are often shared and several common frameworks allow pipelines to be described in a common manner for sharing. These still face challenges posed by varying infrastructure and requirements. Cross platform workflow engines are easing issues, as is cloud computing by providing common compute infrastructure.

There is opportunity for the research community to be encouraged (and facilitated) to continue sharing pipelines using containers, common workflow languages and workflow execution engines. However, technical and quality standards need to be consistent to allow this.

The reference genome is the common reference point for all genomic analysis, yet its use is reportedly not standardised, and adoption of the most recent version (GRCh38) is not yet widespread.

## 4.4 The role of self-describing repositories

Australia has a wealth of research and health services organisations with an interest in genomics. When considering the impact of this ecosystem on any solution, one must be cognisant of a variety of drivers:

- Commonwealth vs state/territory objectives and responsibilities
- public vs private institutions
- health service delivery vs research.

To effectively bring these disparate capabilities together, repositories should be self-describing in a computational manner, identifying data available, the consent for the uses of that data, and the capabilities that exist.

A registry of such repositories would be a useful addition. However, a registry's value would be limited without standardisation, as registries of non-curated, non-standardised data are of limited use. Standardisation at the right level needs to be achieved despite the diverse nature of data stores.

Ontologies for describing registries can be obtained from GA4GH, European Genome-phenome Archive (EGA) and others. Some of these standards are described in Section 6.

## 4.5 Developing a value framework

The principles in Section 2 include reference to delivering benefits or value to stakeholders. It is useful to consider how those benefits may be defined. In a broad sense, benefits to consumers accrue from benefits to the health system resulting from greater efficiencies. More specifically, concepts of value or benefit in genomics have been described by the CEO of the GA4GH, Peter Goodhand, as including [56]:

- **Diagnostic benefit:** The identification of pathogenic or likely pathogenic variants in known disease genes.
- **Clinical benefit:** Changes in the medical or surgical management of patients because of the diagnosis being made. For example, the assignment of therapies (therapeutic benefit) or improvements in the management of patients in the absence of therapy assignment (management benefit).
- **Clinical trial benefit:** Changes related to the improvement of clinical trial operations.
- **Personal benefit:** The presence of non-clinical outcomes important from a personal viewpoint to a person with a genetic disease or who is affected by a genetic disease. These outcomes may relate to the intrinsic value of information, the knowledge about the condition and the opportunity to plan for the family or the future.

These classes of benefits are also reflected in a report by the World Economic Forum [57]. These benefits are focused on direct or immediate clinical benefit, and this is understandable in the context of health service delivery. However, basic research importantly provides other indirect benefits in areas such as basic biology and preclinical domains that can lead to later developments that deliver clinically focused benefits [58].

## 4.6 High-level requirements

This Blueprint has identified several high-level requirements, which are summarised below:

- **Standards-based interoperability –** National agreements on a standards-based approach to the management of genomic data management will allow interoperability between systems, be they research or health delivery services. As Australia exists within an ecosystem that includes significant global repositories, these standards also need to consider the international use of standards.
- **Interdependence between research and health delivery** – Genomic medicine is influenced by the discoveries made by the research community, while researchers can benefit from access to clinically derived genomic data. However, both communities have differing drivers and requirements for systems to support their work. Any national approach needs to consider both communities to the benefit of both.
- **Limited decentralisation of genomic medicine data repositories** – The Australian health system is largely delivered through the state and territory health departments, who operate under their own legislation and regulation. Many (but not all) jurisdictions are developing their own solutions to manage genomic data in support of genomic medicine, and a single centralised approach likely would not meet the needs of these groups, especially considering existing jurisdictional legislation and regulation. However, standards-based development will promote interoperability between them and support a federated approach to accessing genomic medicine repositories.
- **Researchers need access to genomic data** – Australian researchers operate on comparatively small data sets when compared to the potential data sets generated by health service delivery. Often this data is sourced from international repositories and may not therefore represent the diversity of Australia’s population. Access to the volume of genomic, clinical and phenotype data generated in clinical settings may have significant benefits for the research community. However, individuals providing this data need to know the implications of sharing their data and what it is going to be used for, including any potential secondary uses for commercial purposes and potential disclosures.
- **National approach to research capabilities** – A national coordinated approach to genomic research capabilities would allow for individual innovation within a strategic national plan for genomic data management. While not all research repositories can be combined, there is value in establishing a few standards-based genomic research repositories to reduce the cost of creating multiple repositories. This will reduce the cost of duplicate data storage and potentially minimise data transfer requirements for compute capabilities.
- **Address specific needs of Aboriginal and Torres Strait Islander peoples and others** – Aboriginal and Torres Strait Islander peoples have specific needs regarding the management of data and the benefits derived from that data. When examined, these needs may reflect a ‘gold standard’ to treat data for all Australians. Other priority cohorts exist within the Australian healthcare sector, each of which may have specific needs regarding the management of their data and the benefits derived from that data.
- **Improvements in privacy, consent and security** – A national standards-based approach based on a strong governance framework will support the data required to allow consistent application of privacy controls, consent mechanisms and security profiles. Social licence for using genomic data will be gained and maintained by delivering on the privacy, consent and security expectations of the community.

- **Consent mechanisms** – Improvements in managing consent, including standards-based ontologies to support consent capture and the potential use of dynamic consent are needed to support strong governance and build community trust genomics in Australian.
- **Building for the future** – While the focus of this Blueprint is on genomic data, potential value will be lost if other 'omic data beyond genomics is not considered when building solutions. As a rapidly developing science with new technologies emerging regularly, the Blueprint needs to be open and maintained to leverage the design for additional data types as they become available.

## 4.7 Non-human genomics

While the focus of this Blueprint is on human genomics, non-human genomics play a significant role in health service delivery and health research. Examples of organisations working in this space include:

- The National Microbial Genomics Framework 2019-2022 is the first national strategic document for microbial genomics in Australia. It provides a nationally consistent and strategic view for integrating microbial genomics in the Australian public health system, and for identifying microbial genomics policy issues and challenges [59].
- The Australian Infectious Diseases Research Centre undertakes research on parasite, viral, bacterial and fungal infectious diseases, including the use of clinical genomics [60].
- NSW Health Pathology are using genomics to better understand the origin of pathogen-based outbreaks and how infectious diseases spread and working to better understand how bacteria like staphylococcus aureus (Staph) are becoming resistant to antibiotics [61].
- Research at the Doherty Institute includes the study of antibiotic resistance, microbial pathogenesis and HIV evolution, the role of epigenetics in T-cell developmental biology and the human genetic susceptibility to tuberculosis and typhoid fever [62].
- In addition to the research being conducted by the Doherty, the Microbiological Diagnostic Unit Public Health Laboratory (MDU PHL) conducts NATA accredited clinical testing for pathogen genomics. The unit provides a service for pathogen 'tracing' in addition to pathogen detection [63].

While there are differences in how non-human genomic data is managed (such as reduced need for privacy constraints), human and non-human data sets share many of the same information management approaches, and consideration of non-human genome requirements for data repositories may provide opportunities for later benefits.

This is particularly relevant at this time given the worldwide COVID-19 pandemic, in which genomics has an important role in tracing infections and development of both treatments and preventive vaccines.

# 5 Proposed logical architecture

Several leading architecture approaches are based on the concept of three levels of models: Conceptual, Logical and Physical. These can be seen in framework such as the Zachman Framework [64], The Open Group Architecture Framework (TOGAF) [65] and the Archimate Modelling Language [66]. Logical models describe how a solution will work in an abstract manner, in terms of the functions it performs, and the nature of the data/information is processes. They support communications in a tangible way with enough detail to allow meaningful discussion, but without the specific constraints of any one implementation (or physical architecture) [67].

This document has outlined a set of principles that will guide implementations and defined a genomic data categorisation framework to describe the data and information to be managed. The logical architectures that follow describe a national approach to the management of genomic data and information, in line with these principles and data categorisations, and informed by national and international examples that exist or are being implemented.

It is important to also understand that the architectures proposed allow for variation in implementation and the reality that the genomic sector includes organisations at different levels of digital maturity and progress towards genomic technology and use. This is in line with principle *DM07: Pragmatism*.

## 5.1 Data and compute capabilities

Genomics has brought, and will continue to bring, new meaning to the expression ‘big data’. It has been described as a ‘four-headed beast’, referring to the four issues of data acquisition, storage, distribution and analysis [68]. The need to acquire, store and move data at genomic scales has increasingly led to using cloud-based technologies.

While historically hesitant to use cloud storage, governments and health services are increasingly leveraging these technologies as they have become more secure and mainstream. The Australian Government’s Digital Transformation Agency (DTA) has adopted a cloud-first approach to government services [69], and many state and territory governments are following similar strategies.

Cloud capabilities include not only data storage but also data analysis (known as ‘compute’) capabilities. Leveraging high-performance computing has been critical to analyse the data at scale represented by genomic data. Between commercial cloud computing and research capabilities provided by organisations such as the National Computational Infrastructure (NCI) [70] and the National Collaborative Research Infrastructure Strategy (NCRIS) [71], the capability to match big data with big compute is critical to delivering genomics at scale in Australia.

While using these technologies is largely an implementation level consideration, these capabilities should be part of the consideration of any national logical architecture.

## 5.2 A logical model for genomic medicine

As the potential source of increased volumes of genomic data to be managed, Figure 5 provides a high-level model for a genomic medicine environment. This is focused on the clinical application of genomic technologies and knowledge.

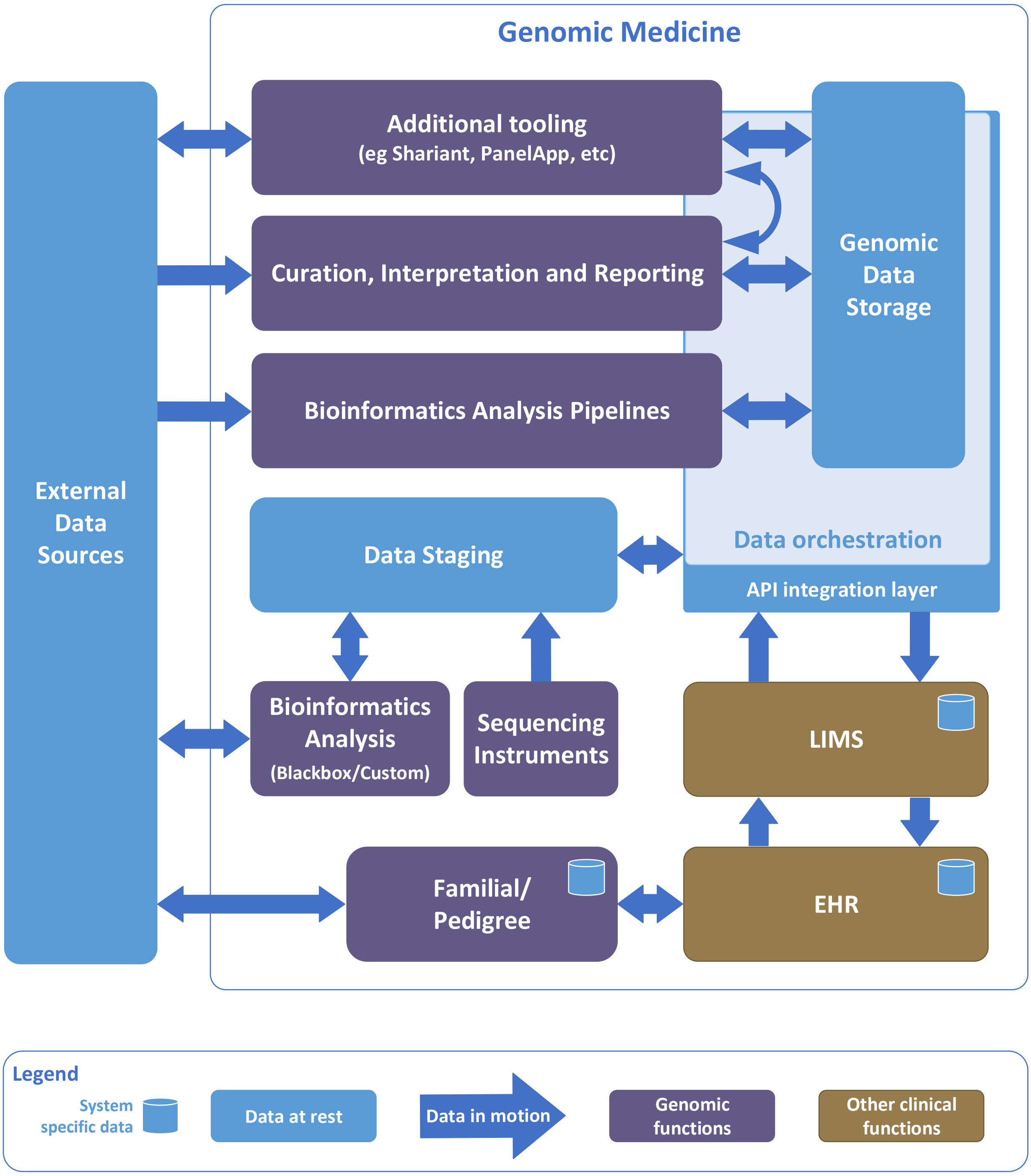


*Figure 5: A logical model of genomic medicine data management*

The architecture above has several functional components, including:

- **Existing EHR and LIMS systems** that operate within typical health services. Ideally these systems are integrated and digital, providing the link between clinical systems and the diagnostic (or predictive) use of genomics.
- **Systems to support familial and pedigree** data within genetic counselling services, ideally integrated with the EHRs within a health service for the exchange of clinical and patient administration data.
- **Genomic sequencing technologies** including optional bioinformatics analysis capabilities. These generate the genomic sequencing data for later analysis.
- Further **bioinformatics analysis capabilities** (if required) and capabilities to support the **annotation, classification, curation and interpretation** of genomic data.
- Support for the **exchange of genomic knowledge** such as reference genomes and the classification of variants (to name only two). A variety of additional tools support such exchanges.

The architecture also recognises several repositories of data, including:

- **Local system-focused repositories** that hold data for specific functions, including EHRs and LIMS. Data from these systems will need to flow through the ecosystem for later use in genomic processes and may be held locally on using cloud storage.
- **Data may be staged** using local or cloud storage once generated from the sequencing instruments for initial processing and access by the core genomic data systems.
- Storage of the **core genomic data** used by the genomic processes. These are shown logically grouped but may comprise multiple individual databases, file stores and other repository formats. Critical to national adoption, they will require consistent **standards-based interfaces (APIs)** to other systems and will need to provide **data orchestration** within and between repositories and systems. While local storage may be used, based on feedback from clinicians and researchers in Australia and internationally, there is an increasing preference for the core genomic data store to be cloud-based.
- An important part of the data landscape is **external data repositories**, which will be the source of critical information for genomic activities. These external data repositories are also the target of data flows leaving a health service to share genomic knowledge with others nationally and internationally.

## 5.3 A logical model for genomic research

While there may appear to be similarities between genomic medicine and genomic research, key differences need to be considered. Figure 6 illustrates a logical architecture for research genomics.

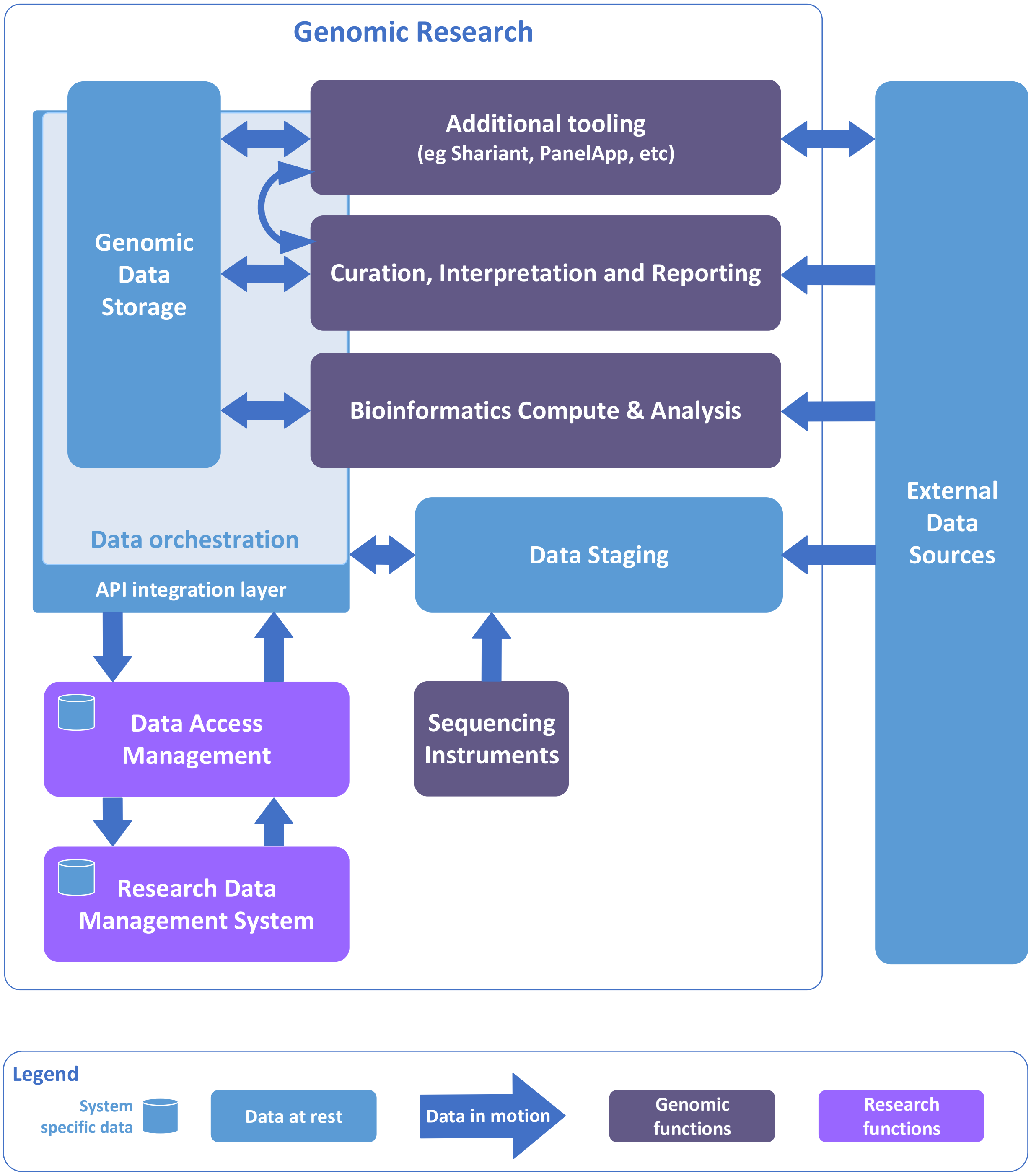


*Figure 6: A logical model of research genomics data management*

The architecture above has several functional components, including:

- **Research data management systems** (RDMS) that operate within a research organisation to manage the data and processes of a research project. They may hold data locally or may manage externally hosted data.

- **Data access management** (DAM) systems that support the management of genomic and clinical data, including access requests and fulfilment.
- **Genomic sequencing technologies** used to generate the genomic sequencing data for later analysis.
- **Bioinformatics compute capabilities** to undertake genomic analysis and broader research analysis of the genomic data.
- Capabilities to support the **annotation, classification, curation and interpretation** of genomic data.
- Support for the **exchange of genomic knowledge** such as reference genomes and the classification of variants (to name only two).

The architecture also recognises several repositories of data, including:

- **Local system-focused repositories** that hold data for specific functions, including the RDMS and DAM. These are especially important for managing the non-genomic aspects of a research project and may be held locally on using cloud storage.
- **Data may be staged** using local or cloud storage once generated from the sequencing instruments for access by the core genomic data systems.
- Storage of the **core genomic data** used by the research processes. These are shown logically grouped but may comprise multiple individual databases, file stores and other repository formats. Critical to national adoption, they will require consistent **standards-based interfaces (APIs)** to other systems and will need to provide **data orchestration** within and between repositories and systems. While local storage may be used, there is a general preference for the core genomic data store to be cloud-based.
- An important part of the data landscape is **external data repositories**, which will be the source of critical information for genomic activities. These externals data repositories are also the target of data flows leaving a research organisation to share genomic knowledge with others nationally and internationally.

## 5.4 Patterns of interactions

Given the logical architectures for genomic medicine and genomic research described in the previous sections, several interactions patterns can be identified. These patterns describe how systems can pass data and queries. Not all patterns operate at the same level of maturity, and this allows for an evolving ecosystem.

### 5.4.1 Point-to-point requests for data

This pattern is typical of traditional approaches for data from clinical sources, but also applies to request between research groups sometimes. The process is:

- The data requestor defines the data sets required and passes the request to the Data Provider.
- The request is assessed based on consent associated with the data and planned usage by the data requestor. A data access group may be involved to assess or approve requests.
- The data provider collates the data and passes it to the data requestor.
- The data requestor is responsible for computational analysis.

Note, where data linkage is required, the process will be context dependent. Within research data sets, the data requestor may be responsible for undertaking data linkages. With identified government-held data, the data requestor would usually only be responsible for seeking approval for data linkage to be undertaken. The data linkage itself would usually be undertaken by either the Data Custodian (e.g. the creator of the clinical data) or by another organisation that is approved by the Data Custodian to undertake the linkage on their behalf (e.g. the AIHW) because it requires the management and use of personal identifiers.

While some aspects of this pattern can be automated, it is commonly a manual process requiring resources to be invested by the Data Provider to support the request.

### 5.4.2 Synchronisation of data sets

This pattern supports data availability across organisations, improving redundancy and reducing latency for processing. This model is most applicable in a research context but may also be implemented in clinical environments. The federated node model used by the EGA is an example of this pattern [72].

This pattern requires:

- Both parties to have agreed a set of standards to be applied for the storage and transport of the data.
- Data governance practices regarding standards for security, privacy and other factors for protecting the data.
- Agreed mechanism for the control of access to data based on consent associated with the data and planned usage by the recipient. A data access group may be involved to assess or approve requests.

### 5.4.3 Remote query

This pattern provides a more sophisticated approach to provision of data to Data Requestors. Under this model:

- The data requestor identifies sources of the data under consideration. This may occur through published catalogues or knowledge by the researchers.
- The data requestor defines the query to be executed over the data sets and passes the request to the Data Provider.
- The request is assessed based on consent associated with the data and planned usage by the data requestor. A data access group may be involved to assess or approve requests.
- The Data Provider executes the query over the data repository they hold and passes the results to the data requestor. This requires the Data Provider to provide compute capability suitable for execution of the request.
- The data requestor is responsible for computational analysis.

As with the point-to-point request, data linkage is dependent on the context.

This pattern requires agreements on computational capabilities and data transport mechanisms and is like the Researcher Environment approach provided by Genomics England [73].

### 5.4.4 Federated queries

This pattern extends on the remote query model by providing the request to multiple data providers. While procedurally like the remote query pattern, this pattern is suited where more than one Data Provider is involved and requires agreement across several providers of data to avoid complexities involved in different connection and exchange arrangements.

### 5.4.5 Self-describing repositories

This pattern supports repositories that can provide data (and optionally compute) capability that support other patterns described above by providing a documented capabilities interface that allows their capabilities to be queried and used by other participants.

## 5.5 Increasing genomic data interoperability

Not all jurisdictions or research organisations are starting with the same capabilities or priorities, nor can they all move with the same speed. A mix of technologies and capabilities will continue to be the norm over the short to midterm, with those organisations with capability and capacity leading development.

The following are examples of possible stages in the development towards a mature genomic data ecosystem for Australia. In these diagrams, integrations are

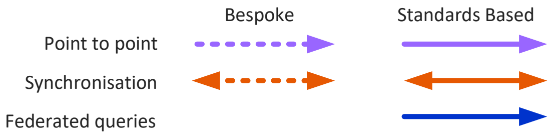


*Figure 7: Legend for following diagrams*

### 5.5.1 Current state

Figure 8 illustrates the current states of genomic data capabilities, which largely comprise point-to-point connectivity with standards being used inconsistently. Some standards-based sharing occurs, but this needs to be extended to a broader range of sites and data types.

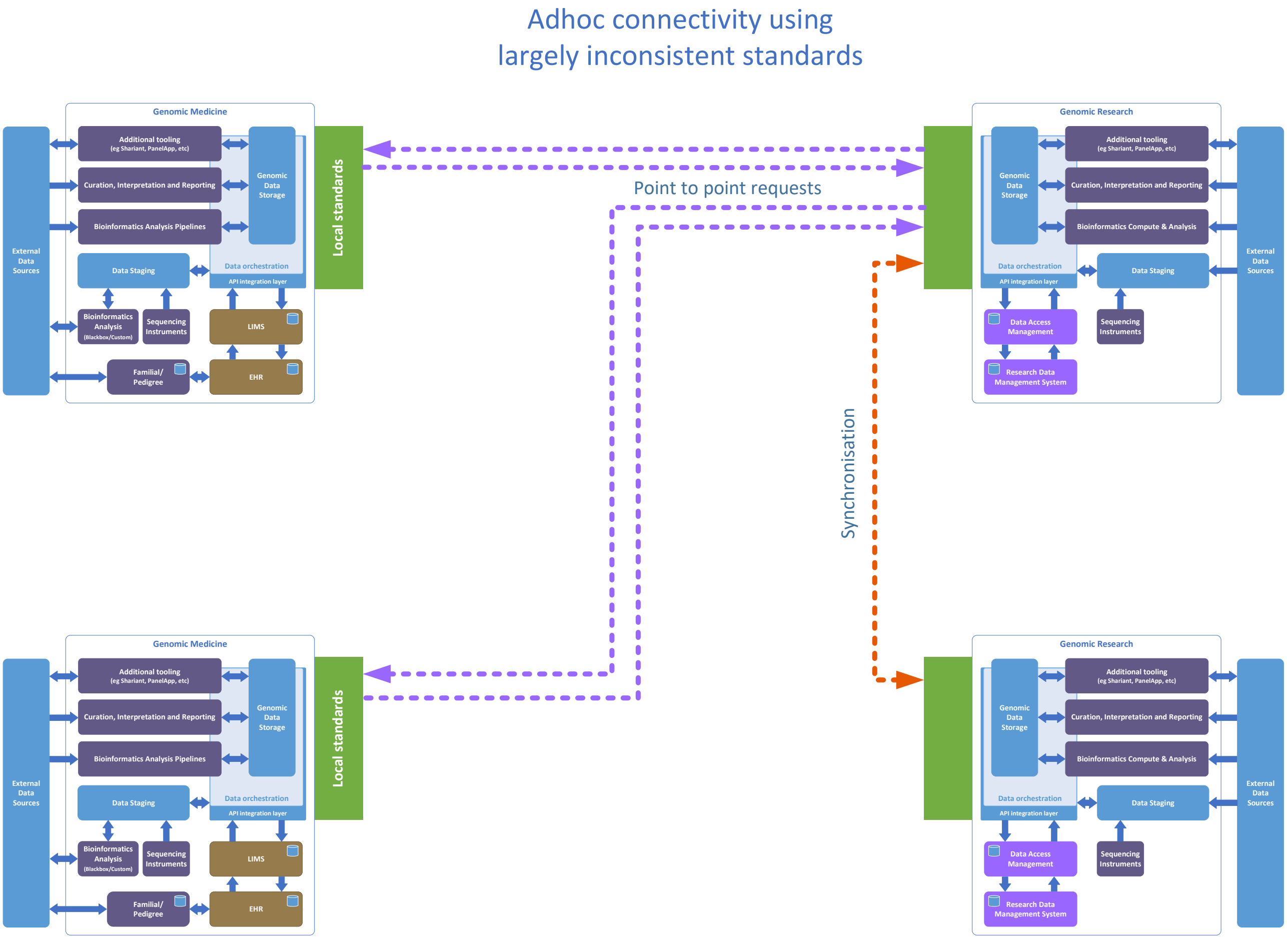


*Figure 8: Example of typical current state capabilities*

This model lacks a nationally coordinated ecosystem: organisations operate within their own constraints. While these groups are exchanging data and ideas, scalability and flexibility are common issues. Point-to-point interfaces are frequently bespoke and need to be negotiated and established on a case by case basis, making scalability problematic. As they can be negotiated based on locally agreed standards rather than national standards, interoperability between negotiation groups relies on whether they have adopted the same standards (or the same implementation/profile of the same standard). Adoption of nationally agreed standards allows groups to exchange data beyond the partners originally envisaged.

Some limited synchronisation of data sets between research groups may exist, providing limited sharing of data. Research groups are largely leading the trials of connectivity, with clinical systems not yet focused on data exchange as a core service.

### 5.5.2 Establishing a standards-based genomic data ecosystem

As maturity increased and a nationally coordinated ecosystem evolves, developing agreed standards promotes better integration, as illustrated in Figure 9

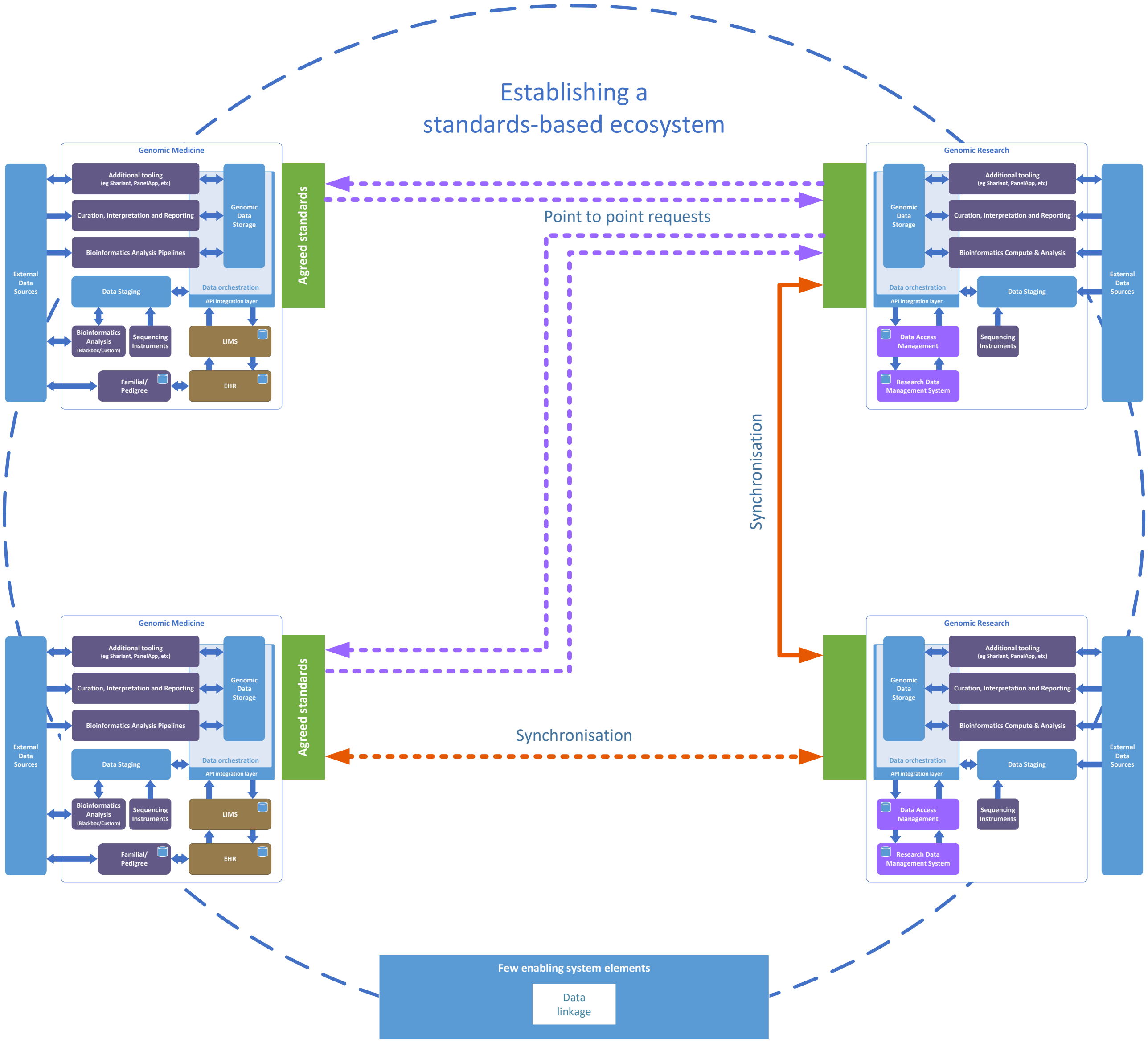


*Figure 9: Establishing an ecosystem based on standards and with national collaboration*

National coordination promotes collaboration between jurisdictions and research organisations to agree a set of national standards for interfaces. Another outcome is the development of consistent requirements so aligned solutions emerge. This also allows for enabling systems to support the initial stages of integration between systems.

Exchanges are still largely manually facilitated, but governance arrangements can be established, especially within jurisdictions to support the provision of genomic data.

### 5.5.3 Standards-based integration

The next stage of maturity supports a suite of nationally agreed standards, compute capability available to support federated queries across multiple data providers and more systems that enable integration and data use across the ecosystem. This is illustrated in Figure 10.

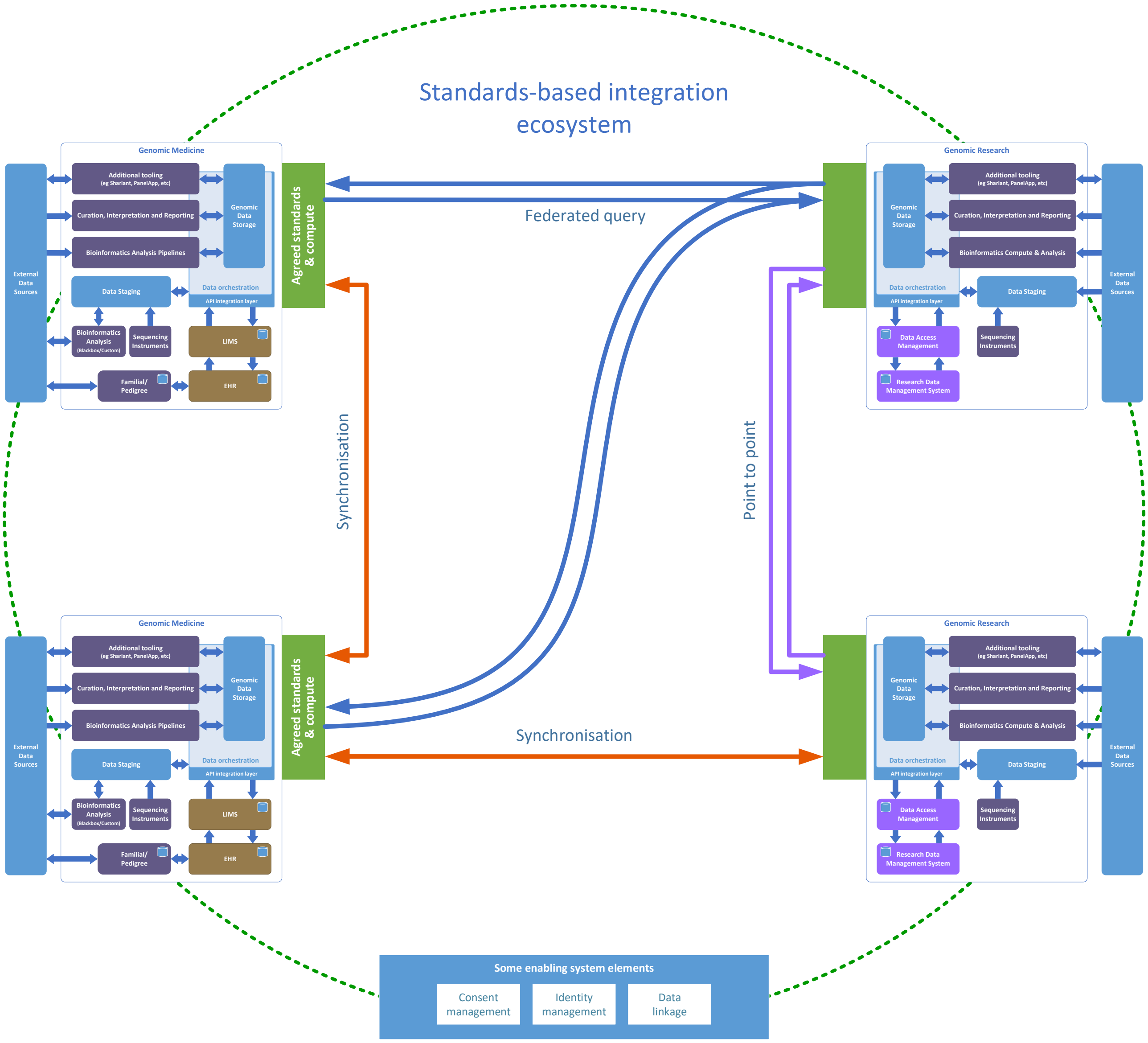

Standards-based integration ecosystem
Genomic Medicine
External Data Sources
Additional tooling (eg Shariant, PanelApp, etc)
Curation, Interpretation and Reporting
Bioinformatics Analysis Pipelines
Genomic Data Storage
Data Staging
Data orchestration
API integration layer
Bioinformatics Analysis (Blackbox/Custom)
Sequencing Instruments
LIMS
Familial/ Pedigree
EHR
Agreed standards & compute
Federated query
Genomic Research
Genomic Data Storage
Additional tooling (eg Shariant, PanelApp, etc)
Curation, Interpretation and Reporting
Bioinformatics Compute & Analysis
Data orchestration
API integration layer
Data Staging
External Data Sources
Data Access Management
Sequencing Instruments
Research Data Management System
Synchronisation
Point to point
Genomic Medicine
External Data Sources
Additional tooling (eg Shariant, PanelApp, etc)
Curation, Interpretation and Reporting
Bioinformatics Analysis Pipelines
Genomic Data Storage
Data Staging
Data orchestration
API integration layer
Bioinformatics Analysis (Blackbox/Custom)
Sequencing Instruments
LIMS
Familial/ Pedigree
EHR
Agreed standards & compute
Synchronisation
Genomic Research
Genomic Data Storage
Additional tooling (eg Shariant, PanelApp, etc)
Curation, Interpretation and Reporting
Bioinformatics Compute & Analysis
Data orchestration
API integration layer
Data Staging
External Data Sources
Data Access Management
Sequencing Instruments
Research Data Management System
Some enabling system elements
Consent management
Identity management
Data linkage

*Figure 10: The third level of maturity sees more enabling systems and compute capabilities*

Federated queries reduce the need to exchange large data sets but requires compute capabilities close to the data. Nationally supported capabilities such as identity management are leveraged to support the use of data. A nationally agreed approach to consent allows for dynamic determination of whether access is appropriate.

### 5.5.4 Standards-based interoperability-enabled ecosystem

At this highest level of maturity, the genomics data ecosystem has moved from a focus on integration to an ecosystem based on interoperability standards.

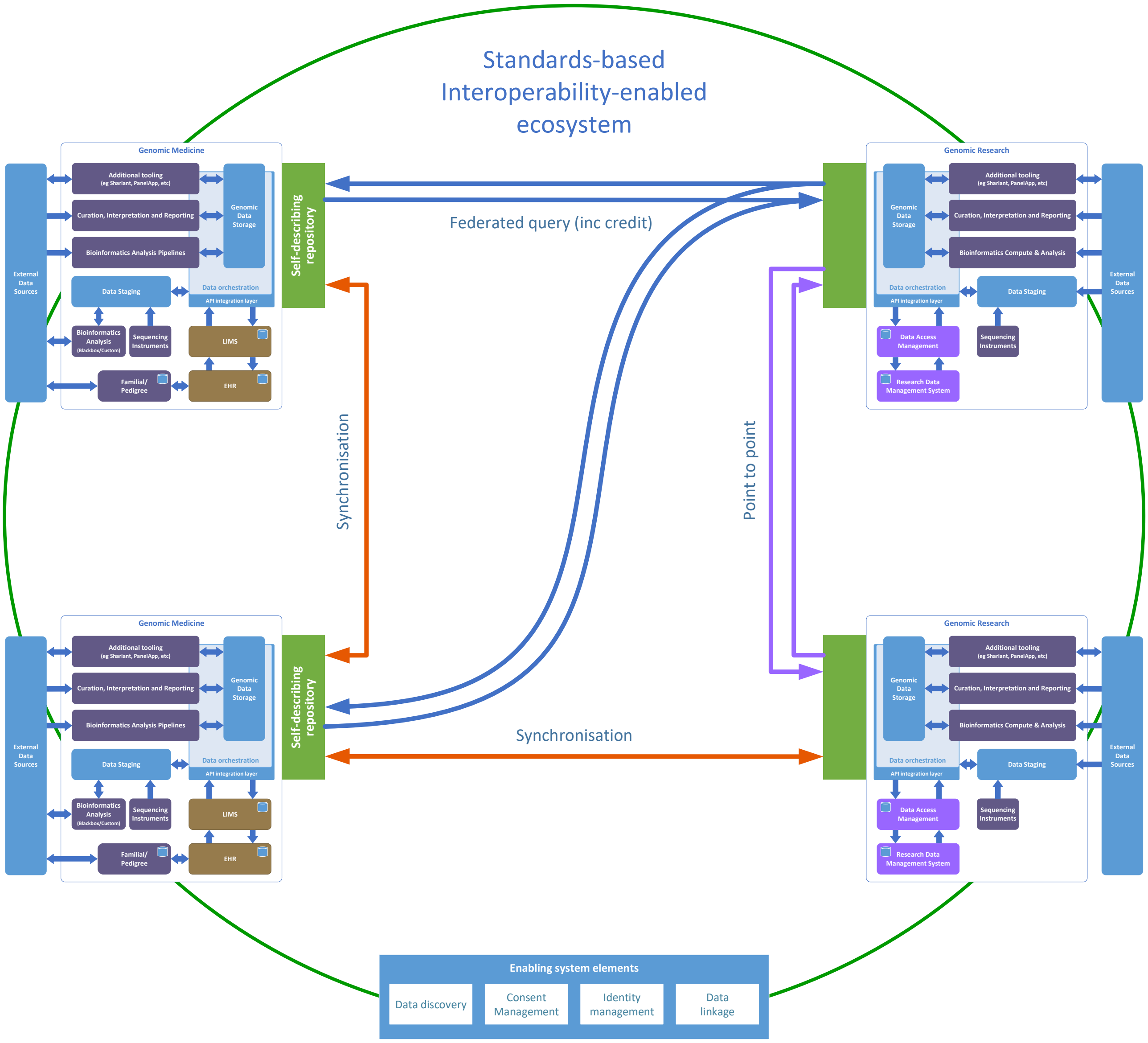

Standards-based Interoperability-enabled ecosystem
Genomic Medicine
External Data Sources
Additional tooling (eg Shariant, PanelApp, etc)
Curation, Interpretation and Reporting
Bioinformatics Analysis Pipelines
Genomic Data Storage
Data Staging
Data orchestration
API integration layer
Bioinformatics Analysis (Blackbox/Custom)
Sequencing Instruments
LIMS
Familial/ Pedigree
EHR
Self-describing repository
Genomic Research
Genomic Data Storage
Additional tooling (eg Shariant, PanelApp, etc)
Curation, Interpretation and Reporting
Bioinformatics Compute & Analysis
Data orchestration
API integration layer
Data Staging
Data Access Management
Sequencing Instruments
Research Data Management System
External Data Sources
Federated query (inc credit)
Synchronisation
Point to point
Genomic Medicine
External Data Sources
Additional tooling (eg Shariant, PanelApp, etc)
Curation, Interpretation and Reporting
Bioinformatics Analysis Pipelines
Genomic Data Storage
Data Staging
Data orchestration
API integration layer
Bioinformatics Analysis (Blackbox/Custom)
Sequencing Instruments
LIMS
Familial/ Pedigree
EHR
Self-describing repository
Synchronisation
Genomic Research
Genomic Data Storage
Additional tooling (eg Shariant, PanelApp, etc)
Curation, Interpretation and Reporting
Bioinformatics Compute & Analysis
Data orchestration
API integration layer
Data Staging
Data Access Management
Sequencing Instruments
Research Data Management System
External Data Sources
Enabling system elements
Data discovery
Consent Management
Identity management
Data linkage

*Figure 11: A mature 'marketplace' ecosystem using standards-based interoperability*

This stage also introduces the concept of a 'marketplace'. This should not be confused with the commercialisation of genomic data, but rather a basis of exchange between data providers and data consumers that supports an equitable distribution of the costs associated with operating such an ecosystem.

Data providers with sufficient capabilities have established self-describing repositories that allow for a national approach to data discovery and capability availability.

## 5.6 A draft roadmap for implementation

This Blueprint lays out a possible future state logical architecture to support genomic information management for both genomic medicine and genomic research. However, healthcare is a complex environment subject to many complicating factors, including funding, jurisdictional priorities and continuing advancements in technology. It would be disingenuous to assume that the approaches outlined here could be implemented as a single leap.

Instead, a staged approach is suggested to allow for the variations in experience and capacity across the jurisdictions and research institutions to adopt such change. Horizons are described below including indicative activities that would support eventual implementation of a national approach to genomic information management. Such an approach would reflect a learning health system 'in which science, informatics, incentives and culture are aligned for continuous improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience' [10].

These phases are visualised in per Figure 12, and labelled conceptually to orient initial activities that leverage and plan, build on identified foundations and translate over time to an operational ecosystem.

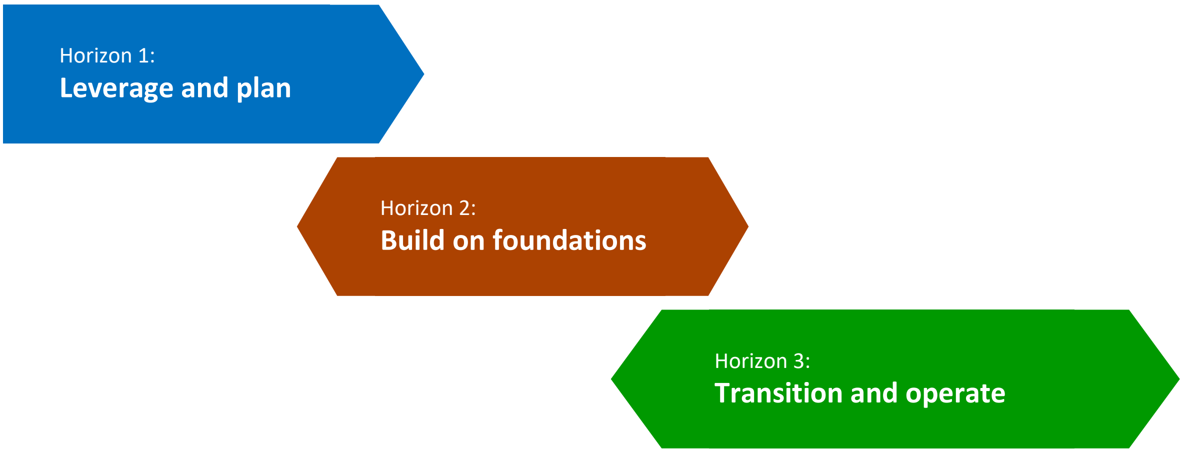


*Figure 12: A phased approach to national genomic information management*

For each horizon, four broad areas of possible activities are identified:

- **Governance activities** that will support coordination of context specific activities elsewhere and encompassing collaborative co-designed solutions acknowledging policy drivers and market forces.
- **Medical genomic activities** that will maximise the availability of genomic data and incorporate the existing benefits available from research.
- **Genomic research activities** that will leverage the available clinical genomic data to drive discoveries.
- **Infrastructural activities** required to support the delivery of the outcomes in the other two areas.

The activities that are described in the following sections are indicative and will need to be validated by appropriate funding bodies and governance groups.

### 5.6.1 Horizon 1 – Leverage and plan

The first conceptual horizon would start now and acknowledges the current landscape. Existing programs in both healthcare and research arenas are delivering outcomes and these need to be supported with the outcomes leveraged to provide learnings for the broader sector. Activities should be driven by those jurisdictions and organisations currently leading implementation but should engage with others to ensure that learnings are shared, and such groups can factor these learnings into their planning. It should be a shared future to realise the true vision for a national approach to genomics information management.

| Area | Indicative 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.<br>• 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.<br>• Consideration should be given to whether a national or jurisdictional **Data Custodian/Steward** are required to provide oversight of how data is managed, accessed and shared.<br>• **Confirm or amend the roadmap elements of this** national approach to genomic information management.<br>• **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.<br>• 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).<br>• 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.<br>• Agree an **interoperability capability model** that allows for organisational self-assessment in support of planning and funding decisions.<br>• Agree **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.<br>• Establish a cross-jurisdictional working group to **standardise access to familial and pedigree data** for clinical purposes.<br>• Establish **national agreements for genomic data sharing for clinical purposes**, leveraging existing clinical data sharing agreements working with private and public providers.<br>• Establish an **agreed approach to capture or mapping of phenotype data** within clinical systems to support genomic diagnosis, predictions and research.<br>• Support **ongoing operation and expansion of variant curation repositories and tools** (e.g. Shariant) to support genomic medicine. |

| Area | Indicative activities |
|---|---|
| **Genomics research** | • Establish **national agreements for genomic data sharing for research**, leveraging existing data sharing agreements.<br>• Establish a **national research consent mechanism** for genomic data utilising strong credentialing for participants with dynamic approaches to ongoing engagement<br>• Continue **trials of research data sharing** with leading clinical groups, leveraging existing genomic programs, to establish baselines and learnings for later implementations.<br>• Establish **national arrangements** to consider Australia's access to and use of global genomics data assets, our dependencies and role on the world stage. |
| **Infrastructure** | • 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[3] (ideally cross-jurisdictional) as well as existing and emerging jurisdictional solutions[4]. A study of clinical/research partnerships[5] would be beneficial.<br>• Develop a **standards-based, interoperable approach to cloud adoption** to support storage and retrieval of genomic data in both medical and research domains.<br>• Work with international groups (such as GA4GH) to agree **standards for self-describing repositories** that can identify their content and capabilities.<br>• 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.<br>• Establish **standards for federated query** across genomic data repositories.<br>• Work with international groups to agree **standards for international research data sharing**. |

### 5.6.2 Horizon 2 – Build on foundations

The second conceptual horizon should build upon the existing capabilities and those created in Horizon 1 to create or enhance national infrastructure and capabilities to work with it. This phase should see the roll out of the capabilities at jurisdictional and national levels to support a national genomic information network operational framework.

[3] An example of such a partnership is work done between QIMR Berghofer and the Garvan Institute.

[4] The GenoVic solution developed by Melbourne Genomics is an example

[5] Partnerships between PathWest and Harry Perkins or Canberra Clinical Genomics and the Centre for Personalised Immunology are examples

| Area | Indicative activities |
|---|---|
| **Governance** | • Develop a **national genomic information network operational framework** to guide operational activities supporting national data exchange. This will provide the ongoing operational processes required to manage a cross-organisational network.<br>• Trial operations of the **national genomic information network operational framework**.<br>• **Monitor emerging technologies** that will influence the national approach to genomic information management.<br>• **Review the outcomes from Horizon 1** and update the roadmap to reflect changes in priorities and emerging technologies, and correct and refine governance encouraging agility and dynamism in the system.<br>• Establish the **measures for innovation and value** assessment, to provide the ability to understand and measure the value or benefits to consumers, research and the health system as described in the principle *GM04: Sustainable genomics*. |
| **Medical genomics** | • Establish **access of familial and pedigree data** between jurisdictions to support genetic counselling services.<br>• Adoption of **standards-based, interoperable, self-describing repositories** continues in line with system adoption strategies.<br>• Standardised **phenotypic data capture** and exchange begins between clinical systems and pathology systems and research.<br>• Genomic **data sharing for clinical purposes** established using national agreements.<br>• Support **continued development of variant curation repositories and tools** (e.g. Shariant) to adapt to expanding requirements. |
| **Genomics research** | • Genomic **data sharing for research** established using national agreements.<br>• A **national consent mechanism** is operational to support research data sharing.<br>• Apply the learnings from **trials of research data sharing** with leading clinical groups to the broader clinical/research community.<br>• Build on existing tools to **improve variant classification and curation technologies**. |
| **Infrastructure** | • Adopt national interoperability for cloud infrastructure<br>• **Trial federated query standards** across repositories to support a national genomic information network operational framework.<br>• Expand a **shared, cloud-based repository for genomic research** data across at least two jurisdictions to establish baseline and learnings to inform future implementations.<br>• Work with international groups to operationalise **international research data sharing**. |

### 5.6.3 Horizon 3 – Transition and operate

The last conceptual horizon remains to be determined, being the most likely to change to reflect changes in the national priorities and the emergence of new genomic (and other 'omic) technologies. Horizon 3 should deliver the operationalisation of capabilities established in previous horizons into mature services that support all genomic fields, that is forward reaching and responsive in a less than predictable future. When coupled with other emerging technologies, approaches and clinical advancements within the health system, realise a learning health system.

| Area | Indicative activities |
|---|---|
| **Governance** | • Monitor and review the **national genomic information network operational framework** to ensure it remains fit for purpose.<br>• Continue **monitoring emerging technologies** that will influence the national approach to genomic information management.<br>• **Review the outcomes from Horizon 2** and update the roadmap to reflect changes in priorities and emerging technologies. |
| **Medical genomics** | • Adoption of **standards-based, interoperable, self-describing repositories** continues in line with system adoption strategies.<br>• Genomic **data sharing for clinical purposes** continues using national agreements. |
| **Genomics research** | • Genomic **data sharing for research** continues using national agreements supported by the **national consent mechanism**.<br>• **Research data sharing** across clinical groups is operational across all jurisdictions.<br>• Build on existing tools to **improve variant classification and curation technologies**. |
| **Infrastructure** | • Continue roll out and standardisation of national interoperability for cloud infrastructure.<br>• Leverage **federated query standards** across repositories to support a national genomic information network operational framework.<br>• Expand a **shared, cloud-based repository for genomic research** data across all jurisdictions to complete the national genomic information network operational framework.<br>• Monitor and leverage **international research data sharing**. |

### 5.6.4 Who should be involved?

The activities outlined in this roadmap cover a broad range of functions and will require a broad range of stakeholders to be engaged. Groups that should be considered include:

- the Commonwealth Department of Health
- other Commonwealth departments and agencies, including:
    - Services Australia
    - Australian Institute of Health and Welfare
    - National Health and Medical Research Council
    - Australian Commission on Safety and Quality in Health Care
    - Australian Digital Health Agency
    - HealthDirect Australia
- Jurisdictional health departments and their digital teams
- the genomic Alliances (Australian Genomics, Queensland Genomics, Melbourne Genomics, GA4GH)
- research groups and institutes
- peak bodies for relevant clinical disciplines (e.g. Royal College of Pathologists of Australia (RCPA))
- standards development groups such as Integrating the Healthcare Enterprise (IHE) and HL7
- international groups already involved in genomics (e.g. Genomics England, EGA, Genomics Canada)
- technology industry participants (e.g. Illumina, Roche, Google, AWS) and peak bodies
- private health insurance organisations
- organisations wanting to commercialise genomics (e.g. pharmaceutical companies and others).

As noted in the governance activities within each horizon, governance groups will need to be established to guide and prioritise this work. The exact structure and hierarchy of these groups would need to be determined to define what activities receive oversight by which groups.

The above list includes several organisation types with a commercial interest in genomics. The level of involvement of such organisations will be dependent on achieving a social licence for such activities and broader understanding of the benefits to consumers, researchers and the health system.

## 5.7 Implementing solutions against this logical architecture

The logical architecture should act as a common language independent of actual implementations that can compare such implementations. Individual implementations should be able to be mapped against the architecture, noting that some implementations may not require all components of a model.

To illustrate this approach, Figure 13 shows how the GenoVic system delivered by Melbourne Genomics to support clinical genomics corresponds to the logical architecture for genomics medicine.

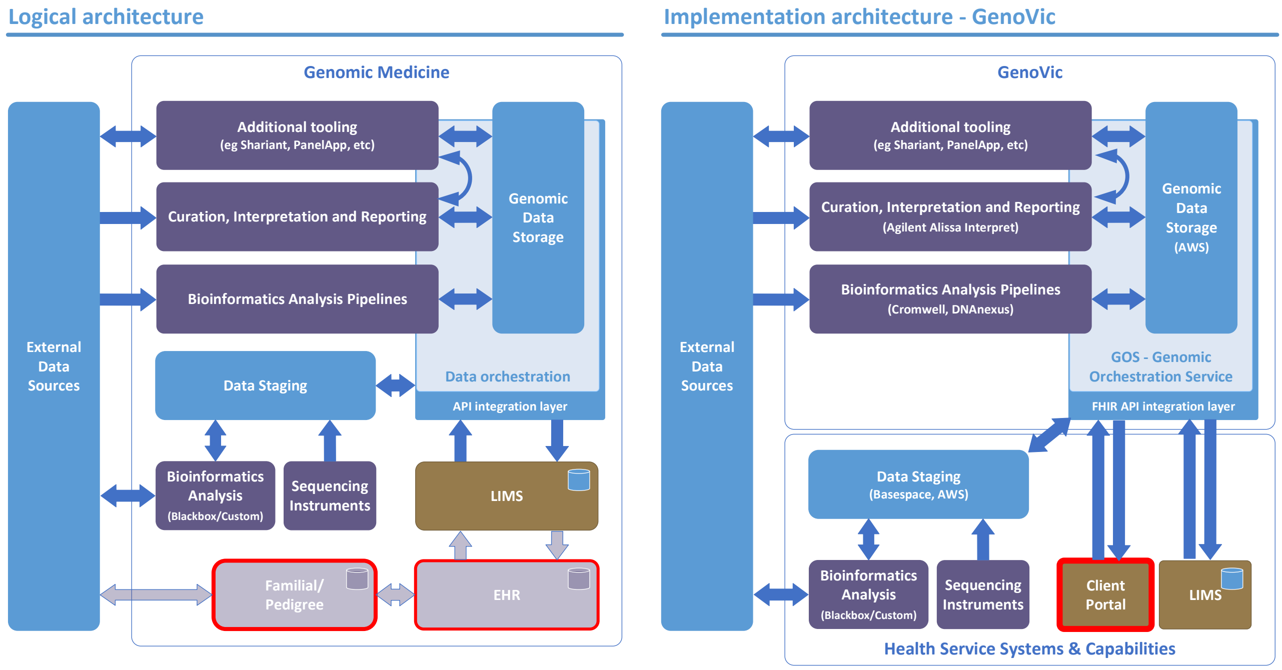

Logical architecture
Genomic Medicine
External Data Sources
Additional tooling (eg Shariant, PanelApp, etc)
Curation, Interpretation and Reporting
Bioinformatics Analysis Pipelines
Genomic Data Storage
Data orchestration
API integration layer
Data Staging
Bioinformatics Analysis (Blackbox/Custom)
Sequencing Instruments
LIMS
Familial/ Pedigree
EHR
Implementation architecture - GenoVic
GenoVic
External Data Sources
Additional tooling (eg Shariant, PanelApp, etc)
Curation, Interpretation and Reporting (Agilent Alissa Interpret)
Bioinformatics Analysis Pipelines (Cromwell, DNAnexus)
Genomic Data Storage (AWS)
GOS - Genomic Orchestration Service
FHIR API integration layer
Data Staging (Basespace, AWS)
Bioinformatics Analysis (Blackbox/Custom)
Sequencing Instruments
Client Portal
LIMS
Health Service Systems & Capabilities

*Figure 13: Comparing an implementation against the logical model*

The following should be noted:

- In the GenoVic implementation, specific technologies and tools are called out and these reflect the current state solution rather than what is in scope for the future state solution.
- The strength of the modular architecture of GenoVic means that additional tools (i.e. for curation or pipelines) and EHRs can be included in the solution in the future.
- For the GenoVic implementation, clinical information such as phenotype, ethnicity and familial relationships is supported and utilise controlled terminologies and ontologies through standards such as AeHRC's Ontoserver.
- GenoVic provides a client portal to manage the solution for those health services which have not integrated their LIMS solution.

A similar approach can be taken with other implementations in both the genomic medicine or genomic research areas.

## 5.8 Determining the correct approach

The roadmap provided above provides an archetype for an approach that could be implemented over a multi-year timeframe. Further input will be required from the proposed coordination and governance groups. The establishment of such a governance group is one of the first activities required, and a formalised roadmap should be delivered early in that group's existence after discussions with funding bodies.

Such a governance group should include all stakeholder groups: clinicians, public and private healthcare delivery organisations, researchers, policy makers, funders, industry and consumers. Aboriginal and Torres Strait Islander people are acknowledged and recognised for their important and essential role in this group. Other stakeholders may be considered, as outlined in Section 5.6.4.

When considering the architectural approach, not everything could or should be centralised. Those aspects of the national genomic information network that need to be centralised should be established by a central authority, especially those elements that traditionally are not done well at the local level. However, some elements are naturally best handled at the edge of the network.

However, the growing capabilities of cloud technologies are changing our perception of what 'national' or 'centralised' systems need to look like. Combined with substantial national technical and research capabilities, the governance group and national participants should be open to new approaches to the concept of federated systems driven by advances in virtualisation, interconnectedness and learning systems. Organisations with existing capabilities such as BioCommons, AARNet and the Australian Access Federation are identified here to demonstrate support for such a model.

Considerations of a federated, interoperable system include:

- Different jurisdictional data management requirements can be accommodated because individual data owners/custodians can tailor requirements and policies, within accepted standards, through an individualised dashboard.
- Funding, priorities and policy drivers reflect variation and inequity between jurisdictions. Central management of some essential functions through a federated model would lower barriers to uptake by resource-poor jurisdictions. Federation would also enable scale and cost levelling as it does in other areas of national activity.
- Centralised, distributed, large or small compute capabilities either part of or outside the system are all compatible with a federated model based on agreed standards.

A hybrid solution employing advanced technology and best practice approaches could leverage the most of centralised and highly decentralised models.

Identification of the core elements that could be centralised to enable effective coordination of distributed data stores will require quality metrics, data structures, metadata schemes, APIs, analytical approaches and security.

To orchestrate such a network of systems, there should be a national governance group with robust legal and governance structures. This group should have strong links to both the research community and healthcare delivery, but equally must engage with the community to establish a social licence (a trust-based relationship as per *CN02: Trust*) to support this work. The nature of genomic technologies and approaches is still emergent and dynamic. These details will be much better addressed over time once the governance group and relationships are established.

# 6 Genomic data governance framework

Data governance can be defined as 'a system of decision rights and accountabilities for information related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods.' [74]

Key ethical, legal and social issues (ELSI) relating to the collect, use and management of genomic information need to be addressed as a priority. This will be a critical enabler for a national approach to genomic information management, including building capability and capacity.

Such ELSI issues can be addressed through a national genomic data governance framework. They include, but are not limited to:

- consent and waivers of consent
- privacy and its limitations
- genetic discrimination
- secondary use of data including for commercial purposes and disclosure for law enforcement purposes
- return of findings
- ownership
- custodianship
- sovereignty
- intellectual property and provenance.

A national genomic data governance framework needs to address both clinical and research data governance. Most of the international frameworks available for data governance (and specifically data sharing) are focused on research uses and less on clinical reuse (which should not to be confused with clinical research). The needs of researchers, clinicians, policy makers and individuals may not align and must be balanced.

Examples of existing data governance frameworks across Australian government and academic institutions include:

- The **AIHW Data Governance Framework**, which applies to data assets in the AIHW data catalogue, including data collections and pointers to other data holdings [75].
- The **National Archives of Australia Information and Data Governance Framework**, which provides a basis for decisions and activities relating to their information and data assets [76].
- The **National Blood Authority Data and Information Governance Framework**, which defines the National Blood Authority (NBA) governance principles and arrangements for the NBA's management of data and information, and for the NBA's dealings with data stakeholders in the blood sector [77].
- The **Australian Commission on Safety and Quality in Health Care Data Plan 2016-19**, which supports the priorities agreed to in the Commission's Work Plan, and to outline data requirements to ensure these responsibilities and objectives are met [78].
- The **Standing Committee on Screening** have released the *Genomic Tests in Population based Screening Programs: Statement*, which outlines principles for data management in population screening programs that use genomic tests [179].

- The **Policies and standards for data governance at UNSW**, which reflect the University of New South Wales's approach to institutional data and information as a strategic asset which requires management to support the University's operations [79].
- The **Patron Data Governance Framework** at the University of Melbourne, which supports the secondary use of health data for research through their *Data for Decisions* program [80].

While each framework addresses data governance in general rather than genomics specifically, they are generally consistent in the broad approach.

In genomics, the GA4GH have developed a considerable body of standards and guidance materials in their Regulatory & Ethics Toolkit [81]. Some of this work has also been leveraged by the World Economic Forum's *Genomic Data Policy Framework and Ethical Tensions* [29], which specifically provides principles on consent, data privacy, data access and benefit sharing.

The following sections describe the key components required for a national data governance framework to support an advancing genomics system.

## 6.1 Data lifecycle management

All data has a lifecycle that requires management of the data and information as it is collected, used and eventually retired.

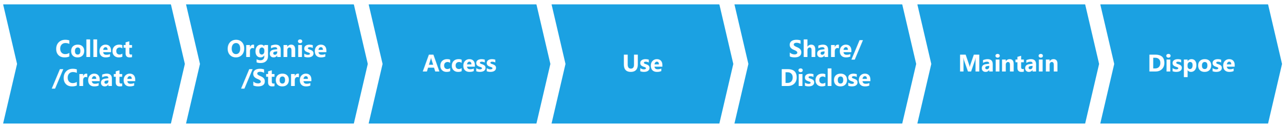


*Figure 14: Lifecycle of data*

At each stage several questions must be considered:

- **Collect/create** – How is the data collected or created? What is the provenance of the data?
- **Organise/store** – Where and how is the data being stored? Is the data immediately available or held in a format for long term archiving (e.g. tapes), unavailable for immediate access/computation?
- **Access** – How will the data be accessed? Is it static/unchanging or is it transactional in nature?
- **Use** – How will the data be used? Is this in line with the consent granted?
- **Share/disclose** – With whom will the data be shared? Is this in line with the consent granted?
- **Maintain** – How is the quality of the data maintained? What metadata needs to be recorded? How is data preserved?
- **Dispose/Reuse** – How is the data disposed of / reused at the end of its primary purpose and how is this determined?

While these are common to much data, additional concerns apply to genomic data, as outlined below.

## 6.2 Data aspects of consent

Managing informed consent in line with principle *CN03: Informed consent* is a key aspect of gaining trust within the community for genomic data use (also see principle *CN02: Trust*). Consent is frequently considered solely as an individual decision (especially from a legal position), but consent can be influenced through communities of interest.

While a full analysis of consent is outside the scope of this Blueprint, establishing and understanding the way consent provides data to be managed is critical to both ethical and future healthcare outcomes.

Factors that need to be considered include:

- Free, prior and fully informed consent is a cornerstone of research ethics, as outlined in the *National Statement* [6] and principle *GR02: Ethical data and provenance*.
- Traditional paper-based consent mechanisms do not transition or scale well in a digital environment [82].
- Consent gained for diagnosis or treatment in a clinical setting is different to the consent required for research. Consent data structures need to support both mechanisms.
- Using ethical broad-based consent from participants to support research can be challenging, however, this issue is found and addressed in biobank solutions [83].
- Australian Genomics and others are addressing the delivery of dynamic consent which allows research participants to remain engaged with their consent over research, including the right to withdraw consent completely for projects [16].
- The GA4GH has developed a Data Use Ontology (DUO) for defining data use that can support computable consent [84].

Consent data required to support data sharing includes:

- the data that can be shared
- with whom the data can be shared (not-for-profit versus commercial use; geographic restrictions)
- the use made of the shared data (conditions or restrictions on use)
- the period for which the individual will share that data.

Although broader issues related to consent are being addressed elsewhere, it is important to note the requirements of machine actionable processes that support consent. The digital front-end of consent may vary among studies, clinical settings, institutions, investigators and jurisdictions. To be effective for large, persistent genomic data stores, these digital front-end processes need to work with underlying data structures and metadata schemes. The need for interoperability should be the key principle guiding the design of digital front ends, whether these are in person, online, dynamic or patient accessible. GA4GH have developed *Machine Readable Consent Guidance* [85] based on their DUO standard which may assist.

It should be noted in some cases, ethics committees or data custodians may grant consent waivers to de-identified data sets for certain approved research purposes. When recording consent data, such consent waivers need to be considered as an option.

## 6.3 Data sovereignty

There are two aspects to consider when discussing data sovereignty: that which applies to the data of Aboriginal and Torres Strait Islander peoples, and that which applies in an Australian or jurisdictional setting.

### 6.3.1 Indigenous data sovereignty

Indigenous communities around the world know both the value and potential harm of genomic research for their communities [12], [13]. Calls for data sovereignty by Indigenous Peoples are well established [86]–[89]. Work by the Lowitja Institute [90] and by the National Centre for Indigenous Genomics [91] relates specifically to the needs of Aboriginal and Torres Strait Islander peoples in genomics.

Factors that need to be considered include:

- historical truth telling in genomics with respect to eugenics and denial of cultural identity
- the role of informed individual and collective consent, including the possibility of withdrawal and renewal
- risks of discrimination from research including implications for employment and insurance if people are found susceptible to certain conditions
- sharing benefits of research with participants

- the need for community consultation to prevent inclusion of specific findings and research publication
- participation in consumer advisory groups to support ethical and culturally appropriate approaches to research
- specific needs regarding the handling and return of bio-specimens
- existing community priorities for genetic research.

### 6.3.2 Jurisdictional data sovereignty

Besides sovereignty over data for Aboriginal and Torres Strait Islander peoples, there are broader data sovereignty issues for national and international sharing and storage of data.

There is a complex mix of rights, permissions and ownership that need to be considered when establishing data sharing and storage arrangements. This is further complicated by multiple stakeholders involved in the generation of genomic data, including:

- the patient whose sample is being analysed
- the clinician requesting the test
- the laboratory sequencing the data
- the organisation who paid for the test and/or associated data collection
- the research group which leads the research program
- the clinicians who collect and contribute or derive the related clinical dataset
- the secondary researcher who creates reanalysed data

The GA4GH have established terminology describing the roles, and this is used in publications such as their *Data Privacy and Security Policy* [92]. Consistent terminology such as that used by the GA4GH can improve communication and avoid overloaded words such as 'ownership' which can have varying legal and moral interpretations.

Commonwealth, state and territory legislation and regulation can also mandate the handling of health data regarding storage locations and sharing across borders. Onshore repositories for the deposition of research data may be needed to enable data sovereignty while simultaneously meeting the requirements for data access imposed by journal editorial policies and international agreements.

## 6.4 Data ownership, commercialisation and legal considerations

There is substantial commercial potential in genomics data for development of value-added services, drug development and augmented clinical services [93]. Commercial drivers can generate capital to develop a secondary economy based on genomics data. Consumer genomics companies such as 23andMe are spearheading large-scale research projects [94] in collaborations with pharmaceutical companies to combat large societal problems.

Whether or not this commercial model will succeed in Australia, these providers demonstrate the power of a consumer-centric model linking genomics research and healthcare, that may have lessons for developing an effective architecture of systems and relationships within the Australian public sector.

Additionally, there is the very reasonable driver for research institutions to commercialise the results of their research. However, there are debates around the validity of patents over genetic discoveries (as opposed to applications of these discoveries) [95].

Finally, there is the issue of 'ownership' of data. Some principles in this Blueprint assert the importance of patient (or community) involvement in governance of data and the deriving of benefit from its provision, including:

- *CN07: Benefit from use*

- *CN04: Right to access*
- *CN05: Use of data/portability*
- *IG01: Collective and individual benefit*
- *IG02: Authority to control.*

These speak to:

- a right to access or use, rather than ownership
- a right to benefit from the outcomes of research, which may not include a monetary benefit.

The role of those that provide the samples from which data is derived should be a factor in research design, consistent with the principle *CN01: Person-centred focus*.

This Blueprint does not attempt to address the legal or intellectual property issues commonly labelled 'data ownership', noting that such ownership may vary over time; the initial analysis may be handled by a different entity to that which stores data, and later re-analysis may be undertaken by yet another party. Rights to data access by relevant parties (patient, treating physician, researcher, industry, etc.) need to be addressed within this context.

To address data ownership at a national level would require a review of legislation, regulation and policy across both Commonwealth and jurisdictional contexts. Melbourne Genomics have undertaken a review from a Victorian perspective (which considers some other jurisdictions). Australian Genomics have also undertaken investigations. A national approach will need to look at all Commonwealth and state/territory legislation, regulation and policy to determine how these can be harmonised to support national use of genomic data.

## 6.5 Privacy

Handling personally identifiable information (PII) such as health data is controlled by many legal instruments, including Commonwealth legislation, regulation and policy (such as the APPs [16]), state and territory health and privacy legislation [96], and even the GDPR of the European Union [19], [20].

Within clinical contexts, the storage, use and privacy considerations of PII are routinely dealt with. In principle *GM06: Genomic data is clinical data*, genomic data related to patients should be treated with the same importance as other clinical data regarding confidentiality and rights to access the information.

For research, the data may be held in identified or de-identified forms, depending upon the research and the requirements of data linkage. The debate over whether genomic data such as genomic sequence data are regarded as de-identified has many interpretations [97]. Genomic data must be part of any privacy review for research projects, in line with current practice in Australia.

The GA4GH has developed a detailed *Data Privacy and Security Policy* that addresses privacy and security practices for data sharing in clinical and research contexts [92]. From a privacy perspective, this policy document addresses:

- the legal aspects of data processing
- risks and safeguards for data privacy
- consent and related legal matters
- restrictions on re-identification of research data
- disclosing identifiable data in a public setting
- the retention and disposition of data
- constraints on data access
- transparency of policies and processes
- accountability for data privacy
- special steps for data related to populations perceived as vulnerable.

A consistent set of policies, procedures and standards addressing privacy across all jurisdictions should allow data sharing to be achieved. Alignment to the GA4GH *Data Privacy and Security Policy* would also support a broader international consistency.

## 6.6 Security

Security of data is critical to obtain the social licence to make use of genomics in Australia (see principle *CN02: Trust*). A full analysis of possible security models and encryption technologies is outside the scope of this Blueprint. However, as implementations are undertaken consideration of these issues will be necessary:

- the implications of security of cloud-base solutions
- what encryption methods are used for data at rest and data being transmitted
- whether or not data is stored in an identifiable manner or if personal identifiers are stored separately
- implications of and control of data linkages to other data sets.

The GA4GH has established an initial set of security standards and policies through one of its work streams. The GA4GH Security work stream advocates implementers to apply 'defence in depth to protect the high-value data we rely upon to accelerate the acquisition and application of biomedical knowledge' [98].

As noted previously, the GA4GH *Data Privacy and Security Policy* addresses privacy and security practices for data sharing in clinical and research contexts [92]. From a security perspective, this policy document addresses:

- procedural approaches that can be taken by organisations to mitigate or avoid security risks
- technical measures that can be implemented
- physical measures that can reduce security risks.

A consistent set of policies, procedures and standards addressing security requirements across all jurisdictions should be established to allow data sharing to be achieved. Alignment to the GA4GH *Data Privacy and Security Policy* would also support a broader international consistency.

Note that based on feedback from clinicians and researchers in Australia and internationally, historical concerns around security of cloud storage mechanisms have reduced substantially. Security in the cloud leverages the significant investments made by cloud service operators and large international corporations [99].

## 6.7 Data sharing

If consent and privacy issues are addressed, any sharing of data needs to occur within context and a framework of agreements. There are five contexts for genomics data:

- clinical use for delivering clinical care
- clinical reuse, such as using genomic data for one patient to address the healthcare needs of a related patient
- research purposes
- laboratory validation work
- education, training and outreach purposes.

Templates for these agreements have been developed nationally and internationally. Nationally, agreements for genomic data sharing have been developed by Australian Genomics (focusing on research data sharing) [100] and Melbourne Genomics (focusing on clinical reuse as well as research data sharing) [101], as well as more general data sharing agreements available from government and university organisations. Internationally, guidance is available from GA4GH [28]. A national approach will require an agreed set of templates that support the contexts and address the potential intellectual property issues associated with academic research.

Data sharing is considerably more complex than just reaching a contractual agreement. There are many processes involved; considerations for national implementations; and future standards and tools being developed, for implementing data sharing successfully and at scale. For example:

- Processes and systems for managing data access applications and approvals: technical systems (e.g. DUOS, REMS), implementing machine readable codes or ontologies (e.g. DUO) to automate/ streamline data sharing and processing access applications; Governance processes, such as establishment and operating data access committees (DACs).
- Future tools and standards such as GA4GH passport and visas, and library card systems, that would support data sharing processes at scale, and for national implementations.
- Policies and processes for managing incidental findings or secondary findings, when on-sharing data for secondary purposes.
- Considerations around sharing ‘aggregated’ data vs ‘individual level’ data
- Processes to ingest and collect data at scale, for example, from sequencing laboratories to support aggregation and subsequent sharing of data; includes technical transfer of large-scale datasets, acquisition of comprehensive metadata; data transfer agreements (different legal and governance considerations to data sharing agreements).
- Authentication and authorisation challenges and requirements for supporting data access, and particularly for national data sharing: managing researcher identity, institutionally verified authorisation for access to data. There will need to be an exploration of Australian requirements for national systems around this, similar to international models (e.g. Elixir Authentication and Authorisation Infrastructure).

The above discussion is largely focusing on sharing data between clinical contexts and research. However, a one-way moving of data from the laboratory system to a data repository that can be accessed by researchers does not address other data sharing requirements. The value in sharing is not in this one-way exchange of information but in a two-way exchange that gets information from other laboratories back into the laboratory's analysis pipeline to aid in curation. This has been identified as a key driver for developing Shariant [102] and other information sharing tools. Clinical laboratories are time poor and need the evidence accessible from within their interpretation system. This is also in line with international collaborative efforts.

## 6.8 Data quality, provenance and metadata management

In the genomic data categorisation framework in Section 3, there are many types of data that need to be managed and governed to support the use of genomic data and information in Australia.

Beyond the quality control processes required to ensure the quality of the resulting data discussed in Section 3, when considered at a national level, such concerns are increased [33]. Standardisation of approaches to the collection, storage and management of quality control data, where possible, will support national applications of this data.

Similarly, using metadata management systems for the retention and analysis of provenance data and other forms of genomic metadata will be critical to a robust research capacity.

## 6.9 Data retention

Data is often seen differently by researchers and clinicians. Researchers collect data with a clear intent about its use in research, whereas health services have traditionally collected data as a by-product of the delivery of care, especially what is typically considered ‘administrative health data’ [103], [104].

Traditional clinical practice generally sees the role of the data as being complete once a diagnosis (or prognosis) is achieved. For the public health system, the ongoing cost of managing genomic data may not be trivial. While traditional health economics has difficultly forecasting the benefit to be actualised through this expenditure, the World Economic Forum and others are investigating the value proposition presented by retention of genomic data [105].

Despite this, public laboratories in Australia are generally retaining data longer than the minimum times set by regulation or accreditation [106]. The duration and retention rules associated with genomic and other clinical data varies according to the data involved, and whether the patient was a child or adult.

The genomic data categorisation framework outlined in Section 3 can support a set of nationally agreed retention rules that meet the requirements of both medical genomics and genomics research communities. The adoption of a national approach in line with principle *DM03: Genomic data retention* will ensure that the cost of genomic data management within healthcare systems are understood and budgeted. These costs can be offset by economic models that recognise the value of research to reducing the costs of healthcare provision and guide an ecosystem approach to data storage and management.

## 6.10 Governance structure

A national approach to genomic information management requires good governance (see principle *DM05: Strong governance models*) that can be applied with consistency. This will require an operational model that will support the diverse requirements of the clinical and research sectors. The *Data Management Body of Knowledge* (DMBOK) [107] describes six common models:

- **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.

The complexities of the Australian healthcare and research sectors suggest that a federated or hybrid model are most appropriate. A decision on such a model will be needed to support a governance framework for genomic data.

Regardless of the model selected, involvement by those that provide samples from which data is derived need to be considered, to ensure that benefits are delivered (*CN07: Benefit from use*), access is available (*CN04: Right to access*) and individuals and groups retain control over using their data (*IG02: Authority to control*).

### 6.10.1 Workforce implications

Although workforce development is beyond the scope of this Blueprint, a national approach to genomic information management will require skills resources to coordinate, collaborate and govern the way data and information is managed. This will include skills from a wide variety of disciplines, including data managers, system administrators, technology specialists, data scientists, medical scientists, clinicians, researchers and other diagnostic professionals.

When developing the governance model for a national approach, consideration of these skill requirements will support the ability of the sector to leverage the value to be gained through use of genomic data.

# 7 Standards and interoperability

The *National Digital Health Strategy* [108] calls out interoperability (and data quality) as one of seven strategic priorities for the Australian health sector. Interoperability is 'the ability of different information systems, devices and applications (systems) to access, exchange, integrate and cooperatively use data in a coordinated manner, within and across organisational, regional and national boundaries, to provide timely and seamless portability of information and optimise the health of individuals and populations globally.' [109]

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 across the spectrum of care, within all applicable settings and with relevant stakeholders, including the individual [109].

Effective data standards are necessary to support a national approach to genomics information management. To support interoperability, a standards-based approach needs to be taken. This is consistent with action 5.3 in the NHGPF [1]: *Develop nationally agreed standards for data collection, safe storage, data sharing, custodianship, analysis, reporting and privacy requirements*, and action 20 of the *Implementation Plan* [2]:

> *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.*

It is also worth noting that the Australian Digital Health Agency is running an interoperability program [110] which is pursuing a broad engagement agenda and has already published a document entitled *A Health Interoperability Standards Development, Maintenance and Management Model* [111]. While not specifically addressing genomics, many of the areas under consideration will bear directly on genomic data standardisation.

This section outlines core standards considered important to this effort. Consideration will need to be made regarding possible different requirements for clinical collections compared to research collections of genomics data.

## 7.1 Global Alliance for Genomics & Health

The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organisation, seeking to enable responsible genomic data sharing within a human rights framework.

The GA4GH is addressing the differences between healthcare and research in areas of policy, language and funding and bring down the technical barriers to interoperability. A key approach is to establish standardised, accredited processes in medical genomics that will then allow a pivot to support research genomics. This approach has been trialled in Sweden and Canada. Australian Genomics is also a GA4GH Driver Project, piloting the GA4GH tools and standards in Australia [112].

There are three broad areas of work with related standards and projects:

- **Genomic Data Toolkit** provides open standards for genomic data sharing:

  - The **GA4GH Data Use Ontology (DUO)** allows users to semantically tag genomic datasets with usage restrictions, allowing them to become automatically discoverable based on a health, clinical or biomedical researcher's authorisation level or intended use.
  - The **Data Repository Service** API is a standard for building data repositories and adapting access tools to work with those repositories. The API allows data consumers to access datasets regardless of the repository in which they are stored or managed.
  - The **GA4GH Passport** specification aims to support data access policies within current and evolving data access governance systems.
  - **Phenopackets** provides information models with different levels of complexity to enable high-level clinical phenotype information and deep clinical phenotype information to be exchanged.
  - The **RNAget API** v1 provides a means of retrieving data from several types of RNA.
  - The **Service Info** API is an endpoint for describing GA4GH service metadata.
  - The **Service Registry** API provides information about other GA4GH services, primarily to organise services into networks or groups and service discovery across organisational boundaries.
  - The **Tool Registry Service** is a standard API for exchanging tools and workflows to analyse, read and manipulate genomic data.
  - The **Beacon API** can be implemented as a web-accessible service that users may query for information about a specific allele.
  - The **CRAM file format** is an efficient storage format for read data, achieving lossless compression better than BAM, while maintaining full compatibility.
  - **Crypt4GH** is a file format that can store data in an encrypted and authenticated state.
  - The **Family History Tool Inventory** is a catalogue of family history tools available for documenting family health history information.
  - **htsget** is a genomic data retrieval specification that allows users to download read data for subsections of the genome in which they are interested.
  - The GA4GH **refget API** enables access to reference genomic sequences without ambiguity from different databases and servers using a checksum identifier based on the sequence content itself.
  - **SAM/BAM File Formats** v1 are specifications for storing next-generation sequencing read data
  - **Variant Benchmarking Tools** provides standardised benchmarking methods and tools are essential to robust accuracy assessment of next-generation sequencing variant calling.
  - The **Variation Representation** specification provides a flexible framework of computational models, schemas and algorithms to precisely and consistently exchange genetic variation data across communities.
  - The **Workflow Execution Service (WES)** API provides a standard that lets users run a single workflow (defined using Common Workflow Language (CWL) or Workflow Description Language (WDL)) on multiple different platforms, clouds, and environments and be confident that it will work the same way.

- The **Regulatory & Ethics Toolkit** provides ready-to-use regulatory and ethics guidance for genomic and health-related data sharing:

  - The **GA4GH Framework for Responsible Sharing of Genomic and Health-Related Data** provides a principled and practical framework for the responsible sharing of genomic and health-related data. It contains foundational principles and core elements for responsible data sharing and is guided by human rights, including the right to benefit from the progress of science, and privacy, non-discrimination and procedural fairness.
  - The **GA4GH Consent Policy** aims to guide sharing genomic and health-related data so it respects autonomous decision-making while promoting the common good of international data sharing.

  - The **GA4GH Privacy and Security Policy** aims to guide sharing genomic and health-related data so it protects and promotes the confidentiality, integrity, and availability of data and services, and the privacy of individuals, families and communities whose data are shared.

- The **Data Security Toolkit** provides ready-to-use Data Security for genomic data sharing:
  - The **Data Security Infrastructure Policy** describes the security infrastructure policy and technology recommended for stakeholders in the international genomic data sharing ecosystem.
  - **Authentication & Authorisation Infrastructure** is a GA4GH-approved Standard which leverages OpenID Connect (OIDC) Servers for authenticating the identity of researchers desiring to access clinical and genomic resources from data holders adhering to GA4GH standards, and to enable data holders to obtain security-related attributes of those researchers.

## 7.2 HL7 Standards

Founded in 1987, Health Level Seven International (HL7) is a not-for-profit, ANSI-accredited Standards Development Organisation providing a comprehensive framework and related standards for the exchange, integration, sharing and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services.

There are three generations of HL7 standards [113] that need to be considered:

- **V2:** As one of the most widely implemented standards for healthcare information in the world, the Version 2 Messaging Standard was first released in October 1987. Version 2.7, representing the latest update, was published in 2011. This messaging standard allows the exchange of clinical data between systems. It supports a central patient care system and a more distributed environment where data resides in departmental systems [114]. While considered by some a legacy standard, significant active implementations exist in hospital systems and pathology systems worldwide and in Australia.
- **CDA:** The HL7 Version 3 Clinical Document Architecture (CDA®) is a document mark-up standard that specifies the structure and semantics of 'clinical documents' for exchange between healthcare providers and patients. It defines a clinical document as having these six characteristics: Persistence, Stewardship, Potential for authentication, Context, Wholeness and Human readability. A CDA can contain any clinical content. Typical CDA documents could include Discharge Summaries, Imaging Reports, Pathology Reports. The most popular use is for inter-enterprise information exchange [115]. It is a core component underpinning the My Health Record system in Australia.
- **HL7 FHIR:** FHIR® is an interoperability standard intended to facilitate the exchange of healthcare information between healthcare providers, patients, caregivers, payers, researchers and others involved in the healthcare ecosystem. It consists of a content model in 'resources' and a specification for the exchange of these resources in the form of real-time RESTful interfaces as well as messaging and documents [116]. While a recent standard, it is receiving much attention worldwide. Specific resources have been developed for genomics [117] and consent [118].

## 7.3 Observational Health Data Sciences & Informatics (OHDSI)

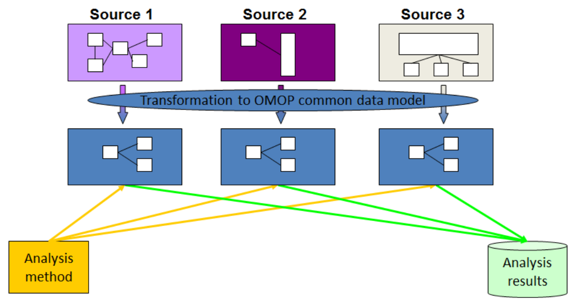


*Figure 15: Conceptual approach to using a CDM*

The OHDSI program is a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics, with over 2500 users across six continents. With more than half a billion patient records and a common data model, OHDSI specifies strict coding, privacy and security standards that all collaborators have agreed to adopt.

The work enables the unification of data from multiple sources to provide the data mass, reproducibility testing and statistical power required by analytics, especially those powered by Artificial Intelligence (AI) methods.

Based on the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), it allows for the systematic analysis of disparate observational databases. Data within those databases is transformed into a common format (data model) and a common representation (terminologies, vocabularies, coding schemes). Systematic analyses using a library of standard analytic routines written based on the common format can then be performed.

While the core CDM model is focused on clinical data from EHRs, an extension to support genomic data has been proposed called G-CDM. A working group within OHDSI develops this work [119].

The Transformational Data Collaboration (TDC) [120] is an initiative under the auspices of the Australian Health Research Alliance (AHRA) [121]. The TDC has established an Australian chapter of the OHDSI [122], with the goal of:

> “To utilise the unique open and collaborative nature of AHRA to help develop and support national data initiatives where an open, inclusive and non-competitive environment is required.”

## 7.4 Functional Genomics Data (GFED) Society

The GFED Society works with other organisations to accelerate and support the effective sharing and reproducibility of functional genomics data. They facilitate the creation and use of standards and software tools that allow researchers to annotate and share data easily [123].

They have a broad focus on any data generated using any functional genomics technology applied to genomic-scale studies of gene expression, binding, modification (such as DNA methylation), and other related applications. Projects undertaken include:

- The formulation of the minimum information about a microarray experiment required to interpret and verify the results (MIAME).
- Developing the Minimum Information about a high-throughput SEQuencing Experiment standard for Ultra High-Throughput Sequencing experiments.
- A simple spreadsheet-based, MIAME-supportive format for microarray experimental data called MAGE-TAB, based on a richer a data exchange and object modelling format known as MAGE.
- A stand-alone desktop application to help bench biologists annotate biomedical investigations and their resulting data.
- Developing ontologies for microarray experiment description and biological material (biomaterial) annotation.

- Engaging with and supporting the efforts of other relevant standards organisations.

## 7.5 Metadata standards

Metadata standards, such as those developed in AIHW's METeOR [124], describe the expected meaning and acceptable representation of data for use within a defined context. Metadata standards are endorsed for use within an organisation or across Australia, improving the quality, relevance, consistency and comparability of national information about the health and welfare of Australians. The drivers for standards development arise from the need for better information, whether it is statistical, administrative, clinical or other information.

The AIHW works closely with government and non-government organisations to improve adherence to data standards in administrative data collections and to promote national consistency and comparability of data and reporting.

METeOR, as Australia's repository for national metadata standards for health and welfare, should be a component of existing national infrastructure for housing metadata standards. Further, METeOR should be part of the solution for accessible genomics data standards, including metadata that facilitates data linkage.

## 7.6 Data access approaches

Genomics England guides the release and use of data from its systems, which limits access to individual patient data [125]. The EGA provides data access agreements and provides guidelines for the submission and use of data to the archive. This includes a suite of metadata regarding submissions [126], [127]. These will need to be assessed for relevance to Australia and compliance with the principles laid out in this document.

Equally, existing work has been undertaken within Australian Genomics and Melbourne Genomics (and others) regarding data sharing and access agreements [100], [101], [128]. These should be leveraged to move towards a consistent approach for Australia.

## 7.7 Other standards

Refer to Appendix A for additional standards that may apply to specific data types.

## 7.8 Assessing standards and interoperability maturity

The Healthcare Information and Management Systems Society (HIMSS) describes [109] four levels of interoperability:

- **Foundational (Level 1):** Establishes the interconnectivity requirements needed for one system or application to securely communicate data to and receive data from another. This is consistent with the definition of point-to-point interfaces described in Section 5.4.1.
- **Structural (Level 2):** Defines the format, syntax and organisation of data exchange including at the data field level for interpretation. Standardisation assists in achieving this level.
- **Semantic (Level 3):** Provides for common underlying models and codification of the data including the use of data elements with standardised definitions from publicly available value sets and coding vocabularies, providing shared understanding and meaning to the user.
- **Organisational (Level 4):** Includes governance, policy, social, legal and organisational considerations to facilitate the secure, seamless and timely communication and use of data both within and between organisations, entities and individuals. These components enable shared consent, trust and integrated end user processes and workflows.

Regardless of whether they are state and territory health services, research institutions or commercial entities, organisations exist at varying levels of capability (and indeed can vary even within a single organisation). The ability to understand the capability maturity of an organisation will assist in identifying the need for investment and prioritisation of activities and allow for realistic assessments of when or if an organisation can participate in a national approach to genomic information management.

# Appendices

# Appendix A. Genomic workflows and associated data

To understand 'genomic data', we need to look at in the context of the processes that create and consume it. While those already engaged with genomics may consider this self-evident, many clinicians within the healthcare sector and most other non-clinical participants will require a basic grasp of genomics and the processes involved in producing data.

This appendix acts as a primer for those readers, and to help define the term used. It outlines the key processes in massively parallel sequencing, the data sources required as inputs and the data generated. Where appropriate, alternative processes or concepts will be identified. Note other forms of sequencing exist, resulting in similar data flows.

## A.1. Genomic variations

There are several sources of variation between human genomes[129], which may cause disease in certain cases:

- **Microscopically visible changes in chromosomes** are typically of over 5 million bases and may include entire chromosome duplications or missing chromosomes. These are present in less than 1% of the population and are almost always pathogenic.
- **Copy Number Variations (CNVs)** are common changes in regions of the genome, including duplications and deletions (typically between a thousand to five million bases). Most people have around a hundred CNVs in their genome, some of which may predispose individuals to disease but most of which are benign.
- **Single nucleotide polymorphisms (SNPs)** are common changes to single base code in the DNA. People typically have around 3 million SNPs in their genome, some of which may predispose individuals to disease but most of which are benign.

## A.2. Types of genomic testing

A variety of tests are used to identify the variations described above. Genomic testing can be broadly divided into two categories: cytogenic tests and molecular tests.

### A.2.1. Chromosomal testing

Chromosomal or cytogenic testing examines the chromosomes to determine if extra, missing, or rearranged chromosomes or genes exist [130]. Techniques used include:

- **Karyotyping:** Chromosome analysis or karyotyping evaluates the number and structure of a person's chromosomes to detect abnormalities. A karyotype examines a person's chromosomes to determine if the right number is present and to determine if each chromosome appears normal [131].
- **Fluorescence in situ hybridisation (FISH):** This test uses fluorescent probes to evaluate genes and/or DNA sequences on chromosomes. This technique can show extra gene copies (duplicated or amplified genes) and genetic sequences missing (gene deletions) or have been moved (translocated genes) [132].

- **Array Comparative Genomic Hybridisation (aCGH):** Comparative genomic hybridisation (CGH) is a method for detecting copy number polymorphisms and chromosomal imbalances in the genome. Resolution of this method can range from 25 to 200 000 bases [133].
- **SNP Arrays:** SNP arrays have better resolution than either FISH or CGH and are more economical than genomic testing [134].

### A.2.2. Molecular testing

While cytogenic testing remains an important approach for some diagnostic purposes, the RCPA found a substantial decrease compared to molecular testing between 2011 and 2017 [135].

The remainder of this document largely focuses on using massively parallel sequencing, and the data and processes associated with this technique. However, where appropriate, references may be made to these other techniques.

When considering genomic sequencing, three approaches need to be considered:

- **Exome sequencing:** This process sequences all the pieces of an individual's DNA that provide instructions for making proteins (exons). Together, all the exons in a genome are known as the exome, and the method of sequencing them is known as exome sequencing [136].
- **Genome sequencing:** Researchers have found that DNA variations outside the exons can affect gene activity and protein production and lead to genetic disorders. Genome sequencing determines the order of all the nucleotides in an individual's DNA and can determine variations in any part of the genome [136].
- **Panels:** Both exome sequencing and genome sequencing can provide more data to analyse than is sometimes necessary. Panels derive clinically relevant target sequences from sequencing technology [137].

## A.3. Genomic medicine versus genomic research

When considering genomics, it becomes readily apparent there are two primary drivers of genomics:

- **Genomic medicine:** An emerging and rapidly changing medical discipline that involves using genomic information about an individual to inform patient diagnosis and care and the health outcomes and policy implications of that clinical use [138].
- **Genomic research:** Focuses on using genomic technologies help researchers investigate the relationships between many sections of the genome and study their combined influence on health and disease [139].

While genomic medicine typically focuses on the individual and genomic research may focus on a single patient or a cohort of individuals, both use sequencing technologies to derive information from human genetic material. This relationship can be seen in Figure 16.

While each step in the figure may be required for both genomic medicine and genomics research, the satisfactory completion of each may be different. Clinical testing has specific requirements for reliability, quality control, reproducibility and software robustness which may not apply to research work. And genomic medicine covers a range of drivers, including diagnosis of rare diseases, testing of somatic tumours to identify specific cancer types (and hence treatments) and using tests of specific molecular markers inform treatment options.

Clinical processes look at individuals and individual variants. Genomics research is not limited to individual variants in individuals, therefore the bioinformatic tools and computational challenges are different between the two fields, even though many of the same processes are used.

While the following focuses on these two fields, it should be noted that translational genomics has aspects of both. Aspects of this are highlighted where appropriate in the following sections.

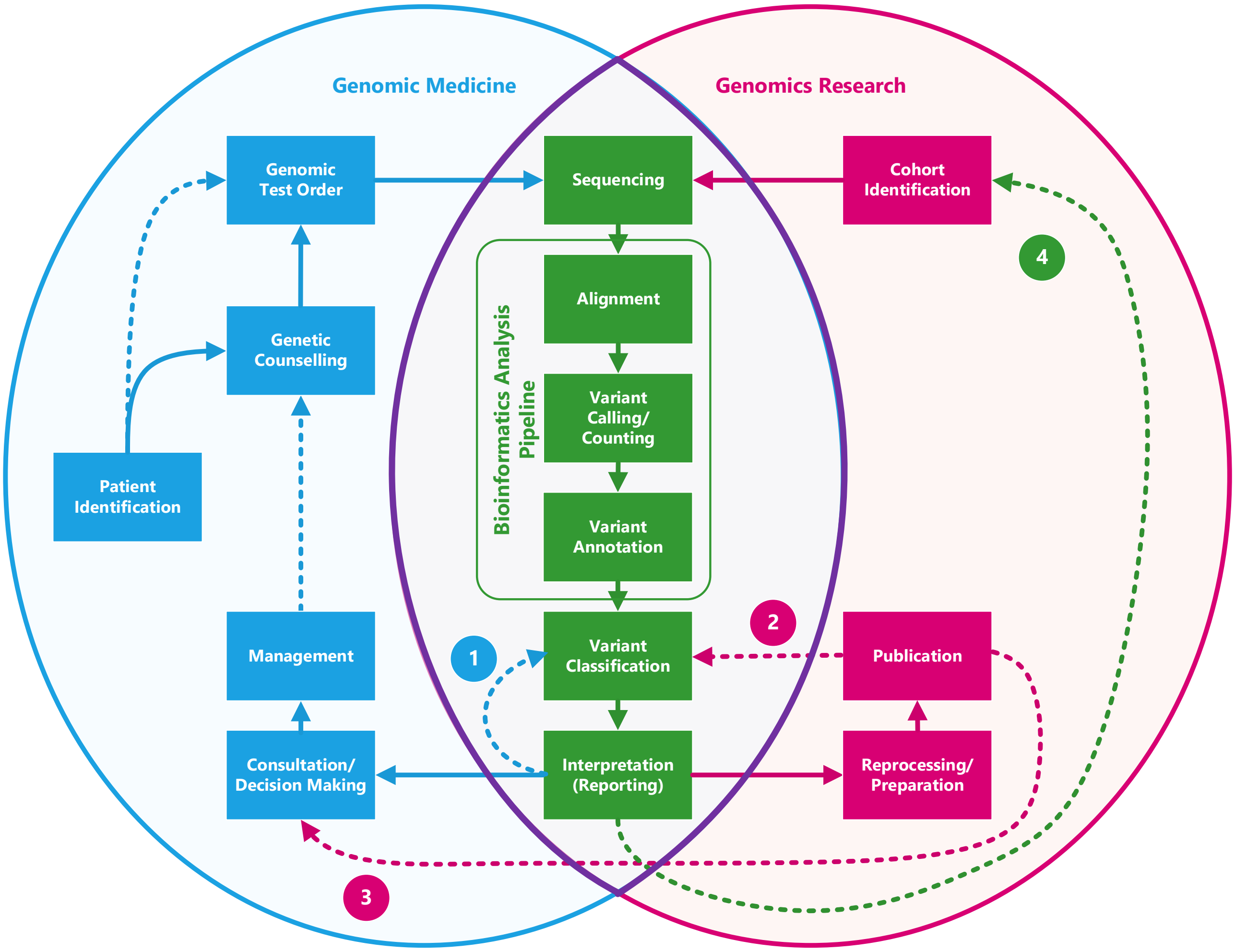


*Figure 16: The intersection between genomic medicine and genomics research*

These sections explore the activities that form the workflows for genomic medicine and genomic research. For each activity a generic description of the processes is provided and the data that comprises the inputs and outputs from that process.

It is impossible in a single diagram to detail all the complexity of both the clinical and research processes. For instance, Figure 16 does not include the development of functional and clinical annotation pipelines, the reference genome (or developing a pan-human genome resource) or population and clinical reference data. For the most part these are listed an inputs or outputs of the processes in the descriptions that follow.

There are four key elements in this document that demonstrate the interconnectedness of the two disciplines:

1 Sharing clinical interpretations (through national and international repositories) is a key input to later variant classification in both clinical and research settings.

2 The publication of research findings is another key input into variant classification and interpretation for both clinical and research settings.

3 The publication of research findings also provides important input into the decision-making and treatment approaches in clinical settings.

4 As genomic medicine becomes more mainstream, the availability of clinical genomic data to support genomic research (with consent) is anticipated to increase.

While the special role of translational genomics in linking research into clinical settings may not be clear in Figure 16, it has elements of both workflows. Clinical outcomes result, but with eventual publication of the results. Generally, translational genomics is about gathering evidence that a proposed research outcome can be applied in clinical care.

Figure 16 includes some aspects of the 'virtuous circle' that links research and clinical worlds, and these can be seen by the data flows between the two domains.

## A.4. Identifying patients (genomic medicine)

In genomic medicine, the focus is an individual patient and possibly their immediate family. The role of identification may occur within the context of primary care or within specialist or acute settings, depending upon the patient's symptoms or clinical question at hand.

Recognising patients with or at risk of genetic conditions can occur when:

- there is a known genetic diagnosis or by recognising signs and symptoms of common genetic conditions
- interpretation of a family history
- results of screening programs indicate genomic counselling is necessary
- patient symptoms cannot be diagnosed using other diagnostic tests
- patients have been referred by their general practitioner after direct-to-consumer genomic testing.

| **Inputs** | |
|---|---|
| Patient history | A detailed patient history will be held within the (electronic) medical records of the treating clinician (primary, secondary or acute) and is fundamental to establishing phenotype data. |
| Family history | A family history may be held within the (electronic) medical records of the treating clinician (primary, secondary or acute) and is the basis of a detailed pedigree to be prepared by genetic counsellors or other specialists. |
| Outputs | |
| Referral for genetic services | Most jurisdiction have a defined referral process to genetic counselling services, and this should include the patient's current conditions/symptoms, a detailed patient history and any family history that may be available. |
| Genomic test order | Sometimes, counselling may be mainstreamed within a specialty (e.g. oncology) and a test order may be the next stage of the process. This will require the appropriate patient data to support the test. |
| Other requirements | |
| Consumer oriented genomic information | An informed consumer can better understand the processes of genomic medicine and provide informed consent (or not).<br>Providing information suitable to consumers and their families is important, such as links to Patient Support and Advocacy Groups and online resources. |

| | |
|---|---|
| Genomic information for treating clinicians (primary care and others) | The Royal Australian College of General Practitioners (RACGP) provides guidelines for primary care clinicians regarding genomics [140] as do other agencies. As genomic medicine becomes more mainstream, it will be important to ensure that primary care clinicians have sufficient information to ensure equity of care and treatment for all patients. |
| Patient identifiers | All clinical systems provide identifiers for patients (and in some jurisdictions this is a jurisdiction-wide identifier), however, the consistency with which this is done can lead to issues between health services (and especially between jurisdictions). |

## A.5. Genetic counselling (genomic medicine)

All jurisdictions provide genetic counselling services, whether they are centralised, decentralised or outsourced. The workforce includes clinical geneticists, other specialists, genetic counsellors and genetic nurses.

The role of genetic counselling includes an education process that seeks to assist affected (and/or at risk) individuals to understand the:

- nature of the genetic disorder and its management/surveillance
- genetic basis, inheritance pattern, and risk to family members
- options for family planning
- the role, options, availability and possible outcomes of genetic testing, including implications for insurance.

Communicating genomics information involves:

- communicating genetic information in an understandable way
- being non-directive and supporting informed decision-making;
- understanding consent and confidentiality issues
- appreciating the emotional, ethical, legal and social impact of genetic information for a patient and their family.

Counselling services also have a role in ordering genetic tests and genomic studies, providing diagnoses to patients and referral to specialist units for further care. The Human Genetics Society of Australasia (HGSA) provides guidelines for the *Process of Genetic Counselling* [141].

As genomic medicine becomes more mainstreamed, some roles described above may be embedded in specialist services, focusing on specific clinical disciplines. This is already common within oncology specialties which may order somatic tests directly.

However, a coordinating role remains to provide stewardship of genomic services across a jurisdiction.

| **Inputs** | |
|---|---|
| Referral for genetic services | As described in the previous section. This is the initiating document for genomic services within most jurisdictions.<br>Note this may not always apply, such as for somatic testing. |
| **Outputs** | |
| Record of consent | Fundamental to working with sensitive health data such as genomics is the application of consent to the process. Recording and accessing consent information will be critical to maintaining trust regarding the storage of genomic data. |

| | |
|---|---|
| | The PRGHG also has carriage of work led from the NSW Ministry of Health to build on Australian Genomics' consent work to develop a national consent form. |
| Family pedigree | Pedigree information is required to support analysis of genomic data, especially for inherited disease).<br>Relevant standards and products include:<br>• PhenoTips [142]<br>• HL7 FHIR Family Member History resource [143]<br>• Standardised Human Pedigree Nomenclature: Update and Assessment of the Recommendations of the National Society of Genetic Counselors [144]<br>• HL7 Version 3 Standard: Genomic medicine; Pedigree, Release 1 [145]<br>• Standard Pedigree Symbols [146] |
| Phenotype information | A clinical phenotype repository holds identified clinical/phenotype data for patients. It is separated from the genomic data and restricted to those people in clinical testing laboratories to address privacy requirements.<br>Relevant standards and products include:<br>• SNOMED CT [147]<br>• Phenopackets on FHIR [148], [149]<br>• PhenoTips [142]<br>• Human Phenotype Ontology [150]<br>• CSIRO SNOMED CT to HPO Mapper [151], [152]<br>• Defining the phenotype in human genetic studies: forward genetics and reverse phenotyping [153]<br>• XGAP: a uniform and extensible data model and software platform for genotype and phenotype experiments [154] |
| **Other requirements** | |
| Consumer oriented genomic information | An informed consumer can better understand the processes of genomic medicine and provide informed consent (or not). Providing information suitable to consumers and their families is important. |
| Genomic information for treating clinicians (primary care and others) | The RACGP provides guidelines for primary care clinicians regarding genomics [140] as do other agencies. As genomic medicine becomes more mainstream, it will be important to ensure that primary care clinicians have sufficient information to ensure equity of care and treatment for all patients. |

## A.6. Genomic test order (genomic medicine)

In clinical settings, the test order is the initiation point for laboratory processes. It is a critical point, where consent and key patient information such as phenotype and family history are conveyed to the laboratory.

This process may be initiated through the genetic counselling service or may be initiated through specialties with mainstreamed genomics and provide embedded counselling services. It may also be initiated from within a pathology service (e.g. from Anatomical Pathology).

| **Inputs** | |
|---|---|
| Record of consent | See 'Record of consent" in Section A.5. |
| Family pedigree | See 'Family pedigree' in Section A.5. |
| Phenotype information | See 'Phenotype information" in Section A.5. |

| **Outputs** | |
|---|---|
| Genomic test order | Raising of a genomic test order is the start of the genomics diagnostic process. The quality and quantity of phenotype information in such orders may have a demonstrable effect on undertaking a genomic study. Existing order entry systems rarely support such phenotype data entry. |

## A.7. Preparation and sequencing (shared)

To sequence the DNA from a sample, steps are required [155] as follows:

- **Extraction of the DNA:** The first step is to extract the DNA from the sample, usually using kits designed for this purpose from the instrument manufacturer.
- **Library preparation:** The length of the fragments to be used is a key decision and will depend upon the application and technology. Fragment lengths of 100 to 300 base pairs are commonly used. The preparation process then adds a specific 'adapter' (a known DNA sequence) to the start of fragments to allow sequencing later. The result is called the genomic library.
- **Target enrichment:** When undertaking panels or exome sequencing, a process called target enrichment is used to reduce data that will be processed later, by selectively extracting only the relevant exome or target gene fragments. Choosing a target enrichment method is an application-driven decision based upon the test and the available technology. Genome sequencing does not require the enrichment step.
- **Sequencing:** A variety of technologies are available to sequence the genomic library, including sequencing by synthesis, ion semiconductor sequencing, single molecule real-time sequencing (SMRT) and nanopore sequencing. The sequencing technology result is recorded in a read file (commonly FASTQ format). While similar, differences exist between the data generated by different technologies.

Recording the detailed processes used is critical, as differences in the process used can limit the ability to combine data for later analysis.

The nature of the technology used will depend upon testing required and the planned and possible future use of the data.

*Table 1: Range of technologies and options for analysis*

| **Technology used** | **Scope of analysis available** | | | | |
|---|---|---|---|---|---|
| | **Single gene** | **Set of genes** | **Clinical genes (mendeliome)** | **All genes** | **Genes & non-gene regions** |
| **Single gene test** | ☑ | | | | |
| **Targeted panel sequencing** | ☑ | ☑ | | | |
| **Clinical exome sequencing** | ☑ | ☑ | ☑ | | |
| **Exome sequencing** | ☑ | ☑ | ☑ | ☑ | |
| **Genome sequencing** | ☑ | ☑ | ☑ | ☑ | ☑ |

| **Outputs** | |
|---|---|
| Read data files | The read data is the raw data generated by the sequencing technology used. The most common file format is FASTQ. FASTQ format is a text-based format for storing both a biological sequence and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character |

| | |
|---|---|
| | for brevity. It is usually constructed from the raw sequence data as part of the first part of the pipeline process.<br>A whole genome sequence can be approximately 150Gb in compressed format, but if the study is for a cancer (for instance), an additional two somatic sequences may be taken in addition to the germline sequence. |

## A.8. Pipeline process (shared)

Bioinformatic analyses invariably involve shepherding files through transformations, called a bioinformatics analysis pipeline or a workflow. Typically, these transformations are done by third-party executable command line software written for UNIX-compatible operating systems. Examples of these workflow management systems or orchestration tools are DNAnexus [156] and Cromwell [157].

Massively parallel sequencing, in which millions of short DNA sequences are the source input for interpreting a range of biological phenomena, has intensified the need for robust pipelines. Analyses involve steps such as sequence alignment and genomic annotation that are both time-intensive and parameter-heavy [158].

Pipelines used in clinical settings need to be accredited for diagnostic use and are relatively stable. In research environments, the pipeline code may be modified regularly as the research progresses, increasing the need and complexity of the source code management.

Developing pipeline code is usually undertaken by bioinformaticians or medical scientists.

| Inputs | |
|---|---|
| Workflow/pipeline code | An orchestration engine is required that executes pipeline commands. Some form of orchestration language is used to define pipeline commands, which the engine then translates to the appropriate vendor-specific commands.<br>These workflow files are usually held in a code repository and retrieved by the orchestration engine during execution. Change management of workflow code is critical metadata to be recorded as part of the provenance of the result. |
| **Outputs** | |
| Process metadata | Throughout the steps of the pipeline process, metadata must be recorded to support the provenance of the resulting genomic data. The metadata (or audit logs if you will) of the process can be significant data streams. |

### A.8.1. Alignment (shared)

Sequence alignment refers to merging fragments from a longer DNA sequence to reconstruct the original sequence. DNA sequencing technology cannot read whole genomes at once instead reading small pieces of between 20 and 30 000 bases, depending on the technology used. The short fragments, called reads, result from shotgun sequencing genomic DNA [159].

The process of alignment is largely automated by pipeline processes under the supervision of bioinformaticians or medical scientists.

| Inputs | |
|---|---|
| Read data files | As per the output from A.7. |
| Reference genome | A reference genome (also known as a reference assembly) is a digital nucleic acid sequence database, assembled by scientists as a representative example of a species's set of genes. |

| | |
|---|---|
| | Changes in the reference genome can cause data that cannot be directly compared with data generated using previous references. Updated reference genomes are one cause for re-analysing historical data.<br>Of note is whether the current reference genome applies in the Australian context given multiculturalism and the diverse genetic nature of the Aboriginal and Torres Strait Islander population. |
| **Outputs** | |
| Aligned/assembled sequences | Once the raw sequence data has been aligned, it is stored for later analysis. The most common file format is BAM which is a compressed format.<br>A BAM file is the binary version of a SAM file. A SAM file is a tab-delimited text file that contains sequence alignment data. These formats are described on the SAM Tools web site [160]. BAM, rather than SAM, is the recommended format for storage.<br>However, a new format (CRAM) has been developed for aligned sequence storage which has a smaller footprint but is compressed. This compression is driven by the reference the sequence data is aligned to. When a lossy compression algorithm is selected, a reduction in the base quality scores can occur [161]. |

### A.8.2. Variant calling/counting (shared)

Having obtained a valid sequence for the sample, those elements that vary from the reference genomic patterns must be identified. Known as variant calling, this process results in data in a format known as VCF.

For certain types of studies, variant counting is an additional process that can be completed.

The process of variant calling is largely automated by pipeline processes under the supervision of bioinformaticians or medical scientists.

| **Inputs** | |
|---|---|
| Aligned/assembled sequences | Variant calling will use the BAM data files (or equivalent) as the source. |
| Reference genome | The reference genome is the baseline against which variants are identified. |
| Other external sources | The process of variant calling also relies on other data sources, including:<br>• known biological annotations of polymorphisms (e.g. dbSNP [162] or the GATK resource bundle [163])<br>• annotations of sequence features required for analysis (e.g. using VEP [164]). |
| **Outputs** | |
| VCF files | The Variant Call Format specifies the format of a text file used in bioinformatics for storing gene sequence variations.<br>The format has been developed with large-scale genotyping and DNA sequencing projects, such as the 1000 Genomes Project.<br>Most of the annotation and curation process is based on the VCF data. |

### A.8.3. Variant annotation (shared)

Variant annotation is assigning information to variants. Many types of information could be associated with variants, from measures of sequence conservation to predictions about the effect of a variant on protein structure and function.

Variant annotation is a crucial step in the analysis of genome sequencing data and involves lookups to databases to gather evidence to inform variant curation and classification. Annotation results can have a strong influence on the classification of variants in the next stage of the process and the final conclusions made during the interpretation stage. Incorrect or incomplete annotations can cause researchers both to overlook potentially disease-relevant variants and to dilute potential variants in a pool of false positives [165]–[167].

The establishment of known variant databases that include information about variants is key to this process. These databases serve as an information storehouse of variants for pathologists and researchers.

The process of annotation is largely automated by pipeline processes under the supervision of bioinformaticians or medical scientists.

| **Inputs** | |
|---|---|
| VCF files | The VCF file is used and updated by later stages of the analysis process, including annotation. |
| Known variant databases | A list of known variants is used to inform the curation process. Known variants allow the pathologist to identify variants known to cause or not cause pathological outcomes.<br>The nature of the variants will depend upon whether germline or somatic variants are being considered. There are several sources for such known variants, depending upon the study under way. Examples include:<br>• **ClinVar** is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence [168].<br>• **COSMIC**, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer [169].<br>• **Shariant** is a technology from AGHA which is a controlled access variant hub and communication platform for real-time sharing of expertise and detailed scientific evidence about clinically curated variants. This is being trialled between Australian genomic sequencing laboratories and clinical services [102].<br>• The **Database of Genomic Variants** provides a useful catalogue of control data for studies aiming to correlate genomic variation with phenotypic data. The database is continuously updated with new data from peer reviewed research studies [170].<br>• **gnomAD:** aggregates data from many studies and provides population frequency and other data [171], [172].<br>• **dbSNP:** Published by National Center for Biotechnology Information, dbSNP contains human single nucleotide variations, microsatellites and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations [162].<br>• **OMIM:** A catalogue of human genes and genetic disorders managed by Johns Hopkins University [173].<br>• **PubMed:** An online resource of biomedical literature produced by the US National Library of Medicine [174]. |
| **Outputs** | |
| Annotated VCF files | Additional columns of data will be added to the VCF files through the annotation process. |

## A.9. Variant classification (shared)

A critical step undertaken by genetic pathologists (and others) is the curation and classification of identified variants in the sequence being studied. Not all variants are pathogenic or relevant to the current study, and the genetic pathologist must select from identified variants to determine those relevant.

This process is not automated, as it requires research, informed value judgements and experience by the genetic pathologist (and others) based on extant knowledge and guidelines. The process of reviewing publications for such information is called document triage and is time-consuming [166]. Curation of data occurs according to defined standards for this process [175].

While manual curation is the gold standard method for curation of variants, it can be time-consuming on a large scale. Sharing variant classification information through repositories such as Shariant [102] is important to reduction in effort and increasing the consistency of curation processes across Australia.

Automation through machine learning to support the prioritisation of variants for consideration has been suggested as a possible way of addressing these concerns [165], [167].

| **Inputs** | |
|---|---|
| VCF files | The VCF file is used and updated by later stages of the analysis process, including annotation. |
| Publications | During the curation process, the genetic pathologist will seek input from a variety of sources, one of which may be contemporary papers published in journals. These may provide information about recent studies of related or similar disease cases, which may indicate the presence or absence of genetic factors. |
| **Outputs** | |
| Curated results | The outcome of the process is a set of curated results with information on variant classification. While the data is based on the VCF file formats, the results may be rendered in a variety of electronic formats within the LIMS, before inclusion in the genomic test report. |

## A.10. Interpretation and reporting (shared)

In genomic medicine, the clinical interpretation of the results is critical to the delivery of a diagnostic report, prepared by a genetic pathologist or clinical geneticist. Sometimes, multidisciplinary teams (MDTs) are used to support the interpretation process. This report and its interpretation are important to:

- help to make/refine a diagnosis
- affect further testing, treatment plans and management strategies
- reveal patterns of inheritance and assess likelihood of genetic disease in relatives
- highlight need for specialist referral
- correct any family misconceptions.

Note that sometimes further analysis may be required, including functional studies, to finalise a report and/or diagnosis. In genomic research, the interpretation process will assess the results against the initial research hypothesis and prepare for later publication of the research outcomes.

| **Inputs** | |
|---|---|
| Curated results | The curated results will be the input to the interpretation stage. |

| **Outputs** | |
|---|---|
| Diagnostic report (clinical) | For genomic medicine, the most common result is the diagnostic report. This may be provided with atomic data to the LIMS but is commonly stored as PDF files or other text attachment files. |
| Research results (research) | For genomic research, the results will most commonly be held in a research data management system. Ideally these are available externally to support later publication. |

## A.11. Consultation/decision-making (genomic medicine)

If genomic testing provides or confirms a diagnosis, clinicians can use this information to consider treatment plans or patient management plans if treatment is not possible. These can then be discussed with the patient to agree a suitable path forward. Further genetic counselling may be required or referral to a specialist clinician.

| **Inputs** | |
|---|---|
| Diagnostic report | Received from the Pathology LIMS for from an external laboratory, this may be loaded to the EHR or the Genomic system used by the Genomic Counsellors (where appropriate). |
| **Outputs** | |
| Agreed patient management plans | Aside from providing a means of retrieving information about a particular patient, the goals of the clinical management plan are to guide clinical problem-solving and care planning, and to enable this plan to be communicated to other health professionals (if necessary). |
| Agreed treatment plans (where appropriate) | Where viable therapies exist, the treatment plan is a documented record of all major aspects of individual patients planned therapy and is an essential reference and communication resource for the patient and all healthcare professionals involved in the patient's care.<br>The treatment plan may be the consensus outcome of a multidisciplinary meeting discussion and reflects decisions made around therapy. The plan should reflect the intent of the treatment and requirements in relation to nursing, allied health and palliative care.<br>Many EHRs provide the capability to record treatment plans. |

## A.12. Management (genomic medicine)

Based upon the clinical and/or genetic counselling, management of the patient would then begin. Where therapies exist, this will include treatment of the patient. Sometimes, results will also inform discussions with family members about their disease risk profiles.

Additional aspects of further management (besides treatment options) include addition of surveillance processes where needed or stopping surveillance where it is unnecessary, and restoration of reproductive management.

| **Inputs** | |
|---|---|
| Treatment/management plans | The treatment or management plans agreed between clinician and patient support ongoing management. |

## A.13. Cohort selection (genomics research)

With genomic research (and translational genomics), the focus is frequently on a cohort of individuals, rather than a single individual, although where testing cannot provide a diagnosis, research activities may be appropriate for a single patient.

Often the data will personal identifiers removed so the researchers involved cannot identify specific individuals.

A research question is identified that may benefit from genomic analysis. This may result from an area of expertise by the investigators or because of a cohort of patients for whom genomic testing has not adequately identified a diagnosis.

Note that sometimes new samples are not required as the process involves re-analysis of existing genomic sequence data.

| **Inputs** | |
|---|---|
| Patient phenotypes | Phenotype data determined from the patient history will support the researchers in looking for patterns of variants that match particular phenotypes. This data may be de-identified. |
| **Outputs** | |
| Identified cohort | The result of the cohort selection process will be a set of inclusion and exclusion criteria to be used to determine what patient data is to be included in the study.<br>Where an existing pool of patient genomic data exists already, the cohort may be defined by defining subsets of genes/variants to be examined during the bioinformatics analysis. |

## A.14. Reprocessing/preparation (genomics research)

For genomic research, the bioinformatics analysis process is followed by a complex set of activities including aggregation, reprocessing, analytics and variations such as patient matching, cohort combining, deposition of data into a post-publication repository and the associated quality control and collation of sequencing metadata.

| **Inputs** | |
|---|---|
| Diagnostic analysis | With research studies, the output will be a set of data for the cohort, which is analysed against the research question. |
| **Outputs** | |
| Evidence datasets | It is good practice to make the datasets used as evidence to be available in some form so independent analysis can be made of the findings. |

## A.15. Publication (genomics research)

For genomic research, the results are then prepared for publication. This is the primary method of increasing the knowledge base for genomics.

Publication of results also occurs regarding clinical activities, not just pure research. This is true of translational research but can also apply to clinical cases studies. This reflects that many clinicians also hold academic affiliations.

| **Inputs** | |
|---|---|
| Diagnostic analysis | With research studies, the output will be a set of data for the cohort, which is analysed against the research question. |
| **Outputs** | |
| Paper for submission | A draft paper for submission to appropriate publications, outlining the findings and evidence. |
| Evidence datasets | It is good practice to make the datasets used as evidence to be available in some form so independent analysis can be made of the findings. Note that making data accessible (as per the FAIR principles) does not equate to the data being openly accessible. Data may be available online via data deposit repositories or require data access requests to the research institution.<br>However, there are challenges regarding the availability of data, including:<br>• the requirements around submitting to EGA type repositories [127]<br>• collecting the needed metadata<br>• the cost/time/resources to upload large genomic data sets<br>• the existence of a Data Access committee to control later requests for access [176] |
| Pipeline code | The code used to support analysis is also provided sometimes (subject to intellectual property considerations). |
| Metadata | To comply with the FAIR Findable principle, data and other artefacts about the project need to be locatable by others through the publication of metadata. Services such as Research Data Australia exist to support these activities [177]. |

# Appendix B. Glossary of Terms & Abbreviations

| Term/abbreviation | Description |
|---|---|
| **ACMG** | American College of Medical Genetics and Genomics |
| **AEHRC** | Australian E-Health Research Centre |
| **AHMAC** | Australian Health Ministers Advisory Council |
| **AIATSIS** | Australian Institute of Aboriginal and Torres Strait Islander Studies |
| **AIHW** | Australian Institute of Health and Welfare |
| **Alleles** | Every person has two copies of each gene, one inherited from each parent. Most genes are the same in all people, but a few genes (less than 1 per cent of the total) are slightly different between people. Alleles are forms of the same gene with small differences in their sequence of DNA bases. These small differences contribute to each person's unique physical features [129]. |
| **API** | Application Programming Interface |
| **APP** | Australian Privacy Principle |
| **ARDC** | Australian Research Data Commons |
| **Autosomes** | In humans, each cell normally contains 23 pairs of chromosomes. Twenty-two of these pairs, called autosomes, look the same in both males and females. The twenty-third pair, the sex chromosomes, differ between males and females. Females have two copies of the X chromosome, while males have one X and one Y chromosome.<br>The 22 autosomes are numbered by size. The other two chromosomes, X and Y, are the sex chromosomes. The picture (at right) of the human chromosomes lined up in pairs is called a karyotype [178]. |
| **AWS** | Amazon Web Services |
| **Bases** | The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and over 99 per cent of those bases are the same in all people. The order, or sequence, of these bases determines the information available for building and maintaining an organism, like how letters of the alphabet appear in a certain order to form words and sentences [178]. |
| **Biochemical assays** | An analytical procedure to detect and quantify cellular processes (e.g. apoptosis, cell signalling) or metabolic reactions. Biochemical assays are a reliable, routinely used procedure that helps in characterising targets and understanding of biomolecular functions [178]. |
| **Bioinformatician** | A person who uses data algorithms and specialised software to analyse biological data, such as DNA or RNA sequences. |
| **Bioinformatics** | The use of algorithms and software to analyse biological data. |

| Term/abbreviation | Description |
|---|---|
| **CARE** | The CARE Principles for Indigenous Data Governance are people and purpose-oriented, reflecting the crucial role of data in advancing Indigenous innovation and self-determination. These principles complement the existing FAIR principles encouraging open and other data movements to consider both people and purpose in their advocacy and pursuits. |
| **Carrier testing** | Carrier testing is used to identify people who carry one copy of a gene mutation that, when present in two copies, causes a genetic disorder. This testing is offered to individuals with a family history of a genetic disorder and to people in certain ethnic groups with an increased risk of specific genetic conditions. If both parents are tested, the test can provide information about a couple's risk of having a child with a genetic condition [129]. |
| **CDA** | Clinical Document Architecture |
| **CDM** | Common Data Model |
| **Centromere** | Each chromosome has a constriction point called the centromere, which divides the chromosome into two sections, or ‘arms’. The short arm of the chromosome is labelled the ‘p arm’. The long arm of the chromosome is labelled the ‘q arm.’ The location of the centromere on each chromosome gives the chromosome its characteristic shape and can help describe the location of specific genes [129]. |
| **Chromosome** | In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called chromosomes. Each chromosome comprises DNA tightly coiled often around proteins called histones that support its structure. [129] |
| **Clinical geneticist** | Physicians who have undergone speciality training in genetics after general professional training (such as paediatrics and oncology) and see referred patients for diagnosis, management, genetic testing and genetic counselling. |
| **Clinical genetics** | The medical specialty which provides a diagnostic service and "genetic counselling" for individuals or families with, or at risk of, conditions which may have a genetic basis. |
| **COAG** | Council of Australian Governments |
| **CSIRO** | Commonwealth Scientific and Industrial Research Organisation |
| **DAC** | Data Access Committee |
| **DAM** | Data Access Management |
| **Diagnostic testing** | Diagnostic testing is used to identify or rule out a specific genetic or chromosomal condition. Often, genetic testing is used to confirm a diagnosis when a condition is suspected based on physical signs and symptoms. Diagnostic testing can be performed before birth or during a person's life but is not available for all genes or all genetic conditions. The results of a diagnostic test can influence a person's choices about health care and the management of the disorder [129]. |
| **DMBOK** | Data Management Body of Knowledge |
| **DNA** | Deoxyribonucleic acid is the hereditary material in humans and most other organisms. Nearly every cell in a person’s body has the same DNA. Most DNA is in the cell nucleus (where it is called nuclear DNA), but a little DNA can also be found in the mitochondria (where it is called mitochondrial DNA or mtDNA) [129]. |
| **DNA sequencing** | Determining the order of DNA building blocks (nucleotides) in an individual's genetic code, called DNA sequencing, has advanced the study of genetics and is one technique used to test for genetic disorders [129]. |

| Term/abbreviation | Description |
|---|---|
| **DTA** | Digital Transformation Agency |
| **DUO** | Data Use Ontology |
| **EGA** | European Genome-phenome Archive |
| **EHR** | Electronic Health Record |
| **ELSI** | Ethical, legal and social issues |
| **Epigenetics** | The study of changes in organisms caused by modification of gene expression rather than alteration of the genetic code itself. |
| **Exome** | Part of the genome formed by exons, the sequences which, when transcribed remain within the mature RNA after introns are removed by RNA splicing. |
| **Exome sequencing** | This method allows variations in the protein-coding region of any gene to be identified, rather than in only a select few genes. Because most known mutations that cause disease occur in exons, exome sequencing is thought to be an efficient method to identify possible disease-causing mutations [129]. |
| **FAIR** | The FAIR Data Principles are a set of guiding principles in order to make data findable, accessible, interoperable and reusable. |
| **FHIR** | Fast Healthcare Interoperability Resources |
| **FISH** | Fluorescence in situ hybridisation is a molecular testing method that uses fluorescent probes to evaluate genes and/or DNA sequences on chromosomes. This technique can show extra gene copies (duplicated or amplified genes), and genetic sequences missing (gene deletions) or have been moved (translocated genes) [129]. |
| **Forensic testing** | Forensic testing uses DNA sequences to identify an individual for legal purposes. Unlike the tests described above, forensic testing is not used to detect gene mutations associated with disease. This testing can identify crime or catastrophe victims, rule out or implicate a crime suspect, or establish biological relationships between people (for example, paternity) [129]. |
| **GA4GH** | Global Alliance for Genomics and Health |
| **GDPR** | The General Data Protection Regulation is a regulation in EU law on data protection and privacy in the European Union and the European Economic Area. It also addresses the transfer of personal data outside the EU areas. |
| **Gene** | A gene is the basic physical and functional unit of heredity. Genes comprise DNA. Some genes act as instructions to make molecules called proteins. However, many genes do not code for proteins. In humans, genes vary in size from a few hundred DNA bases to over 2 million bases. The Human Genome Project estimated that humans have between 20 000 and 25 000 genes [129]. |
| **Gene names** | Scientists keep track of genes by giving them unique names. Because gene names can be long, genes are also assigned symbols, which are short combinations of letters (and sometimes numbers) that represent an abbreviated version of the gene name. For example, a gene on chromosome seven associated with cystic fibrosis is called the cystic fibrosis transmembrane conductance regulator; its symbol is CFTR [129]. |
| **Genetic counsellor** | Healthcare professionals who have undergone speciality training to help individuals, couples and families understand and adapt to the medical, psychological, familial and reproductive implications of the genetic contribution to specific health conditions. |

| Term/abbreviation | Description |
|---|---|
| **Genetic pathologist** | Pathologists who have undergone speciality training in genetics and genomics who provide specialist input to genomic tests including test selection, results and interpretation to aid in the diagnosis, management and treatment of patients with a genetic basis for their disease. |
| **Genetic testing** | Genetic testing is a type of medical test that identifies changes in chromosomes, genes or proteins. The results of a genetic test can confirm or rule out a suspected genetic condition or help determine a person's chance of developing or passing on a genetic disorder. Over 1000 genetic tests are in use, and more are being developed [129].<br>Several methods can be used for genetic testing:<br>• Molecular genetic tests (or gene tests) study single genes or short lengths of DNA to identify variations or mutations that lead to a genetic disorder.<br>• Chromosomal genetic tests analyse whole chromosomes or long lengths of DNA to see if there are large genetic changes, such as an extra copy of a chromosome, that cause a genetic condition.<br>• Biochemical genetic tests study the amount or activity level of proteins; abnormalities in either can indicate changes to the DNA that result in a genetic disorder. |
| **Genome** | The complete set of genetic information in an organism. |
| **Genome sequencing** | Researchers have found that DNA variations outside the exons can affect gene activity and protein production and lead to genetic disorders - variations that exome sequencing would miss. Genome sequencing determines the order of all the nucleotides in an individual's DNA and can determine variations in any part of the genome [129]. |
| **Genomic data** | Refers to data produced from DNA sequencing of a genome. It can be compared with a reference genome. |
| **Genomic knowledge** | Includes information about the interpretation of genomic data and the implications of these findings, as well as relevant non-genomic clinical information. |
| **Genomic medicine** | Is an emerging medical discipline that involves using genomic information about an individual as part of their clinical care (e.g., for diagnostic or therapeutic decision-making) and the health outcomes and policy implications of that clinical use (also used interchangeably with precision medicine, personalised medicine, stratified medicine). |
| **Genotype** | A genotype is an individual's collection of genes. |
| **Germline cell** | The reproductive cells in multicellular organisms. |
| **GFED** | Functional Genomics Data Society |
| **GIDA** | Global Indigenous Data Alliance |
| **GRCh38** | Homo sapiens (human) genome assembly GRCh38 (hg38) from Genome Reference Consortium. |
| **Haematological assays** | A haematology test is a measurement of blood to help diagnose and monitor many conditions. |
| **HGVS** | Human Genome Variation Society |
| **HIMMS** | Healthcare Information and Management Systems Society |
| **HPO** | Human Phenotype Ontology |

| Term/abbreviation | Description |
|---|---|
| **ICD** | International Classification of Diseases |
| **IHE** | Integrating the Healthcare Environment |
| **Immunological assays** | An immunoassay is a biochemical test that measures the presence or concentration of a macromolecule or a small molecule in a solution using an antibody or an antigen [129]. |
| **Incidental findings** | While many more genetic changes can be identified with exome and genome sequencing than with select gene sequencing, the significance of much of this information is unknown. Because not all genetic changes affect health, it is difficult to know whether identified variants are involved in the condition of interest. Sometimes, an identified variant is associated with a different genetic disorder not yet diagnosed (these are called incidental or secondary findings) [129]. |
| **Introns** | Some non-coding DNA regions, called introns, are within protein-coding genes but are removed before a protein is made. Regulatory elements, such as enhancers, can be in introns. Other non-coding regions are found between genes and are known as intergenic regions [129]. |
| **Karyotyping** | Chromosome analysis or karyotyping is a test that evaluates the number and structure of a person's chromosomes to detect abnormalities. A karyotype examines a person's chromosomes to determine if the right number is present and to determine if each chromosome appears normal [129]. |
| **LIMS** | Laboratory Information Management Systems |
| **Metabolomics** | Metabolomics is the scientific study of chemical processes involving metabolites, the small molecule substrates, intermediates and products of metabolism. |
| **Metadata** | A set of data that describes and gives information about other data. |
| **Metagenomics** | Metagenomics is the study of genetic material recovered directly from environmental samples. The broad field may also be called environmental genomics, ecogenomics or community genomics [129]. |
| **Microarray** | Microarray testing is used for a wide variety of purposes. In diagnostic testing it is primarily used to test for the presence in the patient’s DNA (their genome) of either tiny missing sections (called microdeletions) or extra duplicated sections (called microduplications). Microarray testing is more sensitive than conventional chromosome analysis, called cytogenetics or karyotyping. Although both can examine all chromosomes, microarray testing can detect small changes that cannot be seen using a microscope [129]. |
| **Mitochondria** | Mitochondria are structures within cells that convert the energy from food into a form that cells can use [129]. |
| **MRFF** | Medical Research Future Fund |
| **NAGIM** | National Approach to Genomics Information Management |
| **NATA** | National Association of Testing Authorities |
| **NBA** | National Blood Authority |
| **NCI** | National Computational Infrastructure |
| **NCIG** | National Centre for Indigenous Genomics |
| **NCRIS** | National Collaborative Research Infrastructure Strategy |

| Term/abbreviation | Description |
|---|---|
| **NDIS** | National Disability Insurance Scheme |
| **Newborn screening** | Newborn screening is used just after birth to identify genetic disorders that can be treated early in life. Millions of babies are tested each year in the United States. All states test infants for phenylketonuria (a genetic disorder that causes intellectual disability if left untreated) and congenital hypothyroidism (a disorder of the thyroid gland). Most states also test for other genetic disorders [129]. |
| **Next-Generation Sequencing** | NGS, also known as high-throughput sequencing, is the catch-all term used to describe several modern sequencing technologies. These technologies allow for sequencing of DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, and as such revolutionised the study of genomics and molecular biology [129]. |
| **NGS** | See Next-Generation Sequencing |
| **NHGPF** | National Health Genomics Policy Framework |
| **NHMRC** | National Health and Medical Research Council |
| **Non-coding DNA** | Only about 1 per cent of DNA comprises protein-coding genes; the other 99 per cent is non-coding. Non-coding DNA does not provide instructions for making proteins. Once thought 'junk' with no known purpose, it is becoming clear that at least some of it is integral to the function of cells, particularly the control of gene activity.<br>For example, non-coding DNA contains sequences that act as regulatory elements, determining when and where genes are turned on and off. Such elements provide sites for specialised proteins (called transcription factors) to attach (bind) and either activate or repress the process by which the information from genes is turned into proteins (transcription) [129]. |
| **Nucleotides** | DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. The structure of the double helix is somewhat like a ladder, with the base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder [129]. |
| **OAIC** | Office of the Australian Information Commissioner |
| **'omics** | Suffix that refers to the analysis of all the molecules of one type in a cell or tissue. For example, genomics (investigation of all the DNA molecules in a cell), transcriptomics (all RNA molecules), proteomics (all proteins). |
| **ONDC** | Office of the National Data Commissioner |
| **Personalised medicine** | Personalised medicine (also known as stratified or precision medicine) uses this knowledge of genetics to predict disease development, to influence decisions about lifestyle choices or to tailor treatment to an individual. |
| **Pharmacogenetics** | The study of how the actions of, and reactions to, medicines vary with the patient's genes. |
| **Phenotype** | Phenotype is the term used in genetics for the composite observable characteristics or traits of an organism. The term covers the organism's morphology or physical form and structure, its developmental processes, its biochemical and physiological properties, its behaviour and the products of behaviour [129]. |

| Term/abbreviation | Description |
|---|---|
| **PII** | Personally identifiable information |
| **Precision medicine** | Precision medicine is an approach to patient care that allows doctors to select treatments most likely to help patients based on a genetic understanding of their disease. This may also be called personalised medicine. |
| **Predictive and pre-symptomatic testing** | Predictive and presymptomatic types of testing are used to detect gene mutations associated with disorders that appear after birth, often later in life. These tests can help people with a family member with a genetic disorder, but who have no features of the disorder themselves at the time of testing. Predictive testing can identify mutations that increase a person's risk of developing disorders with a genetic basis, such as certain types of cancer. Pre-symptomatic testing can determine whether a person will develop a genetic disorder, such as hereditary hemochromatosis (an iron overload disorder), before any signs or symptoms appear. The results of predictive and pre-symptomatic testing can provide information about a person’s risk of developing a specific disorder and help with deciding about medical care [129]. |
| **Preimplantation testing** | Preimplantation testing, also called preimplantation genetic diagnosis, is a specialised technique that can reduce the risk of having a child with a genetic or chromosomal disorder. It is used to detect genetic changes in embryos created using assisted reproductive techniques such as in vitro fertilisation. In vitro fertilisation involves removing egg cells from a woman’s ovaries and fertilising them with sperm cells outside the body. To perform preimplantation testing, a few cells are taken from these embryos and tested for certain genetic changes. Only embryos without these changes are implanted in the uterus to initiate a pregnancy [129]. |
| **Prenatal testing** | Prenatal testing is used to detect changes in a foetus’ genes or chromosomes before birth. This testing is offered during pregnancy if there is an increased risk that the baby will have a genetic or chromosomal disorder. Sometimes, prenatal testing can lessen a couple's uncertainty or help them decide about a pregnancy. It cannot identify all possible inherited disorders and birth defects [129]. |
| **PRGHG** | Project Reference Group on Human Genomics |
| **Proteomics** | Proteomics is the large-scale study of proteins. The proteome is the entire set of proteins produced or modified by an organism or system. |
| **RDMS** | Research Data Management System |
| **Replication** | An important property of DNA is that it can replicate or make copies of itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical when cells divide because each new cell needs to have an exact copy of the DNA present in the old cell [129]. |
| **RNA** | Regions of non-coding DNA provide instructions for the formation of certain kinds of RNA molecules. Examples of specialised RNA molecules produced from non-coding DNA include transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), which help assemble protein building blocks (amino acids) into a chain that forms a protein; microRNAs (miRNAs), which are short lengths of RNA that block the process of protein production; and long non-coding RNAs (lncRNAs), which are longer lengths of RNA with diverse roles in regulating gene activity [129]. |

| Term/abbreviation | Description |
|---|---|
| **Sanger sequencing** | The original sequencing technology, called Sanger sequencing (named after the scientist who developed it, Frederick Sanger), was a breakthrough that helped scientists determine the human genetic code, but it is time-consuming and expensive. The Sanger method has been automated to make it faster and is still used in laboratories today to sequence short pieces of DNA, but it would take years to sequence all a person's DNA (known as the person's genome). Next-generation sequencing has sped up the process (taking only days to weeks to sequence a human genome) while reducing the cost [129]. |
| **Secondary findings** | See Incidental findings |
| **SNP** | Single nucleotide polymorphisms |
| **SNV** | Single-nucleotide variant |
| **Somatic cell** | Derived from the Greek word soma, meaning “body”. Hence, all body cells of an organism – apart from the sperm and egg cells, the cells from which they arise (gametocytes) and undifferentiated stem cells – are somatic cells. Examples of somatic cells are cells of internal organs, skin, bones, blood and connective tissues. In comparison, the somatic cells contain a full set of chromosomes whereas the reproductive cells contain only half. |
| **Telomeres** | Some structural elements of chromosomes are also part of non-coding DNA. For example, repeated non-coding DNA sequences at the ends of chromosomes form telomeres. Telomeres protect the ends of chromosomes from being degraded during copying genetic material. Repetitive non-coding DNA sequences also form satellite DNA, which is a part of other structural elements. Satellite DNA is the basis of the centromere, which is the constriction point of the X-shaped chromosome pair. Satellite DNA also forms heterochromatin, which is densely packed DNA important for controlling gene activity and maintaining the structure of chromosomes [129]. |
| **TOGAF** | The Open Group Architecture Framework |
| **Transcriptomics** | The transcriptome is the set of all RNA transcripts, including coding and non-coding, in an individual or a population of cells. The term can also sometimes refer to all RNAs, or just mRNA, depending on the experiment [129]. |
| **UNDRIP** | United Nations Declaration on the Rights of Indigenous Peoples |

# Appendix C. References

[1] Australian Health Ministers' Advisory Council, "National Health Genomics Policy Framework," Australian Government Department of Health, 2017 [Online]. Available: https://www1.health.gov.au/internet/main/publishing.nsf/Content/national-health-genomics-policy-framework-2018-2021. [Accessed: 22-Jan-2020]

[2] "Implementation Plan National Health Genomics Policy Framework: driving national action 2018-2021.," Australian Government Department of Health, 2018 [Online]. Available: https://www1.health.gov.au/internet/main/publishing.nsf/Content/national-health-genomics-policy-framework-2018-2021

[3] "WHO definitions of genetics and genomics," 2016. [Online]. Available: https://www.who.int/genomics/geneticsVSgenomics/en/. [Accessed: 15-Feb-2020]

[4] R. L. Ackoff, "From Data to Wisdom," Journal of Applies Systems Analysis, vol. 16, pp. 3–9, 1989.

[5] "Architecture Principles," TOGAF 9.2, 2018. [Online]. Available: https://pubs.opengroup.org/architecture/togaf92-doc/arch/chap20.html. [Accessed: 05-Jun-2020]

[6] "National Statement on Ethical Conduct in Human Research (2007) - Updated 2018," National Health and Medical Research Council, 2018 [Online]. Available: https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018. [Accessed: 16-Jun-2020]

[7] "Guidelines for Ethical Research in Indigenous Studies: May 2000," Australian Institute of Aboriginal and Torres Strait Islander Studies, Canberra, 2000 [Online]. Available: https://aiatsis.gov.au/research/ethical-research/guidelines-ethical-research-australian-indigenous-studies. [Accessed: 17-Jun-2020]

[8] "CARE Principles of Indigenous Data Governance," 2018. [Online]. Available: https://www.gida-global.org/care. [Accessed: 22-Jan-2020]

[9] "Principles for the translation of omics-based tests," National Health and Medical Research Council, 2015 [Online]. Available: https://www.nhmrc.gov.au/about-us/publications/principles-translation-omics-based-tests. [Accessed: 18-May-2020]

[10] M. Smith, R. Saunders, L. Stuckhardt, J. M. McGinnis, C. on the L. H. C. S. in America, and I. of Medicine, "A Continuously Learning Health Care System," in Best Care at Lower Cost: The Path to Continuously Learning Health Care in America, J. M. M. Mark Smith, Robert Saunders, Leigh Stuckhardt, Ed. National Academies Press (US), 2013, p. 450.

[11] "OAIC Home Page." [Online]. Available: https://www.oaic.gov.au/. [Accessed: 16-Jun-2020]

[12] E. E. Kowal, "Genetic research in indigenous health: Significant progress, substantial challenges," Medical Journal of Australia, vol. 197, no. 1. pp. 19–20, 02-Jul-2012 [Online]. Available: https://www.mja.com.au/journal/2012/197/1/genetic-research-indigenous-health-significant-progress-substantial-challenges. [Accessed: 31-Jul-2020]

[13] N. A. Garrison et al., "Genomic Research Through an Indigenous Lens: Understanding the Expectations," Annual Review of Genomics and Human Genetics, vol. 20, pp. 495–517, 2019, doi: 10.1146/annurev-genom-083118-015434.

[14] G. Pratt et al., GENOMIC PARTNERSHIPS Guidelines for genomic research with Aboriginal and Torres Strait Islander peoples of Queensland Genomic Partnerships: Guidelines for genomic research with Aboriginal and Torres Strait Islander peoples of Queensland. 2019 [Online]. Available: www.qimrberghofer.edu.au. [Accessed: 31-Jul-2020]

[15] "United Nations Declaration on the Rights of Indigenous Peoples United Nations," United Nations, 2007 [Online]. Available: https://www.un.org/development/desa/indigenouspeoples/wp-content/uploads/sites/19/2018/11/UNDRIP_E_web.pdf. [Accessed: 11-Aug-2020]

[16] "Australian Privacy Principles," 2020. [Online]. Available: https://www.oaic.gov.au/privacy/australian-privacy-principles/. [Accessed: 16-Jun-2020]

[17] "Privacy in your state." [Online]. Available: https://www.oaic.gov.au/privacy/privacy-in-your-state/. [Accessed: 16-Jun-2020]

[18] "Privacy for health service providers." [Online]. Available: https://www.oaic.gov.au/privacy/privacy-for-health-service-providers/. [Accessed: 16-Jun-2020]

[19] "General Data Protection Regulation (GDPR) Compliance Guidelines," 2020. [Online]. Available: https://gdpr.eu/. [Accessed: 16-Jun-2020]

[20] "Australian entities and the EU General Data Protection Regulation (GDPR)," 2018. [Online]. Available: https://www.oaic.gov.au/privacy/guidance-and-advice/australian-entities-and-the-eu-general-data-protection-regulation/#who-will-the-gdpr-apply-to. [Accessed: 16-Jun-2020]

[21] "About us - ONDC," 2019. [Online]. Available: https://www.datacommissioner.gov.au/about/about-us. [Accessed: 16-Jun-2020]

[22] "Best Practice Guide to Applying Data Sharing Principles," Australian Government Department of Prime Minister and Cabinet, 2019.

[23] T. Desai, F. Ritchie, and R. Welpton, "Five Safes: designing data access for research," University of the West of England Bristol, 2016 [Online]. Available: https://uwe-repository.worktribe.com/output/914745/five-safes-designing-data-access-for-research. [Accessed: 19-May-2020]

[24] http://www.fivesafes.org/, "The Five Safes." [Online]. Available: http://www.fivesafes.org/. [Accessed: 18-May-2020]

[25] "The FAIR Data Principles," 2014. [Online]. Available: https://www.force11.org/group/fairgroup/fairprinciples. [Accessed: 22-Jan-2020]

[26] M. D. Wilkinson, "Comment: The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, 2016, doi: 10.1038/sdata.2016.18. [Online]. Available: https://www.nature.com/articles/sdata201618.pdf?origin=ppub. [Accessed: 22-Jan-2020]

[27] "The FAIR data principles," 2015. [Online]. Available: https://www.ands.org.au/working-with-data/fairdata. [Accessed: 22-Jan-2020]

[28] B. Knoppers et al., "Framework for Responsible Sharing of Genomic and Health-Related Data," 2014. [Online]. Available: https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/framework-for-responsible-sharing-of-genomic-and-health-related-data/#fp. [Accessed: 29-Jan-2020]

[29] "Genomic Data Policy Framework and Ethical Tensions," World Economic Forum, 2020 [Online]. Available: www.weforum.org. [Accessed: 23-Jul-2020]

[30] "Architecture classification | Queensland Government Enterprise Architecture," 2019. [Online]. Available: https://www.qgcio.qld.gov.au/information-on/digital-and-ict-strategic-planning/current-state/architecture-classification. [Accessed: 26-Jun-2020]

[31] et al Winkler, T, "Quality control and conduct of genome-wide association meta-analyses," Nature America, 2014 [Online]. Available: https://keppel.qimr.edu.au/contents/p/staff/Winkler_NatProtoc_1192-1212.pdf. [Accessed: 09-Jul-2020]

[32] C. Fitzgerald, "Next Generation Sequencing Quality," 2019 [Online]. Available: https://www.cdc.gov/labquality/ngs-quality-initiative. [Accessed: 19-Jun-2020]

[33] "The Next Generation Sequencing Quality Initiative | CDC." [Online]. Available: https://www.cdc.gov/labquality/ngs-quality-initiative.html. [Accessed: 19-Jun-2020]

[34] B. Jew and J. H. Sul, "Variant calling and quality control of large-scale human genome sequencing data," Emerging Topics in Life Sciences, vol. 3, no. 4, pp. 399–409, Aug. 2019, doi: 10.1042/etls20190007.

[35] "What are metadata and why are they so important?," 2015. [Online]. Available: https://www.ebi.ac.uk/training/online/course/ebi-metagenomics-portal-submitting-metagenomics-da/what-are-metadata-and-why-are-they-so-im-0. [Accessed: 02-Jul-2020]

[36] "VCF - Variant Call Format," 2020. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format. [Accessed: 11-Jun-2020]

[37] T. A. Grebe et al., "The interface of genomic information with the electronic health record: a points to consider statement of the American College of Medical Genetics and Genomics (ACMG)," doi: 10.1038/s41436-020. [Online]. Available: https://doi.org/10.1038/s41436-020-. [Accessed: 05-Jun-2020]

[38] S. A. Pendergrass and D. C. Crawford, "Using Electronic Health Records to Generate Phenotypes for Research," Curr Protoc Hum Genet, vol. 100, no. 1, p. 80, 2019, doi: 10.1002/cphg.80.

[39] M. K. Breitenstein, H. Liu, K. N. Maxwell, J. Pathak, and R. Zhang, "Electronic Health Record Phenotypes for Precision Medicine: Perspectives and Caveats From Treatment of Breast Cancer at a Single Institution," Clin Transl Sci, vol. 11, pp. 85–92, 2018, doi: 10.1111/cts.12514. [Online]. Available: www.cts-journal.com. [Accessed: 10-Jun-2020]

[40] "Monash Clinical Registries," 2020. [Online]. Available: https://www.monash.edu/medicine/sphpm/registries. [Accessed: 11-Jun-2020]

[41] D. Raths, "Cancer Registry to Link Genomic, Outcomes Data," 2015. [Online]. Available: https://www.hcinnovationgroup.com/home/article/13025961/cancer-registry-to-link-genomic-outcomes-data. [Accessed: 11-Jun-2020]

[42] W. Jiang, T. Z. King, and J. A. Turner, "Imaging genetics towards a refined diagnosis of schizophrenia," Frontiers in Psychiatry, vol. 10, no. JULY. Frontiers Media S.A., p. 494, 12-Jul-2019.

[43] R. lo Gullo, I. Daimiel, E. A. Morris, and K. Pinker, "Combining molecular and imaging metrics in cancer: radiogenomics," Insights into Imaging, vol. 11, no. 1, pp. 1–17, Dec. 2020, doi: 10.1186/s13244-019-0795-6.

[44] Z. Bodalal, S. Trebeschi, T. Dan, L. Nguyen-Kim, W. Schats, and R. Beets-Tan, "Radiogenomics: bridging imaging and genomics," vol. 44, pp. 1960–1984, 2028, doi: 10.1007/s00261-019-02028-w. [Online]. Available: https://doi.org/10.1007/s00261-019-02028-w. [Accessed: 11-Jun-2020]

[45] "Requirements for the Retention of Laboratory Records and Diagnostic Material (Seventh Edition 2018)," National Pathology Accreditation Advisory Council, 2018 [Online]. Available: https://www1.health.gov.au/internet/main/publishing.nsf/Content/health-npaac-docs-RetLabRecDI-2018. [Accessed: 19-May-2020]

[46] S. Richmond, L. J. Howe, S. Lewis, E. Stergiakouli, and A. Zhurov, "Facial genetics: A brief overview," Frontiers in Genetics, vol. 9, no. OCT, p. 462, Oct. 2018, doi: 10.3389/fgene.2018.00462.

[47] "Data sharing (or transfer) agreements: What are they and when do I need one? | Research," 2016. [Online]. Available: https://uwaterloo.ca/research/office-research-ethics/research-human-participants/pre-submission-and-training/human-research-guidelines-and-policies-alphabetical-list/data-sharing-or-transfer-agreements-what-are-they-and-when. [Accessed: 02-Jul-2020]

[48] "Data management plans," 2017. [Online]. Available: https://www.ands.org.au/working-with-data/data-management/data-management-plans. [Accessed: 02-Jul-2020]

[49] "Management of Data and Information in Research: A guide supporting the Australian Code for the Responsible Conduct of Research," National Health and Medical Research Council, 2019.

[50] C. M. O'Keefe, S. Otorepec, M. Elliot, E. Mackey, and K. O'Hara, "The De-Identification Decision-Making Framework," 2017 [Online]. Available: https://data61.csiro.au/en/Our-Research/Our-Work/Safety-and-Security/Privacy-Preservation/De-identification-Decision-Making-Framework. [Accessed: 18-Aug-2020]

[51] "The Australian health system | Australian Government Department of Health." [Online]. Available: https://www.health.gov.au/about-us/the-australian-health-system. [Accessed: 03-Jul-2020]

[52] "About the Australian Research Council," 2018. [Online]. Available: https://www.arc.gov.au/about-arc. [Accessed: 02-Jul-2020]

[53] E. Birney, J. Vamathevan, and P. Goodhand, "Genomics in healthcare: GA4GH looks to 2022," bioRxiv 203554, p. http://dx.doi.org/10.1101/203554, 2017, doi: 10.1101/203554. [Online]. Available: http://dx.doi.org/10.1101/203554. [Accessed: 03-Feb-2020]

[54] T. Keane, "Integrating healthcare and genomics to improve human disease outcomes across Europe," in Australasian Leadership Computing Symposium, 2019.

[55] J. Alper and C. Grossmann, Health System Leaders Working Toward High-Value Care. 2015 [Online]. Available: http://www.nap.edu. [Accessed: 05-Jun-2020]

[56] "Interview with Peter Goodhand, CEO GA4GH." 2020.

[57] "Global Data Access for Solving Rare Disease - A Health Economics Value Framework," World Economic Forum, Feb. 2020 [Online]. Available: www.weforum.org. [Accessed: 15-Jun-2020]

[58] C. Lee, "Not So Basic Research: the unrecognized importance of fundamental scientific discoveries - Science in the News," Harvard University - SITN Blog, 2019. [Online]. Available: http://sitn.hms.harvard.edu/flash/2019/not-so-basic-research-the-unrecognized-importance-of-fundamental-scientific-discoveries/. [Accessed: 31-Jul-2020]

[59] "National Microbial Genomics Framework 2019 – 2022," Australian Government Department of Health, 2019 [Online]. Available: https://www1.health.gov.au/internet/main/publishing.nsf/Content/national-microbial-genomics-framework-2019-2022. [Accessed: 03-Jul-2020]

[60] "Clinical genomics," 2020. [Online]. Available: https://www.aidrc.org.au/clinical-genomics. [Accessed: 03-Jul-2020]

[61] "Infectious Disease Genomics - NSW Health Pathology - Website." [Online]. Available: https://www.pathology.health.nsw.gov.au/clinical-services/genomics/infectious-disease-genomics. [Accessed: 03-Jul-2020]

[62] "Genomics | Doherty Institute | Doherty Website." [Online]. Available: https://www.doherty.edu.au/our-work/cross-cutting-disciplines/genomics. [Accessed: 03-Jul-2020]

[63] "Microbiological Diagnostic Unit Public Health Laboratory : School of Biomedical Sciences." [Online]. Available: https://biomedicalsciences.unimelb.edu.au/departments/microbiology-Immunology/research/services/microbiological-diagnostic-unit-public-health-laboratory. [Accessed: 21-Aug-2020]

[64] J. A. Zachman, "The Concise Definition of The Zachman Framework," 2008. [Online]. Available: https://www.zachman.com/about-the-zachman-framework. [Accessed: 02-Jul-2020]

[65] "TOGAF | The Open Group." [Online]. Available: https://www.opengroup.org/togaf. [Accessed: 02-Jul-2020]

[66] "The ArchiMate® Enterprise Architecture Modeling Language | The Open Group." [Online]. Available: https://www.opengroup.org/archimate-forum/archimate-overview. [Accessed: 02-Jul-2020]

[67] "How to Use Architecture Levels Effectively." [Online]. Available: https://www.orbussoftware.com/enterprise-architecture/archimate/how-to-use-architecture-levels-effectively/. [Accessed: 02-Jul-2020]

[68] Z. D. Stephens et al., "Big data: Astronomical or genomical?," PLoS Biology, vol. 13, no. 7, p. e1002195, Jul. 2015, doi: 10.1371/journal.pbio.1002195. [Online]. Available: https://dx.plos.org/10.1371/journal.pbio.1002195. [Accessed: 02-Jul-2020]

[69] "About cloud.gov.au," 2018. [Online]. Available: https://www.dta.gov.au/our-projects/about-cloudgovau. [Accessed: 02-Jul-2020]

[70] "National Computational Infrastructure," 2020. [Online]. Available: https://nci.org.au/. [Accessed: 30-Jun-2020]

[71] "National Collaborative Research Infrastructure Strategy (NCRIS)," 2019. [Online]. Available: https://www.education.gov.au/national-collaborative-research-infrastructure-strategy-ncris. [Accessed: 30-Jun-2020]

[72] "EGA European Genome-Phenome Archive," 2017. [Online]. Available: https://ega-archive.org/. [Accessed: 14-Apr-2020]

[73] "Genomics England | 100,000 Genomes Project." [Online]. Available: https://www.genomicsengland.co.uk/. [Accessed: 03-Feb-2020]

[74] "Definitions of Data Governance." [Online]. Available: http://www.datagovernance.com/adg_data_governance_definition/. [Accessed: 03-Jul-2020]

[75] "Data Governance Framework," Australian Institute of Health & Welfare, 2019 [Online]. Available: http://www.datagovernance.com/adg_data_governance_definition/,. [Accessed: 16-Jun-2020]

[76] "Information and data governance framework." [Online]. Available: https://www.naa.gov.au/about-us/our-organisation/accountability-and-reporting/information-and-data-governance-framework. [Accessed: 03-Jul-2020]

[77] "National Blood Authority Data and Information Governance Framework," National Blood Authority, Mar. 2015.

[78] “Data Plan 2016-19,” Australian Commission on Safety and Quality in Health Care, Dec. 2016 [Online]. Available: https://www.safetyandquality.gov.au/sites/default/files/2019-07/endorsed-acsqhc-data-plan-2016-19_22-dec-2016.pdf. [Accessed: 03-Jul-2020]

[79] “Policies and Standards | Data Governance.” [Online]. Available: https://www.datagovernance.unsw.edu.au/policies-and-standards. [Accessed: 03-Jul-2020]

[80] “Patron Data Governance Framework Part of the Data for Decisions research initiative,” 2019, doi: 10.26188/5c52934b4aeb0. [Online]. Available: https://doi.org/10.26188/5c52934b4aeb0. [Accessed: 03-Aug-2020]

[81] “Regulatory & Ethics Toolkit.” [Online]. Available: https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/. [Accessed: 03-Aug-2020]

[82] C. Chen, P.-I. Lee, K. J. Pain, D. Delgado, C. L. Cole, and T. R. Campion, “Replacing Paper Informed Consent with Electronic Informed Consent for Research in Academic Medical Centers: A Scoping Review” [Online]. Available: https://www.github.com/wcmc-research-informatics. [Accessed: 09-Jul-2020]

[83] “CTRL,” 2018. [Online]. Available: https://www.australiangenomics.org.au/resources/for-patients/your-personal-platform/. [Accessed: 09-Jul-2020]

[84] “GA4GH - DUO,” 2020. [Online]. Available: https://github.com/EBISPOT/DUO/wiki/Documentation. [Accessed: 27-Jan-2020]

[85] “Machine-readable Consent Guidance,” 2020. [Online]. Available: https://drive.google.com/file/d/102_I0_phOGs9YSmPx7It9CSt1sHFJ87C/view. [Accessed: 04-Aug-2020]

[86] “Principles of Māori Data Sovereignty Definition of terms ‘He rei ngā niho, he paraoa ngā kauae’” [Online]. Available: https://www.un.org/development/desa/indigenouspeoples/. [Accessed: 03-Jun-2020]

[87] “International Indigenous Data Sovereignty,” 2017. [Online]. Available: https://www.rd-alliance.org/groups/international-indigenous-data-sovereignty-ig. [Accessed: 03-Jun-2020]

[88] “History of Indigenous Data Sovereignty.” [Online]. Available: https://www.maiamnayriwingara.org/projects-1. [Accessed: 03-Jun-2020]

[89] A. P. G. Bodkin-Andrews, P. M. Walter, D. V. Lee, P. T. Kukutai, and D. R. Lovett, “Delivering Indigenous Data Sovereignty,” 2019 [Online]. Available: https://aiatsis.gov.au/publications/presentations/delivering-indigenous-data-sovereignty. [Accessed: 03-Jun-2020]

[90] “Genomics | Lowitja Institute.” [Online]. Available: https://www.lowitja.org.au/page/research/research-categories/science-and-health-conditions/genomics. [Accessed: 10-Jul-2020]

[91] “National Centre for Indigenous Genomics,” 2017. [Online]. Available: https://ncig.anu.edu.au/. [Accessed: 03-Feb-2020]

[92] “Data Privacy and Security Policy,” Global Alliance for Genomics & Health, 2019 [Online]. Available: https://www.ga4gh.org/wp-content/uploads/GA4GH-Data-Privacy-and-Security-Policy_FINAL-August-2019_wPolicyVersions.pdf. [Accessed: 03-Aug-2020]

[93] I. R. Wiechers, N. C. Perin, and R. Cook-Deegan, “The emergence of commercial genomics: Analysis of the rise of a biotechnology subsector during the Human Genome Project, 1990 to 2004,” Genome Medicine, vol. 5, no. 9, p. 83, Sep. 2013, doi: 10.1186/gm487. [Online]. Available: http://genomemedicine.biomedcentral.com/articles/10.1186/gm487. [Accessed: 09-Jul-2020]

[94] “Research - 23andMe AU, DE, FR & EU.” [Online]. Available: https://www.23andme.com/en-int/research/. [Accessed: 09-Jul-2020]

[95] “Intellectual property and human genomics,” 2010. [Online]. Available: https://www.who.int/genomics/elsi/ip/en/. [Accessed: 09-Jul-2020]

[96] “Privacy in your state — OAIC.” [Online]. Available: https://www.oaic.gov.au/privacy/privacy-in-your-state/. [Accessed: 03-Aug-2020]

[97] T. Finnegan, A. Hall, and J. M. Skopek, Identification and genomic data. PHG Foundation, 2017 [Online]. Available: https://www.phgfoundation.org/report/identification-and-genomic-data. [Accessed: 19-May-2020]

[98] “Data Security,” 2017. [Online]. Available: https://www.ga4gh.org/work_stream/data-security/. [Accessed: 10-Jul-2020]

[99] "Why Healthcare Is Finally Moving To The Cloud - HIT Consultant." [Online]. Available: https://hitconsultant.net/2019/09/06/why-healthcare-is-finally-moving-to-the-cloud/#.XpkEDMgzauc. [Accessed: 17-Apr-2020]

[100] "Draft Data Sharing Agreement Template," 2019. [Online]. Available: https://www.datacommissioner.gov.au/resources/draft-data-sharing-agreement-template. [Accessed: 09-Jul-2020]

[101] "Data Access and Sharing Agreement," 2019.

[102] "Shariant," 2018. [Online]. Available: https://www.australiangenomics.org.au/resources/tools/shariant/. [Accessed: 27-Jan-2020]

[103] U. Srinivasan, S. Rao, D. Ramachandran, and D. Jonas, "Flying Blind: Australian Consumers and Digital Health," Australian Health Data Series, 2016 [Online]. Available: https://flyingblind.cmcrc.com. [Accessed: 03-Aug-2020]

[104] U. Srinivasan, D. Ramachandran, C. Quilty, S. Rao, M. Nolan, and D. Jonas, "Australian Researchers and Digital Health Volume 2 of the Australian Health Data Series focusses," 2018.

[105] "Breaking Barriers to Health Data Project," 2020. [Online]. Available: https://www.weforum.org/projects/breaking-barriers-to-health-data-project. [Accessed: 03-Aug-2020]

[106] "National Pathology Accreditation Advisory Council Requirements for Human Medical Genome Testing Utilising Massively Parallel Sequencing Technologies", (First Edition 2017) National Pathology Accreditation Advisory Council, 2017.

[107] S. Earley, Ed., Data Management Body of Knowledge, 2nd Editio. Technics Publications for DAMA International, 2017 [Online]. Available: https://dama.org/content/body-knowledge. [Accessed: 26-Jun-2020]

[108] "National Digital Health Strategy." [Online]. Available: https://conversation.digitalhealth.gov.au/. [Accessed: 19-Aug-2020]

[109] "Interoperability in Healthcare." [Online]. Available: https://www.himss.org/resources/interoperability-healthcare. [Accessed: 19-Aug-2020]

[110] "Interoperability: better connections for better care - Australian Digital Health Agency." [Online]. Available: https://www.digitalhealth.gov.au/about-the-agency/digital-health-space/interoperability-better-connections-for-better-care. [Accessed: 19-Aug-2020]

[111] D. Rowlands, "A Health Interoperability Standards Development, Maintenance and Management" [Online]. Available: http://www.jpconsulting.com.au. [Accessed: 19-Aug-2020]

[112] "Driver projects," 2013. [Online]. Available: https://www.ga4gh.org/how-we-work/driver-projects/. [Accessed: 04-Aug-2020]

[113] "Health Level Seven International - Homepage | HL7 International." [Online]. Available: https://www.hl7.org/. [Accessed: 06-Apr-2020]

[114] "HL7 Standards Product Brief - HL7 Version 2 Product Suite | HL7 International." [Online]. Available: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=185. [Accessed: 25-Jun-2020]

[115] "HL7 Standards Product Brief - CDA® Release 2 | HL7 International." [Online]. Available: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=7. [Accessed: 25-Jun-2020]

[116] "HL7 Standards Product Brief - FHIR® R4 (HL7 Fast Healthcare Interoperability Resources, Release 4) | HL7 International." [Online]. Available: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=491. [Accessed: 25-Jun-2020]

[117] G. Alterovitz et al., "FHIR Genomics: enabling standardization for precision medicine use cases," npj Genomic Medicine, vol. 5, no. 1, pp. 1–4, Dec. 2020, doi: 10.1038/s41525-020-0115-6. [Online]. Available: https://www.nature.com/articles/s41525-020-0115-6. [Accessed: 25-Jun-2020]

[118] "Consent," FHIR v4.0.1, 2019. [Online]. Available: https://www.hl7.org/fhir/consent.html. [Accessed: 27-Jan-2020]

[119] "Genomic CDM Subgroup," 2018. [Online]. Available: https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:genetics-sg. [Accessed: 25-Jun-2020]

[120] "Australian Health Research Alliance (AHRA) Network Data Collaboration; addressing Data Quality Frameworks, Clinical Terminology Standardisation & OMOP Common Data Model Implementation."

[Online]. Available: https://medicine.unimelb.edu.au/research-groups/general-practice-research/health-data-science-for-medical-research/the-ahra-transformational-data-collaboration-an-australian-health-research-alliance-ahra-network-datacollaboration. [Accessed: 21-Aug-2020]

[121] "The Australian Health Research Alliance," 2019. [Online]. Available: https://ahra.org.au/. [Accessed: 10-Jul-2020]

[122] "OHDSI-Australia." [Online]. Available: https://ohdsi-australia.org/. [Accessed: 10-Jul-2020]

[123] "FGED: Projects," 2020. [Online]. Available: http://fged.org/projects/. [Accessed: 08-Jul-2020]

[124] "Metadata standards," 2019. [Online]. Available: https://www.aihw.gov.au/about-our-data/metadata-standards. [Accessed: 04-Aug-2020]

[125] "Data access, security and privacy," 2015. [Online]. Available: https://www.genomicsengland.co.uk/understanding-genomics/data/. [Accessed: 08-Jul-2020]

[126] "Data Access - EGA European Genome-Phenome Archive." [Online]. Available: https://ega-archive.org/access/data-access. [Accessed: 08-Jul-2020]

[127] "Submit to EGA." [Online]. Available: https://www.ebi.ac.uk/ega/submission. [Accessed: 10-Jun-2020]

[128] "Data Sharing Agreements," 2020. [Online]. Available: https://www.datagovernance.unsw.edu.au/data-sharing-agreements. [Accessed: 09-Jul-2020]

[129] "Help Me Understand Genetics - Genetics Home Reference," 2020. [Online]. Available: https://ghr.nlm.nih.gov/primer#. [Accessed: 22-Jan-2020]

[130] "Cytogenetics," 2020. [Online]. Available: https://www.labtestsonline.org.au/inside-the-lab/genetic-testing/cytogenetics. [Accessed: 08-Jun-2020]

[131] "Chromosome Analysis (Karyotyping)," 2020. [Online]. Available: https://www.labtestsonline.org.au/learning/test-index/chromosome-analysis-karyotyping#. [Accessed: 17-Feb-2020]

[132] "Fluorescence in situ Hybridisation (FISH)," 2020. [Online]. Available: https://www.labtestsonline.org.au/inside-the-lab/laboratory-methods/fish-(1). [Accessed: 17-Feb-2020]

[133] A. Theisen, "Microarray-based Comparative Genomic Hybridization (aCGH) | Learn Science at Scitable," Nature Education, vol. 1, no. 1, p. 45, 2008 [Online]. Available: https://www.nature.com/scitable/topicpage/microarray-based-comparative-genomic-hybridization-acgh-45432/. [Accessed: 21-Aug-2020]

[134] M. Etebari, M. Navari, and P. P. Piccaluga, "SNPs Array Karyotyping in Non-Hodgkin Lymphoma," vol. 4, pp. 551–569, 2015, doi: 10.3390/microarrays4040551. [Online]. Available: www.mdpi.com/journal/microarrays. [Accessed: 08-Jun-2020]

[135] "Australian Health Genetics/ Genomics Survey 2017 Report of Key Findings," The Royal College of Pathologists of Australia, 2019.

[136] "What are whole exome sequencing and whole genome sequencing? - Genetics Home Reference - NIH." [Online]. Available: https://ghr.nlm.nih.gov/primer/testing/sequencing. [Accessed: 21-Aug-2020]

[137] F. Bewicke-Copley, E. A. Kumar, G. Palladino, K. Korfi, and J. Wang, "Applications and analysis of targeted genomic sequencing in cancer studies," Computational and Structural Biotechnology Journal, vol. 17, pp. 1348–1359, 2019, doi: 10.1016/j.csbj.2019.10.004. [Online]. Available: https://doi.org/10.1016/j.csbj.2019.10.004. [Accessed: 21-Aug-2020]

[138] "Genomic medicine." [Online]. Available: https://www.garvan.org.au/research/genomics/genomic-medicine?gclid=CjwKCAjw5vz2BRAtEiwAbcVIL3ktvg9ZIjnR8MokYnzgyEMqG6eDuhSVF-1spwO8iJYmovu8CeiYFRoCJUsQAvD_BwE. [Accessed: 10-Jun-2020]

[139] "Genomics," 2020. [Online]. Available: https://www.garvan.org.au/research/genomics/. [Accessed: 23-Jan-2020]

[140] "RACGP - Genomics in general practice," 2018. [Online]. Available: https://www.racgp.org.au/clinical-resources/clinical-guidelines/key-racgp-guidelines/view-all-racgp-guidelines/genomics-in-general-practice. [Accessed: 23-Jan-2020]

[141] "Process of genetic counselling (2012 GL02)," 2012 [Online]. Available: www.hgsa.org.au. [Accessed: 09-Jun-2020]

[142] “Workflows for genomic medicine | PhenoTips,” 2015. [Online]. Available: https://phenotips.com/index.html. [Accessed: 23-Jan-2020]

[143] “FamilyMemberHistory,” FHIR v4.0.1, 2011. [Online]. Available: https://www.hl7.org/fhir/familymemberhistory.html. [Accessed: 23-Jan-2020]

[144] R. L. Bennett, K. S. French, R. G. Resta, and D. L. Doyle, “standardized human pedigree nomenclature: Update and assessment of the recommendations of the National Society of Genetic Counselors,” Journal of Genetic Counseling, vol. 17, no. 5. pp. 424–433, Oct-2008.

[145] “HL7 Standards Product Brief - HL7 Version 3 Standard: Clinical Genomics; Pedigree, Release 1 | HL7 International,” 2012. [Online]. Available: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=8. [Accessed: 27-Jan-2020]

[146] “Standard Pedigree Symbols,” 2014 [Online]. Available: https://precisionmedicine.duke.edu/researchers/precision-medicine-programs/risk-assessment/family-history/metree-software/supporting-materials. [Accessed: 27-Jan-2020]

[147] “SNOMED - Home | SNOMED International,” 2019. [Online]. Available: http://www.snomed.org/. [Accessed: 27-Jan-2020]

[148] “Phenopackets”. 2019 [Online]. Available: https://aehrc.github.io/fhir-phenopackets-ig/. [Accessed: 27-Jan-2020]

[149] J. O. B. Jacobsen, P. N. Robinson, and C. J. Mungall, “What is a Phenopacket?,” 2019. [Online]. Available: https://phenopackets-schema.readthedocs.io/en/1.0.0/basics.html. [Accessed: 12-Jun-2020]

[150] “Human Phenotype Ontology,” 2020. [Online]. Available: https://hpo.jax.org/app/. [Accessed: 27-Jan-2020]

[151] A. Metke and G. H. Alliance, “The Human Phenotype Ontolgy in Ontoserver | 1 The Human Phenotype Ontolgy in Ontoserver Technical Report,” 2016.

[152] “CSIRO SNOMED CT to HPO Mapper,” 2020. [Online]. Available: https://genomics.ontoserver.csiro.au/mapper/. [Accessed: 06-Aug-2020]

[153] T. G. Schulze and F. J. McMahon, “Defining the phenotype in human genetic studies: Forward genetics and reverse phenotyping,” in Human Heredity, 2004, vol. 58, no. 3–4, pp. 131–138, doi: 10.1159/000083539.

[154] M. A. Swertz et al., “XGAP: a uniform and extensible data model and software platform for genotype and phenotype experiments,” Genome Biology, vol. 11, no. 3, p. R27, 2010, doi: 10.1186/gb-2010-11-3-r27. [Online]. Available: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r27. [Accessed: 27-Jan-2020]

[155] “Exome Sequencing 101: Part 1 - Library Preparation,” 2018. [Online]. Available: https://www.twistbioscience.com/blog/science/exome-sequencing-101-part-1-library-preparation. [Accessed: 10-Jun-2020]

[156] “DNAnexus Titan,” 2019. [Online]. Available: https://www.dnanexus.com/product-overview/titan. [Accessed: 22-Apr-2020]

[157] “Cromwell,” 2019. [Online]. Available: https://cromwell.readthedocs.io/en/stable/. [Accessed: 22-Apr-2020]

[158] J. Leipzig, “A review of bioinformatic pipeline frameworks,” Briefings in Bioinformatics, vol. 18, no. 3, pp. 530–536, May 2017, doi: 10.1093/bib/bbw020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5429012/. [Accessed: 03-Feb-2020]

[159] Jill Adams, “Complex Genomes: Shotgun Sequencing | Learn Science at Scitable,” Nature Education, vol. 1, no. 1, p. 186, 2008 [Online]. Available: https://www.nature.com/scitable/topicpage/complex-genomes-shotgun-sequencing-609/. [Accessed: 21-Aug-2020]

[160] HTS format specifications. 2020 [Online]. Available: https://samtools.github.io/hts-specs/. [Accessed: 27-Jan-2020]

[161] “What are CRAM files?,” 2020. [Online]. Available: https://www.internationalgenome.org/faq/what-are-cram-files/. [Accessed: 09-Jun-2020]

[162] “SNP - NCBI.” [Online]. Available: https://www.ncbi.nlm.nih.gov/snp/. [Accessed: 09-Jun-2020]

[163] “Resource bundle – GATK.” [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle. [Accessed: 10-Jun-2020]

[164] D. J. McCarthy et al., “Choice of transcripts and software has a large effect on variant annotation,” Genome Medicine, vol. 6, no. 3, p. 26, Mar. 2014, doi: 10.1186/gm543. [Online]. Available: https://genomemedicine.biomedcentral.com/articles/10.1186/gm543. [Accessed: 03-Feb-2020]

[165] S. R. Sallah et al., “Using an integrative machine learning approach utilising homology modelling to clinically interpret genetic variants: CACNA1F as an exemplar,” European Journal of Human Genetics, 2020, doi: 10.1038/s41431-020-0623-y. [Online]. Available: https://doi.org/10.1038/s41431-020-0623-y. [Accessed: 08-Jun-2020]

[166] K. Lee et al., “Scaling up data curation using deep learning: An application to literature triage in genomic variation resources,” 2018, doi: 10.1371/journal.pcbi.1006390. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1006390. [Accessed: 08-Jun-2020]

[167] C. Wu et al., “Using Machine Learning to Identify True Somatic Variants from Next-Generation Sequencing,” Using Machine Learning to Identify True Somatic Variants from Next-Generation Sequencing, p. 670687, Aug. 2019, doi: 10.1101/670687.

[168] “ClinVar,” 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/clinvar/. [Accessed: 27-Jan-2020]

[169] “COSMIC | Catalogue of Somatic Mutations in Cancer,” 2019. [Online]. Available: https://cancer.sanger.ac.uk/cosmic. [Accessed: 27-Jan-2020]

[170] “Database of Genomic Variants,” 2016. [Online]. Available: http://dgv.tcag.ca/dgv/app/home. [Accessed: 27-Jan-2020]

[171] “gnomAD.” [Online]. Available: https://gnomad.broadinstitute.org/. [Accessed: 09-Jun-2020]

[172] K. J. Karczewski et al., “The mutational constraint spectrum quantified from variation in 141,456 humans,” Nature, vol. 581, no. 7809, pp. 434–443, May 2020, doi: 10.1038/s41586-020-2308-7. [Online]. Available: https://doi.org/10.1038/s41586-020-2308-7. [Accessed: 24-Aug-2020]

[173] “OMIM - Online Mendelian Inheritance in Man.” [Online]. Available: https://omim.org/. [Accessed: 09-Jun-2020]

[174] “PubMed.” [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/. [Accessed: 09-Jun-2020]

[175] S. Richards et al., “Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology,” Genetics in Medicine, vol. 17, no. 5, pp. 405–424, May 2015, doi: 10.1038/gim.2015.30.

[176] “Data Access Committees (DACs),” 2020. [Online]. Available: https://www.ebi.ac.uk/ega/submission/data_access_committee. [Accessed: 10-Jun-2020]

[177] “Research Data Australia.” [Online]. Available: https://researchdata.edu.au/. [Accessed: 10-Jun-2020]

[178] “What is Next-Generation DNA Sequencing?,” 2020. [Online]. Available: https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-you-will-learn/what-next-generation-dna-. [Accessed: 17-Feb-2020]

[179] “Standing Committee on Screening’s Genomic Tests in Population based Screening Programs: Statement”, 2019. [Online]. Available: http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/population-based-screening-framework. [Accessed: 17-Aug-2020]