Australian Genomics Data Infrastructures Survey Report

International

October 2020

Australian Genomics

Background

Australia has an opportunity to develop a customised national genomic data infrastructure. There are concurrent initiatives in Australia, both government-funded and private, exploring the opportunity and requirements of a national genomic data infrastructure.

The infrastructure needs to be scalable and flexible to meet future demands, equitably accessible across the country and capable of managing genomic and other health information produced clinically and in research. It should be built to support genomic data sharing efforts and re-analysis.

The Australian Genomics Health Alliance is contributing by gathering ideas and information about approaches to genomic data management. A significant part of this includes considering infrastructure solutions implemented by large-scale genomic initiatives internationally.

Surveys

Infrastructure surveys were developed using the REDCap electronic data capture tool[1], and web-based survey links were sent to representatives from 40 national genomic medicine initiatives. Survey completion was requested within two weeks, with reminders sent after one week. Recipients could forward the survey link to more appropriate individuals for completion, and multiple individuals could contribute to the same survey.

Responses were received from 17 initiatives, representing North and South America, Europe, Africa and Australasia. Infrastructures ranged from large-scale national precision medicine initiatives, research cohorts, service-based platforms for storage and analysis of multiple projects, and variant databases. Three initiatives indicated they were not at a suitable stage of maturity to complete the survey.

The survey response data reflects a wealth of information, knowledge and experience built by national initiatives who, at different stages of maturity, were able to provide valuable insights relevant to the design and development of an Australian infrastructure.

Prepared by Marie-Jo Brion, Matilda Haas and Tiffany Boughtwood for Australian Genomics
October 2020

[1] Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform, 42*(2), 377-381. doi:10.1016/j.jbi.2008.08.010

1 Core Infrastructure Elements

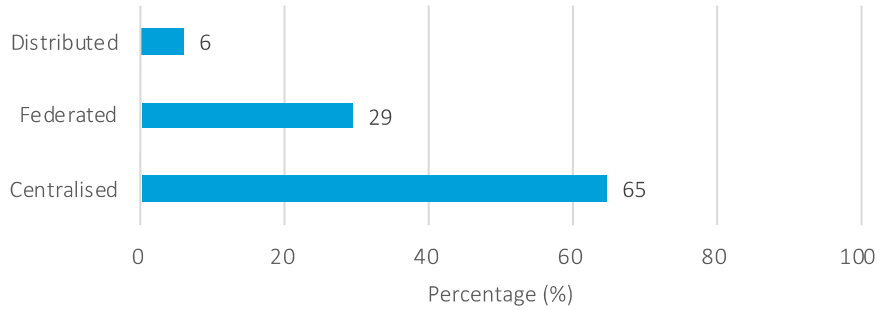
1.1 Infrastructure Model, Type and User Interface

1.1.1 Infrastructure Model

Genomic initiatives predominantly reported having a centralised infrastructure model. This was primarily the option selected by those housing a primary cohort, harmonising datasets towards building a primary cohort, or national precision medicine initiatives.

Those adopting, or working towards, a federated infrastructure were typically 'service' based platforms catering to many independent groups.

Figure 1.1.1 *Infrastructure Model*

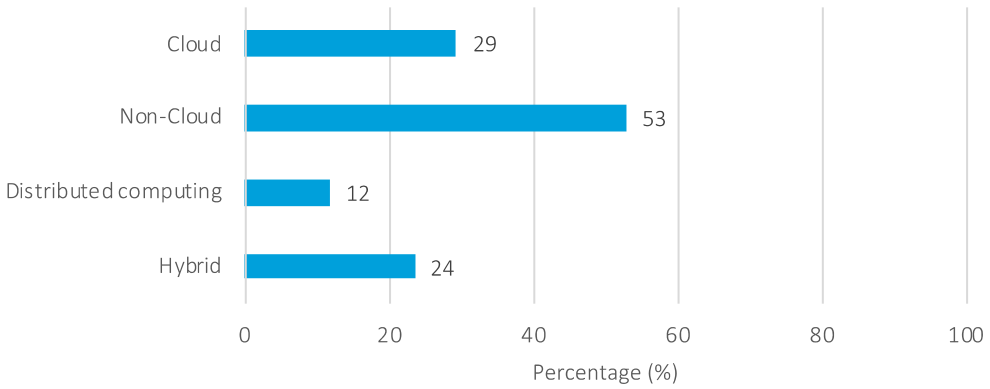


1.1.2 Infrastructure Type

On-premise/non-cloud infrastructure was the most common infrastructure type, with cloud and hybrid including-cloud reported by at least a third of initiatives. Amazon Web Services was the most common provider amongst cloud solutions (3/8).

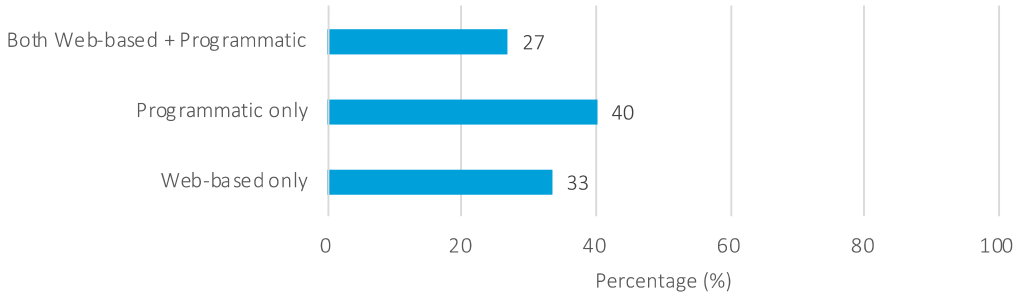
Several non-cloud infrastructures noted plans to transition to cloud in future.

Figure 1.1.2 *Infrastructure Type*



1.1.3 User Interface

Figure 1.1.3 *User Interface*



1.1.4 General Comments

There are a range of infrastructure types and combinations, to suit the different characteristics of each initiative.

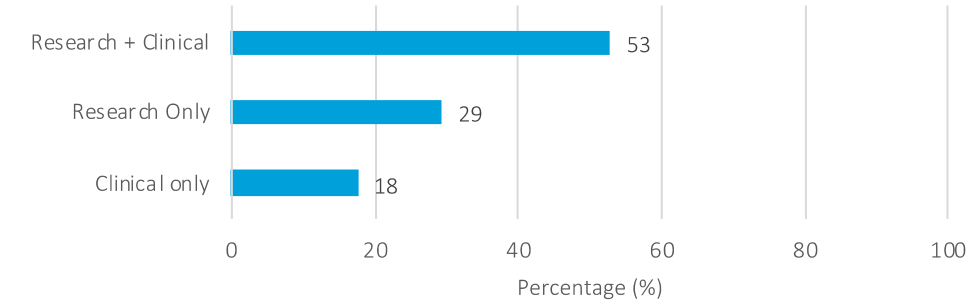
The infrastructures also varied in their maturity, with some not fully implemented, others very established (20 years) and others expanding into new technical phases.

Several initiatives are currently adopting or working to integrate Global Alliance for Genomics and Health (GA4GH) standards, tools, and application programming interfaces, for example the Beacon.

1.2 Data Stored in the Infrastructure

1.2.1 Origin of the Genomic Data

Figure 1.2.1 *Data Origin*

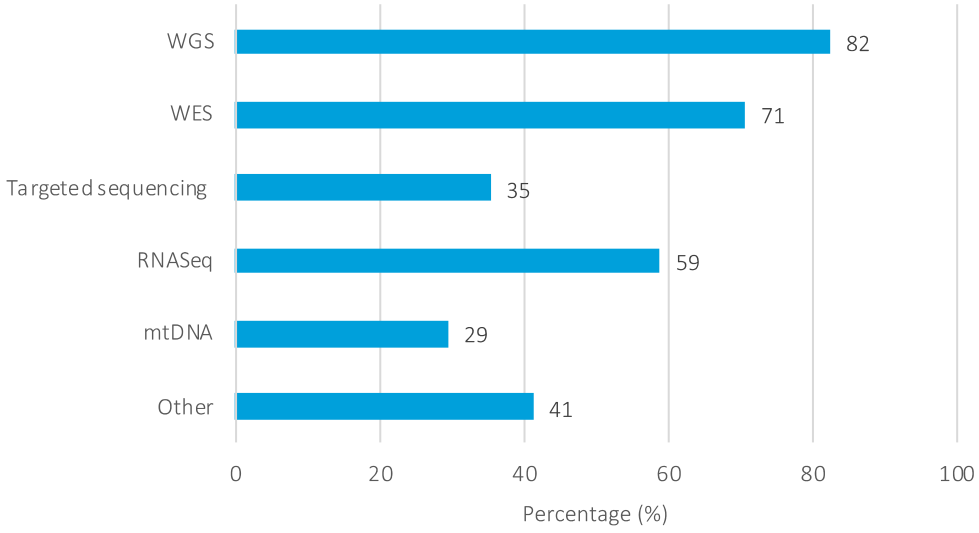


1.2.2 Type of Genomic Data Stored

Most national genomic initiatives store either WGS or WES data.

GWAS and genotyping array a commonly noted 'other' data type.

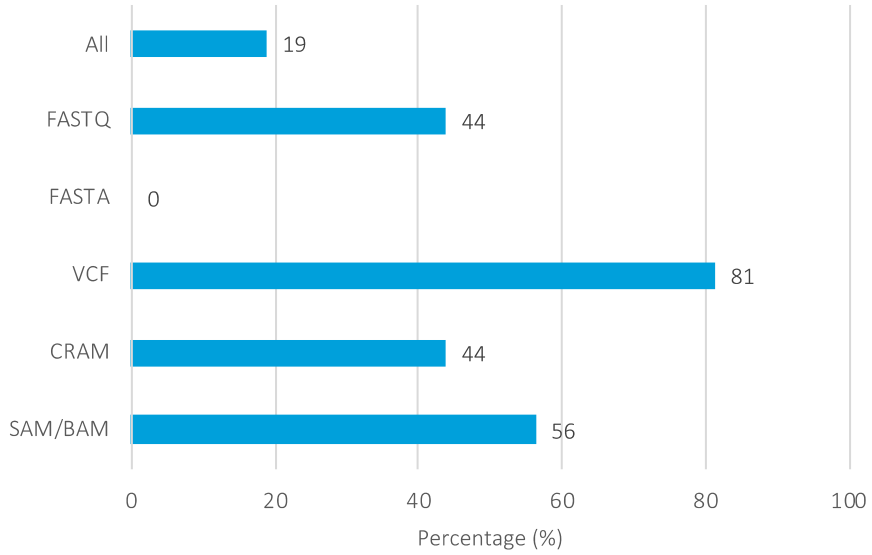
Figure 1.2.2 *Data Type*



1.2.3 File Types Stored

VCFs are the mostly commonly stored file type, and some initiatives only store VCFs. 'Other' data types listed were diverse, including array cel, gVCF, ped and count-level data.

Figure 1.2.3 *File Types*



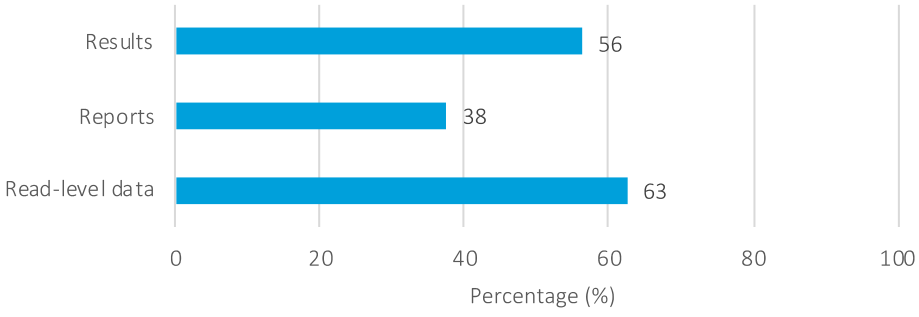
1.2.4 Standard Files Stored Long-term

Read-level data is stored long term by most initiatives. Those who do not store read-level data were primarily variant databases or those only storing VCFs.

Fewer initiatives stored reports and results, those that did were generally the primary cohorts and national precision medicine initiatives.

Several initiatives noted limited capacity for continued long term storage of these data.

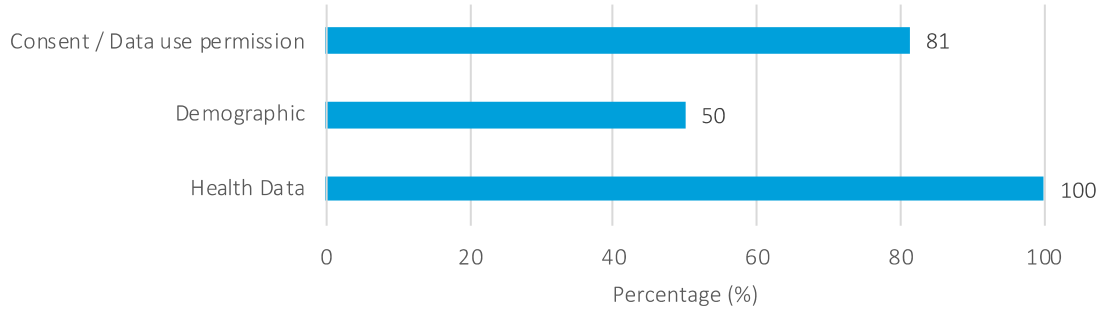
Figure 1.2.4 *Long-term Standard File Storage*



1.2.5 Linkage to Other Data Types

All infrastructures link to health or phenotype data, and most stored consent/data use permissions. Initiatives that stored all three were primary cohort initiatives.

Figure 1.2.5 *Linkage to Other Data*



1.2.6 Standardised Terminologies for Clinical Data

Most initiatives use one or more standardised ontology. However, many noted there was variability within their infrastructure, and between project datasets they store, as to which ontology was used and whether an ontology was applied at all.

HPO was the most commonly cited ontology. Use of ICD-10 classifications was also common.

Storage of health data at the original hospital sites was noted by some initiatives. This meant some limited visibility of data usage and funding requirements.

Challenges with phenotype data included the need for future harmonisation of phenotypes, and difficulties with obtaining quality phenotype data from submitters.

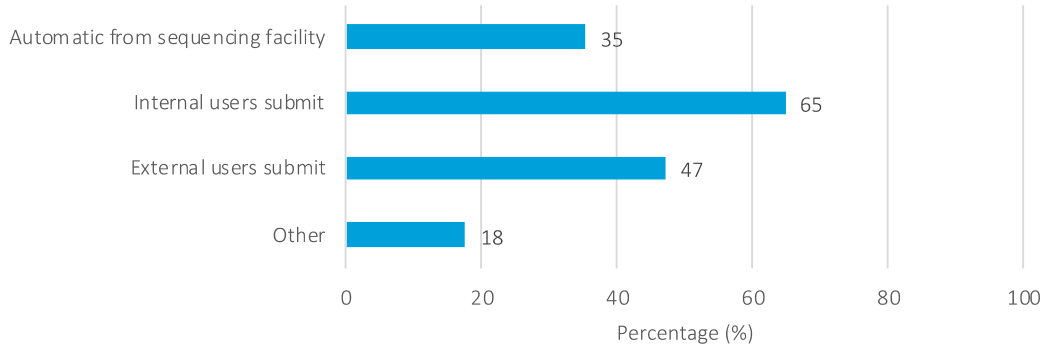
2 Infrastructure Processes

2.1 Data Processing and Handling

2.1.1 Ingesting Data to the Infrastructure

Automated ingestion occurs or is planned, for four of the largest initiatives (with 9,000 TB storage capacity or greater).

***Figure 2.1.1** Data Ingestion*



2.1.2 Harmonisation and Data Compression

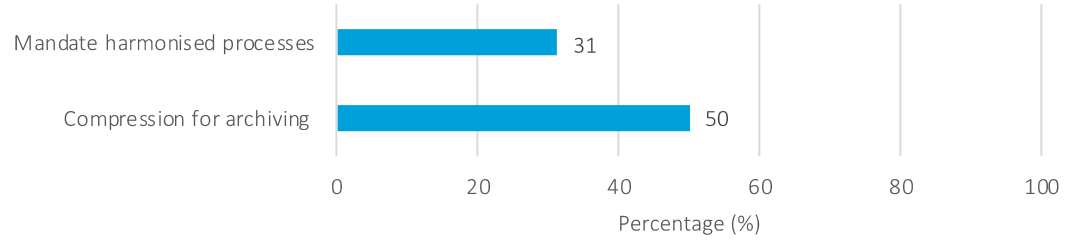
Processing harmonised across data from different sources:

- Only one third of initiatives surveyed harmonise data processing, however responses included a number of initiatives whose data comes from a single source.
- Of those who do not currently harmonise processes, several plan to do so in the future or plan to encourage it, or limit processing options. Several noted the inability to enforce harmonisation.
- Plans to harmonise phenotypes across projects were noted.

Data files compressed for archival storage:

- The use of CRAM and encryption were noted by some initiatives.

***Figure 2.1.2** Harmonisation and Compression*

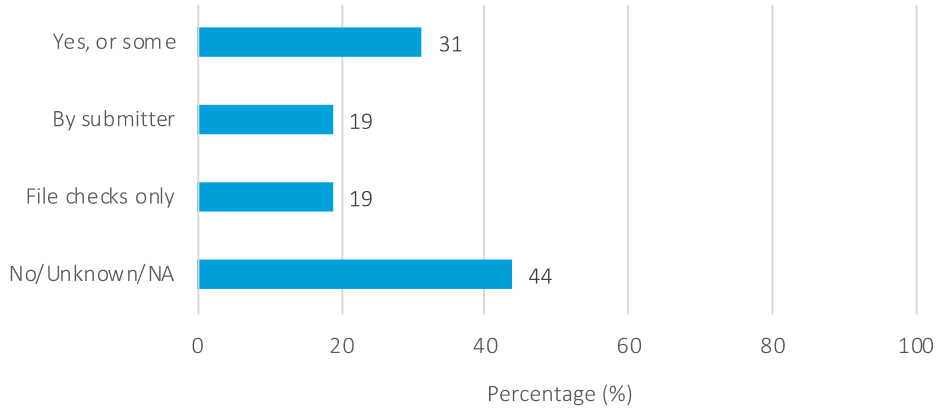


2.1.3 Data Quality Control (QC)

Only about one third of initiatives perform some form of data QC.

Most do no QC, leave QC to data submitters, and/or perform basic file checks only.

Figure 2.1.3 *Data QC*



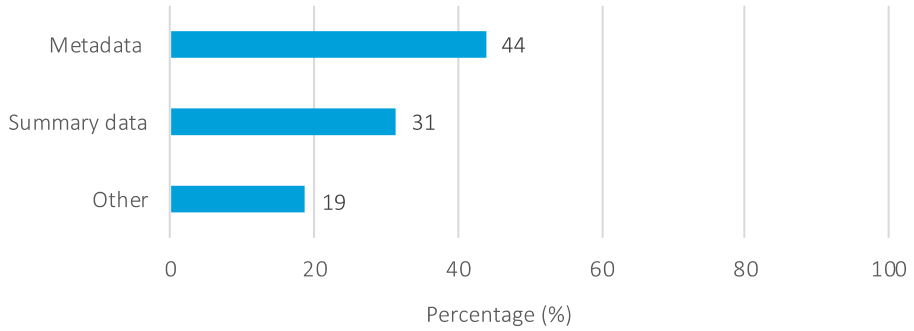
2.2 Data Sharing

2.2.1 Publicly Discoverable Information

Almost half of the surveyed initiatives make metadata publicly discoverable, and one third, summary information. The nature of these varied by project and datasets within infrastructures.

Some noted plans to provide summary statistics and allele frequencies in future or are currently providing some form of aggregated data.

Figure 2.2.1 *Publicly Discoverable Information*

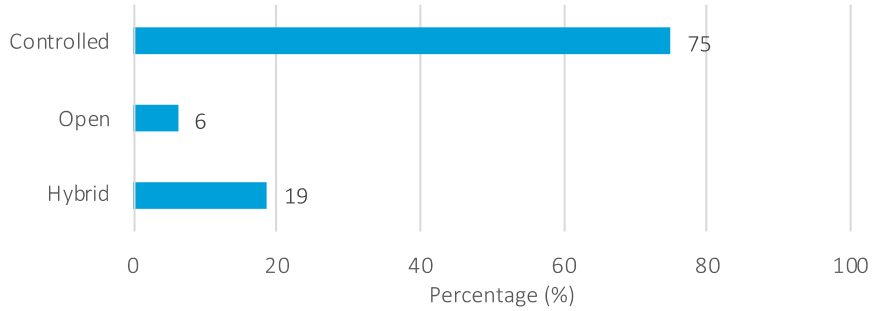


2.2.2 Infrastructure Access Model

Most initiatives implement controlled access models, with a smaller number being hybrid.

Open access was not common, and the one reported instance was linked to a variant database.

Figure 2.2.2 *Access Model*



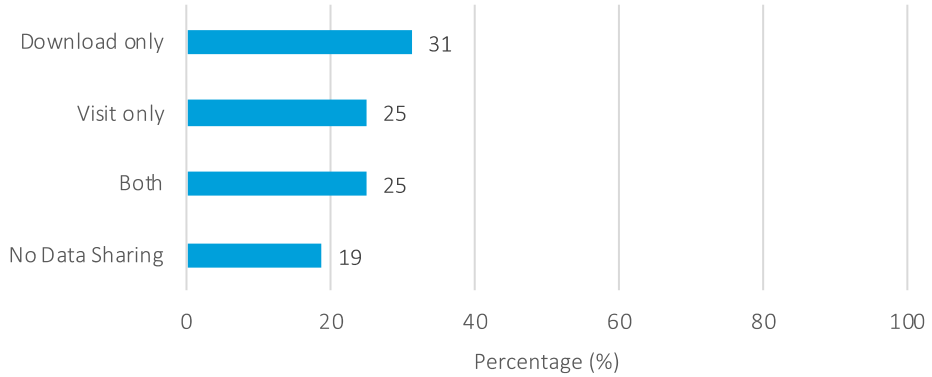
2.2.3 External Data Sharing

Around 80% of the infrastructures support some form of external data sharing, with accessibility dependent on data access committee approvals, and specific permissions.

Three initiatives only provide access for internal or member groups only.

Data download is the most common mechanism of providing external data access. However, data visiting is implemented by a number of the largest initiatives. Data visiting may reflect either running compute in a secure location or accessing aggregated summary data or web-based queries.

Figure 2.2.3 *External Data Sharing*



2.2.4 Workflow Submissions, Querying Contents, and Infrastructure Tools

Can submit workflows, by external users:

- About half support submission of workflows by external users with particular access permissions; of those who do not currently support this, several are actively working towards it or planning to.
- For some initiatives this is not applicable to their infrastructure (e.g. for internal users only, variant databases) and/or the repositories primarily support download-only access.
- Challenges noted include governance and privacy issues, managing approval of workflows, and resources for compute.

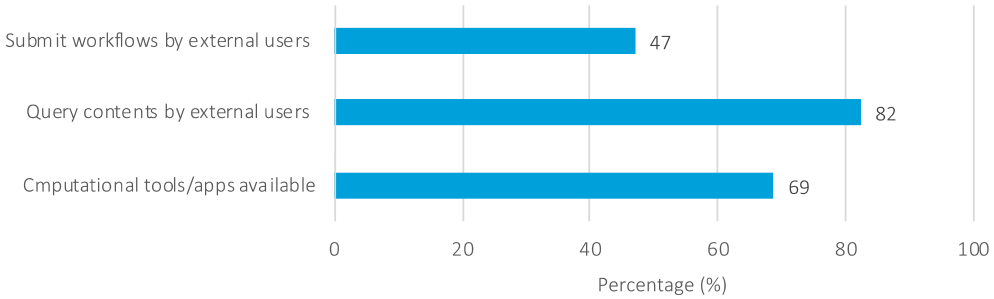
Can query contents by external users:

- Most support capability for external users to query contents of their infrastructure. The type of query or information visible is dependent on specific access permissions. Some allow all users to query non-sensitive information, metadata or processed data. Others require specific permissions.
- Of those that do not, links to public querying portals, such as GA4GH Beacon and Match Maker exchange were noted.

Computational tools and applications:

- Many make computational tools and applications available, primarily for analytics and processing.
- For some, they are not provided in the repository itself, but made available via a separate platform.
- Those that do not, are primarily download access archives or similar.

Figure 2.2.4 *Interoperable Workflows, Queries, and Tools*

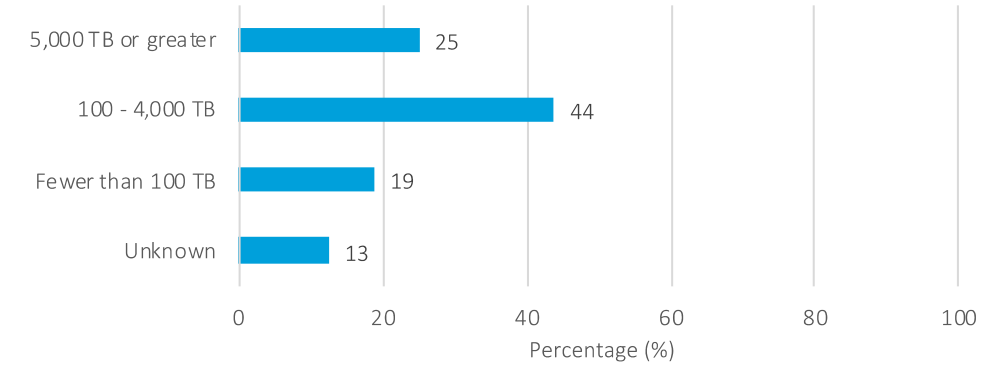


3 Resourcing and Requirements

3.1 Infrastructure Data Storage Requirements

3.1.1 Current Data Usage

Figure 3.1.1 *Current Usage*



3.1.2 Future Funded Storage or Expected Storage Increases

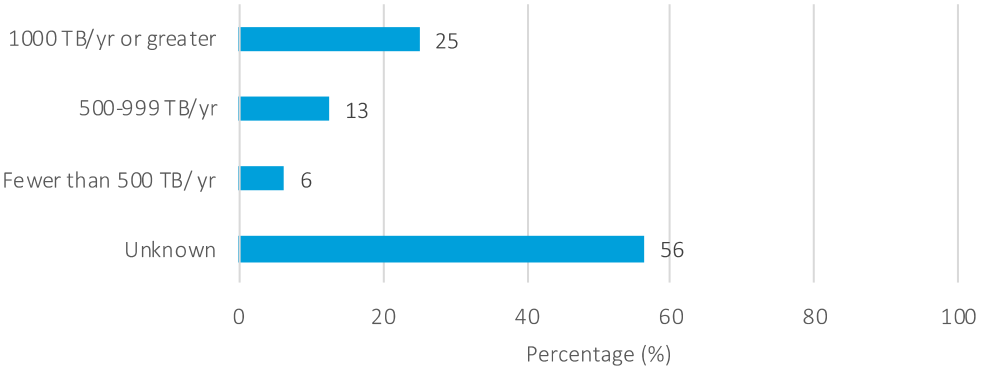
One quarter are initiatives storing extremely large amounts of data (5,000 - 40,000 TB).

Expected increases in storage typically ranged from 25-65% per year, with some expecting to double or triple their annual storage requirements.

One quarter (mostly the larger initiatives) have large expected absolute increases (≥ 1,000 TB/yr).

Future funding availability for storage was unknown for many initiatives, and several noted uncertainty in sustainable funding, and funding as limitation to future scaling and adoption.

Figure 3.1.2 *Future Expected Storage Increases*



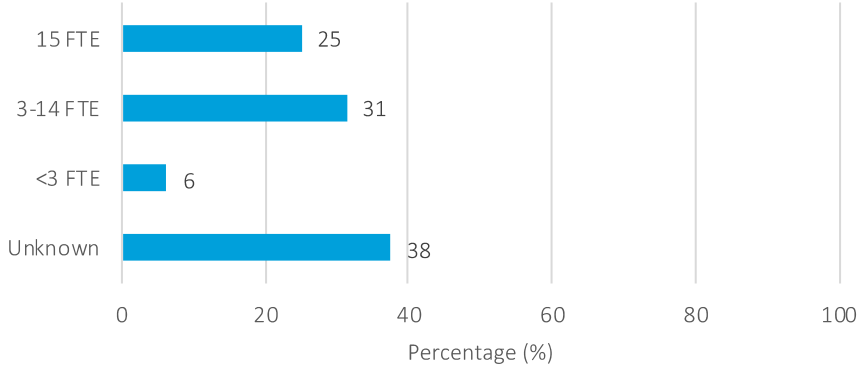
3.2 Operational Requirements

3.2.1 Infrastructure FTE Requirements

FTE requirements for larger initiatives range from 15-50FTE.

Medium to small initiatives are operating on 2-5 FTE.

Figure 3.2.1 *FTE Requirements*



3.2.2 Operating Costs

Infrastructure operating costs for all initiatives are being funded by government and national research funds.

Some initiatives receive supplementation from affiliated universities/institutes, or from specific projects.

3.2.3 Costs for Data Depositors and Infrastructure Users

For most infrastructures, there is no cost for data depositors and data users, with these costs subsidised by the government funding. Some differences were noted for costs charged to commercial/healthcare users, and some charge users compute costs.

4 Evaluation of Current Repository Elements

4.1 Current Challenges to Data Ingestion

Technical challenges and requirements around large datasets:

- Bandwidth, when elasticity is needed
- High speed and secured network, and software for data transportation
- Encryption and transfer of large datasets

Additional challenges:

- Adherence to, and availability of, submission standards for metadata
- Phenotype quality
- Requiring dedicated personnel

4.2 Next Steps for the Infrastructure

- Migrating infrastructures to federated models
- Migrating to cloud-based infrastructure
- Advancing 'new architecture'
- Scaling up infrastructure
- Progressing technical and ELSI (ethical, legal, and social implications) aspects of data sharing

4.3 Limiting Factors to Future Scaling and Adoption

- Funding, funding uncertainty, costs (mentioned by 50%)
- Compute resourcing
- Lack of available solutions for scalable warehousing and genomic databases
- Challenges around data access and governance:
    - data interoperability and ELSI issues
    - data sharing policies that align with different countries
    - building and maintaining infrastructure and security for controlled access data
    - supporting different access models, such as bringing analyses to the data - approvals for usage of infrastructure
    - requirements to retain data locally

4.4 Best Elements of the Organisation's Existing Infrastructure

Retention of data and creating valuable data resources:

- Retention of primary data for future use
- Creating knowledge databases, data platforms or resources for healthcare, researchers, clinicians
- Having permanent dataset identifiers

Adherence to standards:

- Standardisation for future data harmonisation
- Promotes adherence to standards such as GA4GH, and EGA

Enables data access, data harmonisation, and collaboration:

- Quality, simple operations and widespread collaborations
- Return of data generated to the cohort
- Visibility of the data; Submissions to EGA
- Example of data sharing with limited resources

Structural:

- Scalable, secure and elastic, through use of commercial cloud
- Having concrete, usable infrastructure promoting cultural change
- Co-location of storage and compute

4.5 Potential Improvements

General technical improvements:

- Implement a new database
- Usability and flexibility in a secure environment
- Implement efficient and scalable genotype queries
- Expansion to single cell data

Data access:

- Transnational data access, federating infrastructures at a European/international level
- Implementing distributed or federated databases
- User interface to handle data access at more granular levels

Harmonising, standardising and data processing:

- Harmonising data
- Harmonising phenotype data collection
- Processing for allele frequencies
- Standardised ingest mechanisms

Abbreviations and Glossary of Terms

Term	Definition
BAM	*Binary Alignment Map*
CRAM	*Compressed format of BAM/SAM file*
DAC	*Data Access Committee*
ELSI	*Ethical, Legal, and Social Implications*
EGA	*European Genome-Phenome Archive*
VCF	*Variant Call Format*
FASTA	*Text file representing nucleotides or amino acid sequence using single-letter code*
FASTQ	*Text file containing sequence data and quality scores*
GWAS	*Genome-Wide Association Study*
GA4GH	*Global Alliance for Genomics and Health*
mtDNA	*Mitochondrial DNA*
Metadata	*Sequencing, experimental, and computational description associated with a given sample*
RNAseq	*RNA Sequencing*
SAM	*Sequence Alignment Map*
WES	*Whole Exome Sequencing*
WGS	*Whole Genome Sequencing*

Funding Acknowledgements

Australian Genomics is an independent research collaboration launched in 2016 to build the evidence and inform policy for the integration of genomics into mainstream healthcare. It represents 80 organisations including hospitals, research institutes, universities, sequencing laboratories and community groups across Australia. We are funded by the National Health and Medical Research Council (GNT1113531) and the Medical Research Future Fund.