

# **Australian Genomics Data Infrastructures Survey Report**

Domestic

October 2020

**Australian  
Genomics**



<b>Background</b>	<b>2</b>
<b>Surveys</b>	<b>2</b>
<b>1 Core Infrastructure Elements</b>	<b>3</b>
1.1 Infrastructure Model, Type and User Interface	3
1.2 Data Stored in the Infrastructure	5
<b>2 Infrastructure Processes</b>	<b>8</b>
2.1 Data Processing and Handling	8
2.2 Data Sharing	10
<b>3 Resourcing and Requirements</b>	<b>13</b>
3.1 Data Requirements	13
3.2 Operational Requirements	14
<b>4 Evaluation of Current Repository Elements</b>	<b>16</b>
4.1 Current Challenges to Data Ingestion	16
4.2 Next Steps for the Infrastructure	16
4.3 Limiting Factors to Future Scaling and Adoption	17
4.4 Best Elements of the Organisation's Existing Infrastructure	17
4.5 Potential Improvements to the Organisation's Existing Repository	18
<b>5 Considerations for a Future National Genomics Infrastructure (NGIS)</b>	<b>20</b>
5.1 Essential Components of a Future National Infrastructure	20
5.2 Existing Systems and Software to Incorporate in a Future NGIS	22
5.3 Willingness to Pilot a National Genomics Infrastructure Service (NGIS)	22
5.4 Willingness to Contract Services from an NGIS	23
<b>6 International Survey Comparison</b>	<b>24</b>
6.1 Comparisons of Core Infrastructure Elements	24
6.2 Comparisons of Infrastructure Processes	24
6.3 Comparisons of Resourcing and Requirements	24
6.4 Comparisons of Evaluations for Current and Future Repository Elements	25
<b>Funding Acknowledgements</b>	<b>28</b>

## 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 nationally.

## Surveys

Infrastructure surveys were developed using the REDCap electronic data capture tool<sup>1</sup>, and web-based survey links were sent to representatives from 26 Australian organisations managing and 22 using genomic data infrastructure. 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 infrastructure managers and 10 infrastructure users. Infrastructures included university and medical research institutes (6), translational research centres and programs (3), diagnostic testing laboratories (2), infrastructure and data service providers, both research and clinical (3), private/patient genomic data stores (2) and a future planned repository (1).

Results represent the data from infrastructure managers surveys, with supplemental information (where indicated) from infrastructure users. Infrastructure users provided responses relative to primary local infrastructures they use, as well as national or international repositories they submit data to.

The survey response data reflects valuable information, knowledge and experience from domestic organisations and infrastructure users who provided key insights that will be 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*

---

<sup>1</sup> 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

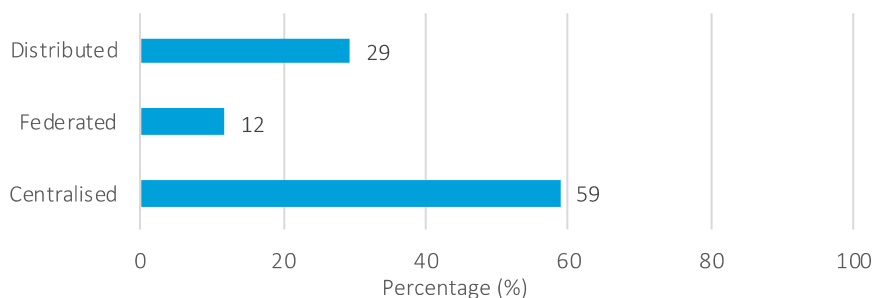
# 1 Core Infrastructure Elements

## 1.1 Infrastructure Model, Type and User Interface

### 1.1.1 Infrastructure Model

Genomic initiatives predominantly reported having a centralised infrastructure model, while a third use a distributed model of multiple datasets across a network.

**Figure 1.1.1** Infrastructure Model

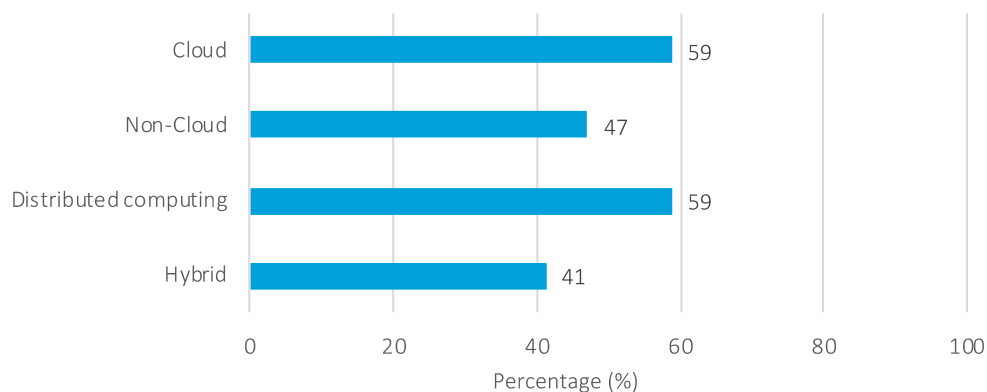


### 1.1.2 Infrastructure Type

Infrastructure types were relatively evenly distributed. The majority of organisations using cloud-based infrastructure also have additional on-premise infrastructure, such as HPC clusters.

Of those nominating cloud types, AWS was commonly used (5/10), with others including Microsoft (1), Google (1), Openstack (1), and a mixed multi-cloud strategy (1).

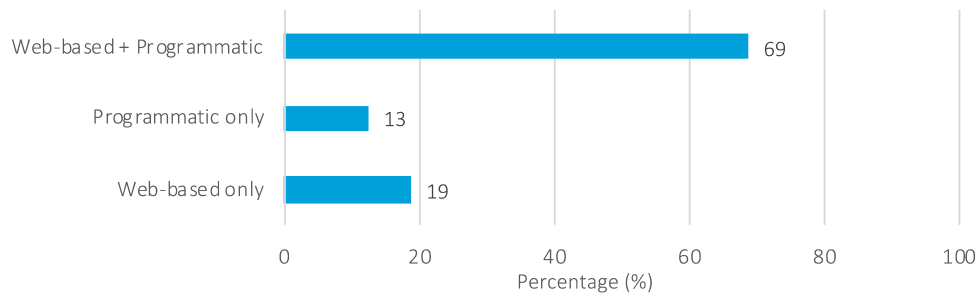
**Figure 1.1.2** Infrastructure Type



### 1.1.3 User Interface

Most infrastructures adopt both web-based and programmatic user interfaces.

**Figure 1.1.3** User Interface

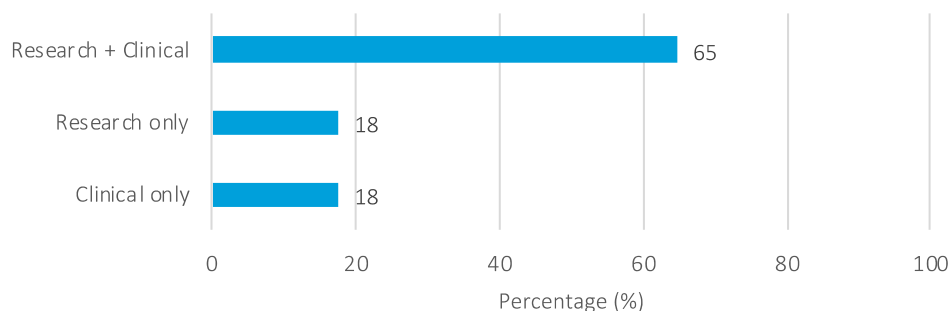


## 1.2 Data Stored in the Infrastructure

### 1.2.1 Origin of the Genomic Data

Most infrastructures are storing human genomic data of both clinical and research origin. Other origins noted include pathogen, animal and plant genomic data.

**Figure 1.2.1 Data Origin**

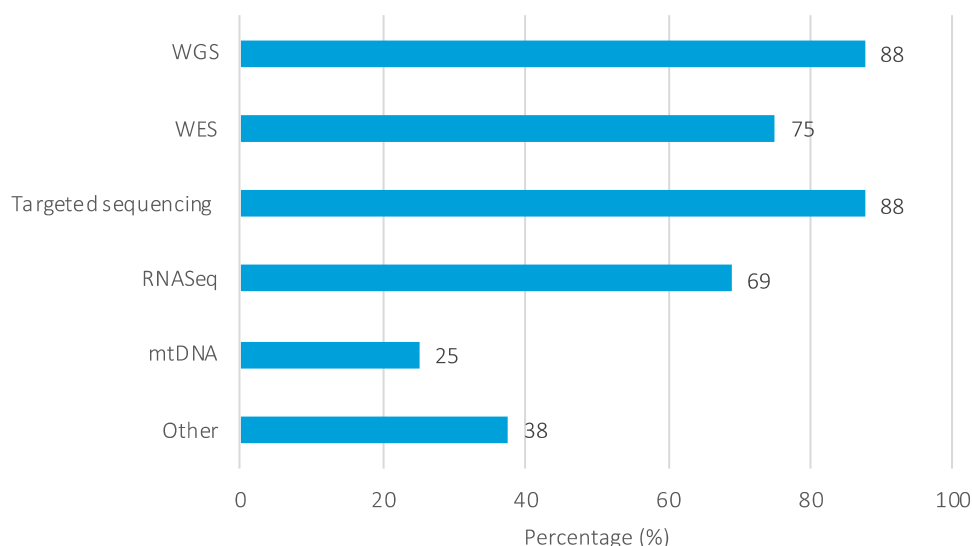


### 1.2.2 Type of Genomic Data Stored

WGS, WES, targeted sequencing data and RNAseq were all commonly stored in the infrastructures.

Other types, noted by infrastructure managers or users, included single cell data, circulating tumour DNA, SNP array, DNA methylation data and from chromatin sequencing assays (ChIPSeq, ATACseq).

**Figure 1.2.2 Data Type**

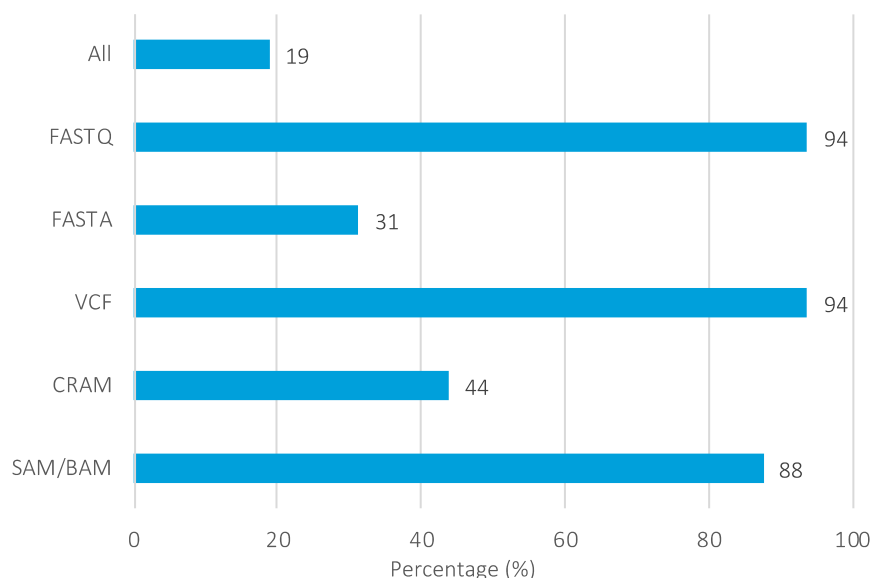


### 1.2.3 File Types Stored

Most infrastructures are storing FASTQ, BAM and VCFs.

More than a third of infrastructures are also storing other file types, which included (as listed by infrastructure managers or users): BED files, QC reports or data, metric and instrument files, BCL data, CSV, TSV, minor allele frequencies, and text output from analysis tools.

**Figure 1.2.3** File Types

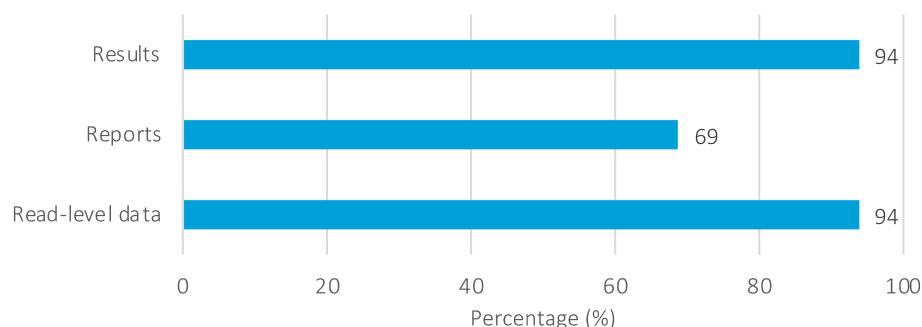


### 1.2.4 Standard Files Stored Long-term

Diagnostic labs and those handling clinical data noted indefinite storage of results and reports.

Research institutes and research programs had a variety of approaches, from no fixed policy or process around data retention (2), storage for 4-7 years (2), and indefinite storage of files (2).

**Figure 1.2.4** Long-term Standard File Storage

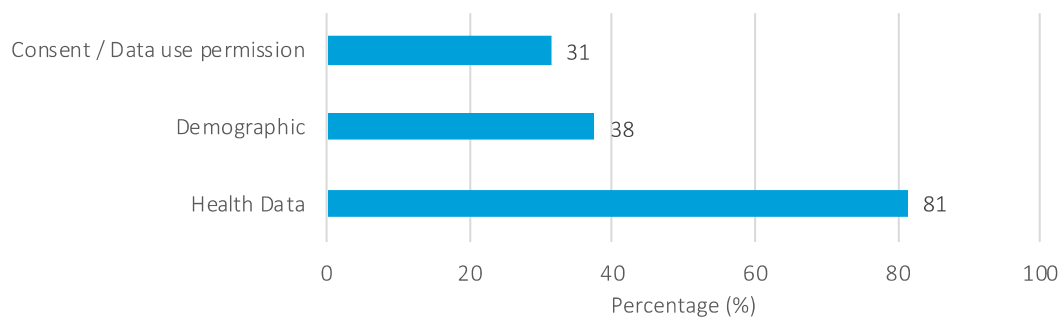


### 1.2.5 Linkage to Other Data Types

Most infrastructures store or link to health data.

Only a third of the infrastructures store consent or data use permissions.

**Figure 1.2.5** Linkage to Other Data



### 1.2.6 Standardised Terminologies for Clinical Data

Most of the infrastructures do not currently store their clinical data in standardised terminologies.

The application of a specific ontology was only named by four organisations - two using HPO, and two using cancer classifications (ICD-9, Oncotree).

Cancer-related infrastructures noted the common ontologies (HPO, SNOMED) were less relevant to cancer.



## 2 Infrastructure Processes

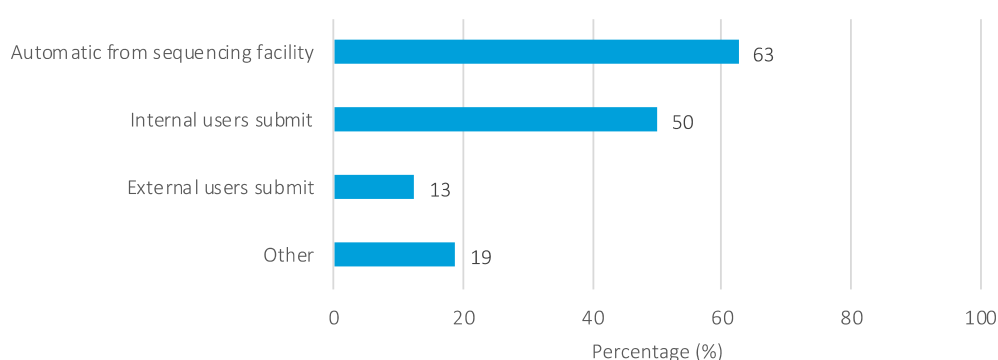
### 2.1 Data Processing and Handling

#### 2.1.1 Ingesting Data to the Infrastructure

Most of the infrastructures either ingest their data directly from the sequencing facility or instrument, or submit data via internal users.

Several infrastructures noted automated ingest of raw data, but internal user submission for processed data.

**Figure 2.1.1** Data Ingestion



#### 2.1.2 Harmonisation and Data Compression

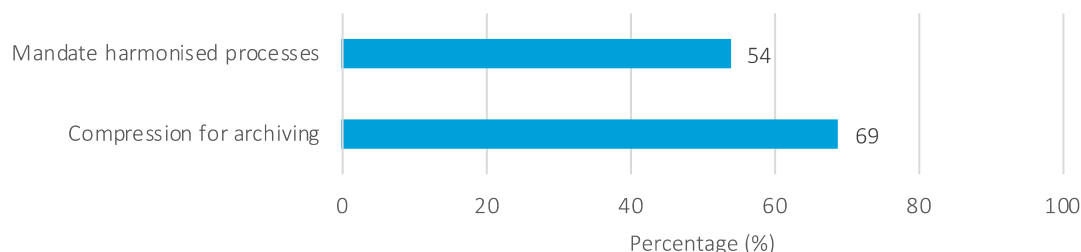
*Harmonised processing mandated for data across different sources:*

- For applicable infrastructures, around half mandate harmonisation of data processing.
- This was noted in relation to aligning to the same reference genome, using the same processing or annotation pipelines, or providing descriptions of workflows with the data.

*Data files compressed for archiving:*

- More than two-thirds of the infrastructures currently compress their data for archiving, using either CRAM or gzip, and several more are planning to do so soon.
- Issues noted included: users often skipping this step, needing to keep FASTQs to perform downstream processing, and unsuitability for somatic workflows.

**Figure 2.1.2** Harmonisation and Compression



### **2.1.3 Data Quality Control (QC)**

Some form of QC is occurring, in all responding infrastructures (13/13). The nature of the QC was variable, including just having md5 checksums for data integrity (3), QC checks and tools on read data (2), to 'standard QC' and 'extensive QC'.

Several noted that QC is performed by the users before submission, rather than by the infrastructure itself, or that it was mostly performed in relation to the sequencing.

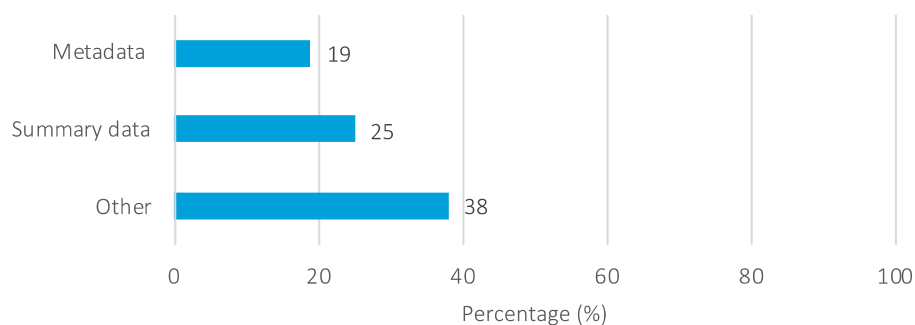
## 2.2 Data Sharing

### 2.2.1 Publicly Discoverable Information

Less than a third of the infrastructures make metadata or summary information from datasets publicly available. Although several infrastructures noted intention to do so.

Other publicly discoverable information from infrastructures included summary statistics through a web-portal, minimal metadata via open discoverable catalogues, and availability of information from research data management platforms.

**Figure 2.2.1** Publicly Discoverable Information

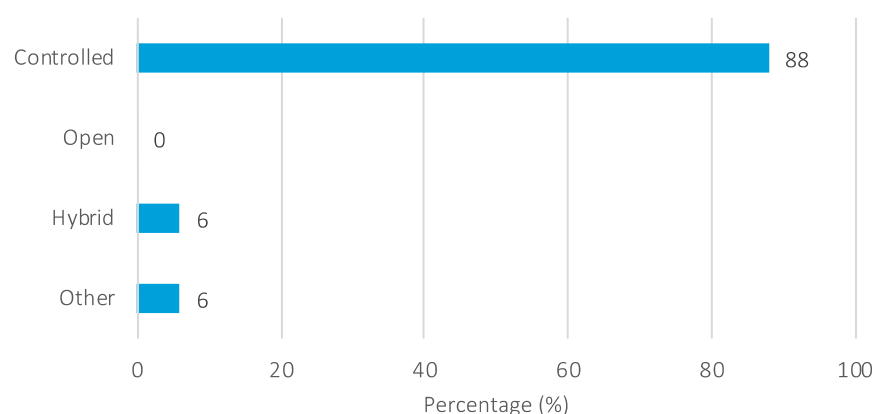


### 2.2.2 Access Model

Most infrastructures operate a controlled access model. Only one (planned) infrastructure intends to operate a hybrid access model. Two infrastructures noted linking to variant-level sharing platforms, such as the Global Alliance Beacon and the Australian Genomics Shariant platform.

Restrictions noted by infrastructures included lack of current ethics to support data sharing, having standardised data access application processes, and lack of a publicly visible data catalogues.

**Figure 2.2.2** Access Model



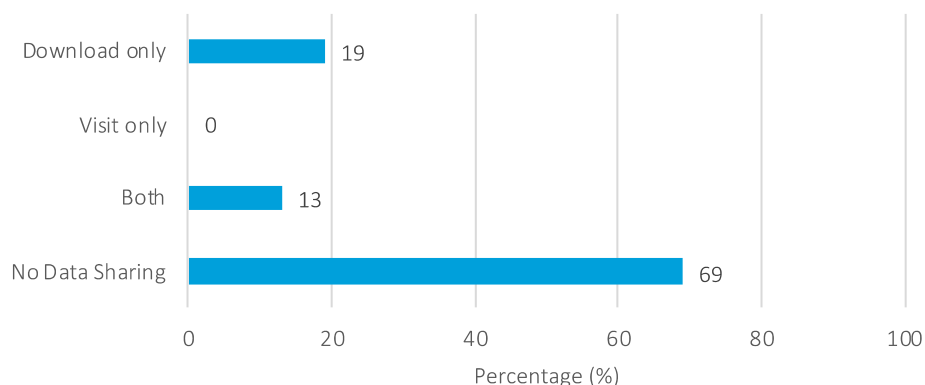
### 2.2.3 External Data Sharing

Over two-thirds of the infrastructures do not support external data sharing.

Of those that do, this occurs by data downloading, or a combination of download and data visiting.

Governance-related issues (lack of appropriate ethics, lack of established governance processes) were noted as key limiting factors to providing external data sharing.

**Figure 2.2.3** External Data Sharing



## 2.2.4 Workflow Submissions, Querying Contents, and Infrastructure Tools

*Can submit workflows, by external users:*

- Only one infrastructure currently supports submission of workflows by external users. However, two noted future interest or plans to do so, and two noted possibility for this, technically.

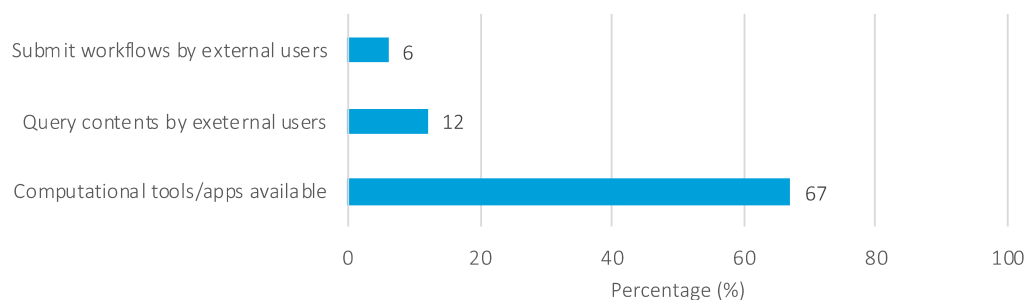
*Can query contents by external users:*

- Most infrastructures do not currently support external querying, with only two organisations currently or partially supporting such queries. Three infrastructures noted future plans to do so.

*Computational tools and applications:*

- Over two-thirds of the infrastructures make computational tools or apps available to users. Various tools were noted, including Galaxy; web notebooks; curation software; cluster access; analysis tools (GATK, STAR, tools for GWAS). Other elements noted included having in-house and commercial tools, and HPC-enabled tools.

**Figure 2.2.4** Interoperable Workflows, Queries, and Tools



### **2.2.5 Authorisation and Access Technologies**

These technologies are being applied in nine infrastructures (60% of responses) and is intended for one planned repository. Five indicated 'none' or 'not applicable'.

The most commonly referenced technologies in use were Australian Access Federation (AAF) (5) and OAuth (3); Others included GSuite log ins, LDAP, VPN and ssh, and institutional controls.

Infrastructures that were service providers, used a variety of external organisation credentials (state health, universities, NCRIS facilities, AAF) to authorise and log into data.

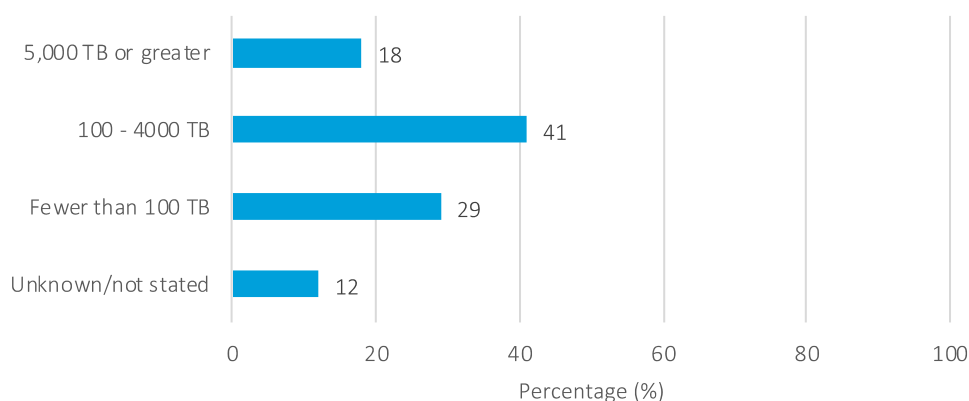
## 3 Resourcing and Requirements

### 3.1 Data Requirements

#### 3.1.1 Current Data Usage

More than half the infrastructures store 100TB or more of genomic data, including three large-capacity infrastructures storing 5PB or more.

**Figure 3.1.1** Current Usage

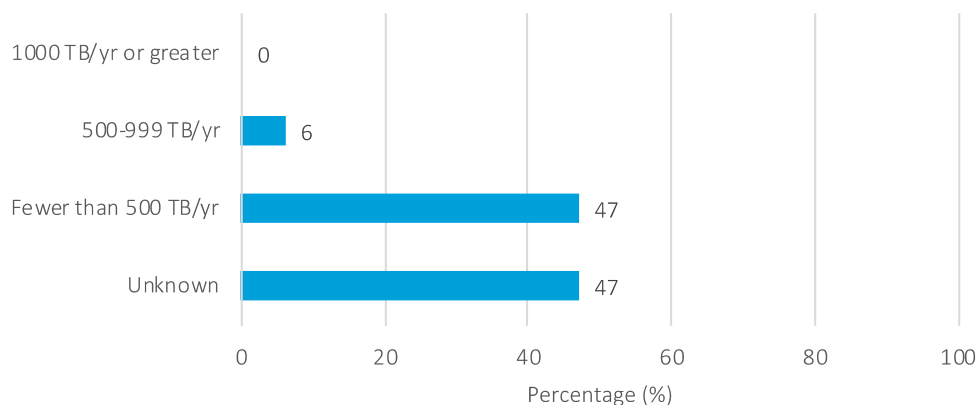


#### 3.1.2 Future Funded Storage or Expected Storage Increases

Around half the infrastructures (4/9 responses) are expecting increases of >20% per year of their current usage.

Funding availability and sources of funding, for these expected increases, were variable across infrastructures: covered by internal operating costs, grants and cost recovery. Several noted uncertainty about future funding for storage.

**Figure 3.1.2** Future Expected Storage Increases



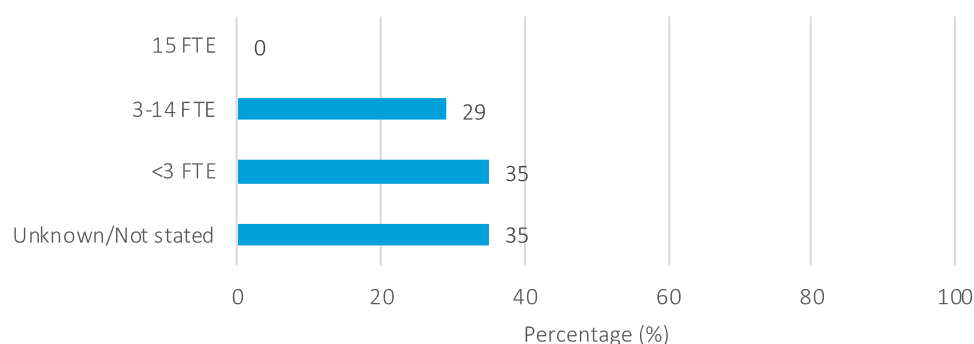
## 3.2 Operational Requirements

### 3.2.1 FTE Requirements

Of the infrastructures who specified their requirements most (8/11) are operating with 3 FTE or less. Three infrastructures required 4 - 8 FTE.

Several noted a single FTE, where allocation was typically split across several different individuals.

**Figure 3.2.1** FTE Requirements



### 3.2.2 Operating Costs

Operating costs were covered by various means, for the different types of organisations.

**Table 3.2.1** Operating Costs

Organisation Type	Primary funding	Supplemental support
<b>Research Infrastructures</b> <i>Universities, Medical Research institutes</i>	Internal core or operational funds	Grants, cost recovery
<b>Clinical infrastructures</b> <i>Diagnostic labs, Clinical platforms</i>	State health departments	Affiliated or member groups, cost recovery
<b>Service-based platforms</b> <i>For research groups</i>	Cost recovery from data owners	Various, including NCRIS

Cost recovery models were a partial or primary approach in 4/16 responses (a clinical organisation, a research institute, a research program, and a private entity).

### 3.2.3 Costs for Data Depositors and Infrastructure Users

**Table 3.2.2** Depositor/User Costs

Process	Cost to Depositor/User	Details
Archiving and Active Storage	No cost in 56% of infrastructures [9/16]	Mostly due to being covered by internal operational costs. Those charging users operate as: flat rate (2), per genome (2), or other approaches (2), e.g. broader levels than per-user, or scaling with use.
Download	No cost in 94% of infrastructures [15/16]	
Analysis	No cost in 75% of infrastructures [12/16]	Two infrastructures that do (or will) charge noted doing so per genome or by computational time. Comments included: <i>"costs subsidised by external service charges"</i> ; <i>"users charged only when internal allocations are exceeded"</i> ; <i>"absence of charges to users is unsustainable"</i>



## 4 Evaluation of Current Repository Elements

### 4.1 Current Challenges to Data Ingestion

Data ingestion challenges were noted by 53% (8/15) of the infrastructure managers, including:

#### *Data format and requirement challenges*

- Lack of integrity checksums accompanying data
- Needing to fit existing table structures
- Data organisation required after upload
- Extremely large data (e.g. germline WGS) being hard to store in traditional databases
- Interoperability

#### *Network and associated challenges*

- Insufficient network bandwidth for scale of data; Slow transfer to/from cloud
- Older instrument operating systems not built to cope with the required transfers
- Hospital firewalls

Absence of ingestion challenges was linked to having automated processes (3); co-located sequencing facility and repository (1); high-speed links to the repository (1).

#### *4.1.1 Infrastructure Users' Perspectives on Data Ingestion*

**Data ingestion challenges were experienced by 50% [4/8] of infrastructure users responding to this question.**

**For some users, challenges are not experienced as ingestion is done by others on their behalf.**

#### **Of those that noted challenges, they included:**

- Time consuming; time taken for transfers from facilities and data size
- Process-related, including the identification of the required data; the varied and manual submission processes; and the fragmented nature of local data storage
- Lack of resources or funding to facilitate upload and sharing
- Costly if using commercial cloud providers

### 4.2 Next Steps for the Infrastructure

#### *Scale up current infrastructure and storage:*

- This was noted by around 2/3 of the infrastructures and included adding cloud and hybrid solutions [7], to on-premise infrastructure.
- Restrictions on ability to scale current infrastructure was noted by three infrastructures – being either technical or human resource-related.

#### *New tools and data:*

- Development of tools and applications for analytics
- Exploring front ends, including commercial solutions
- Including pharmacogenomics data
- Improving standardisation (CRAM) and Metadata

*Adopting Frameworks and strategies (noted by research institutes and research infrastructures):*

- Australian BioCommons strategies for infrastructure, data, and sharing data or pipelines was noted by four research institutes or programs
- Aligning with national and international standards and frameworks

*Progressing data sharing and FAIR data (noted by research institutes and research infrastructure):*

- Better data sharing, making data FAIR (Findable Accessible Interoperable Reusable)
- Authentication and authorisation e.g. using AAF or Elixir AAI
- Using Data Repository Service (DRS) APIs, avoiding data duplication
- Stakeholder engagement on the infrastructure and on the application of FAIR data

### 4.3 Limiting Factors to Future Scaling and Adoption

*Funding and resourcing:*

- 77% (13/17) of the infrastructures noted funding, costs or personnel, as limiting factors to future scaling and adopting. Several noted considering sustainability and business models

*Governance and operational challenges:*

- Requiring sustainable data ownership frameworks
- Implementing appropriate governance for sharing; inter-institutional agreements
- Consensus on the appropriate approach for scaling and adoption
- Managing consent
- Skillset in laboratory users, health department, organisational IT

*Technical challenges:*

- No existing solutions for national authentication and authorisation, national archiving, front ends; and insufficient maturity of cloud systems
- Data curation; data off-boarding; making data FAIR
- Network connectivity
- Limited development resources e.g. software and IT engineers, information, professional resources
- Data security concerns

### 4.4 Best Elements of the Organisation's Existing Infrastructure

*Governance and operations:*

- Privacy and security compliant
- Having centralised infrastructure accessible to multiple organisations and researchers
- Having an established or trusted infrastructure
- Administrated infrastructure for users, co-ordination
- Being local, having in-house servers
- Close relationship to international genome sharing exemplars and to national communities
- Availability of imminent data sharing through Australian BioCommons
- The cost (or no cost) and affordability

#### Technical repository characteristics:

- Flexibility and configurability
  - To move pipelines (*e.g. from on-premise to cloud*)
  - To support future configuration or expansion
  - Modularity of the architecture
  - Ability to add new data with different formats
  - Cloud-based elements
- General repository characteristics of value
  - Scalable processes, noted by four infrastructures
  - Programmatically accessible, having APIs (*e.g. FHIR-based*)
  - Orchestration layers
  - Data fabric architecture and storage technology (*e.g. MeDiCl*)
  - Web interface
  - Data sharing software (*e.g. REMS*)
- Data and processing
  - Infrastructure computational capability, speed and performance (*e.g. for high sequencing volumes, fast instrument data capture and transfer*)
  - Harmonised data processing and availability of comparable data sets
  - Storing and processing data efficiently and at-scale
  - Access to computational tools and optimised pipelines (*e.g. to joint-call, annotate, analyse*)
  - Colocation of data with computational power and software

#### High-level value:

- Delivery of health and biology insights with the available tools and applications
- Supporting cross-disciplinary collaboration and innovation

##### 4.4.1 Infrastructure Users' Perspectives on Best Infrastructure Elements

<b>Usability</b> <ul style="list-style-type: none"> <li>Operates well</li> <li>Has all the required resources</li> <li>Ease of use</li> </ul>	<b>Management</b> <ul style="list-style-type: none"> <li>Managed by others</li> <li>Managed well</li> <li>Good support from facility staff</li> </ul>
<b>Sharing and Access</b> <ul style="list-style-type: none"> <li>Quick and efficient data sharing</li> <li>Facilitated access control</li> </ul>	<b>Other Infrastructure Elements</b> <ul style="list-style-type: none"> <li>Local</li> <li>Expandable</li> <li>Data &amp; compute on same infrastructure</li> </ul>
<b>Cost</b> <ul style="list-style-type: none"> <li>Free to use</li> </ul>	

## 4.5 Potential Improvements to the Organisation's Existing Repository

#### Future and funding considerations:

- More funding to improve the infrastructure, more reliable funding beyond research funds
- Robustness and sustainability of infrastructure
- Implement a roadmap for national services
- Address current fragmentations, centralise resources, community co-ordination

*Data sharing and accessibility:*

- Resolve data sharing and governance and challenges for users (clinical, research, non-expert)
- Improved user permission and access control management
- Improved data discoverability, search tools to query the repository, and fronts for data exploration
- Better sharing of structured metadata with genomic data files
- Electronic management of consent

*Repository elements:*

- Expansion, more compute nodes to execute jobs faster
- Reconfigure a flexible model, to avoid re-building
- Continued modification and expansion
- Migrating to cloud
- GPU/graphical processing
- Cheaper and more scalable for whole genomes
- Better integration
- A back-end storage that handles ingress from instruments
- Improved programmatic interface
- Flexibility in where can be data stored

*Data elements:*

- Improved data quality
- More consistent phenotype capture
- Storage of metadata and clinical data

**4.5.1 Infrastructure Users' Perspectives on Suggested Improvements**

***General Repository Elements***

- Applications and tools for analysis, bring analysis tools to the data
- Replicate existing infrastructures and expand them
- Address significant upload and download challenges associated with external repositories
- Implement a user-friendly interface

***Data access and governance:***

- Improve data discoverability, sharing and management (e.g. standardise processes for storing, managing and locating datasets)
- Implement a national, democratic process for data management
- Improve permission management
- Improve data transfer processes
- Clarity on governance and long-term visibility

***Data elements:***

- Submission of routine clinical genomic data
- More linked phenotype data, to increase the value of the data
- Standardised Clinical data
  - Clinical data captured in HPO at the outset
  - Capacity to machine-read clinical information (e.g. reports) and convert to HPO
  - Capacity to interrogate genomic data using HPO

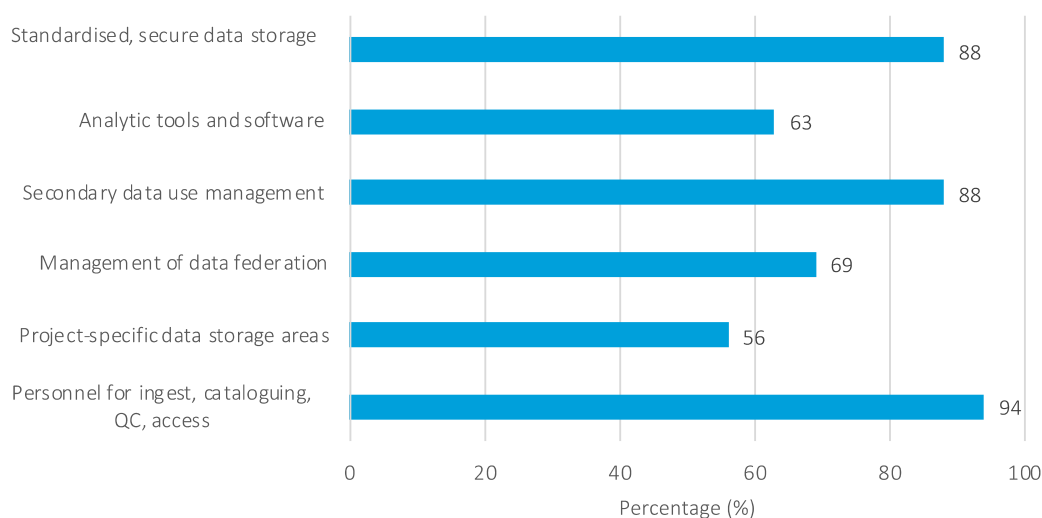
## 5 Considerations for a Future National Genomics Infrastructure (NGIS)

### 5.1 Essential Components of a Future National Infrastructure

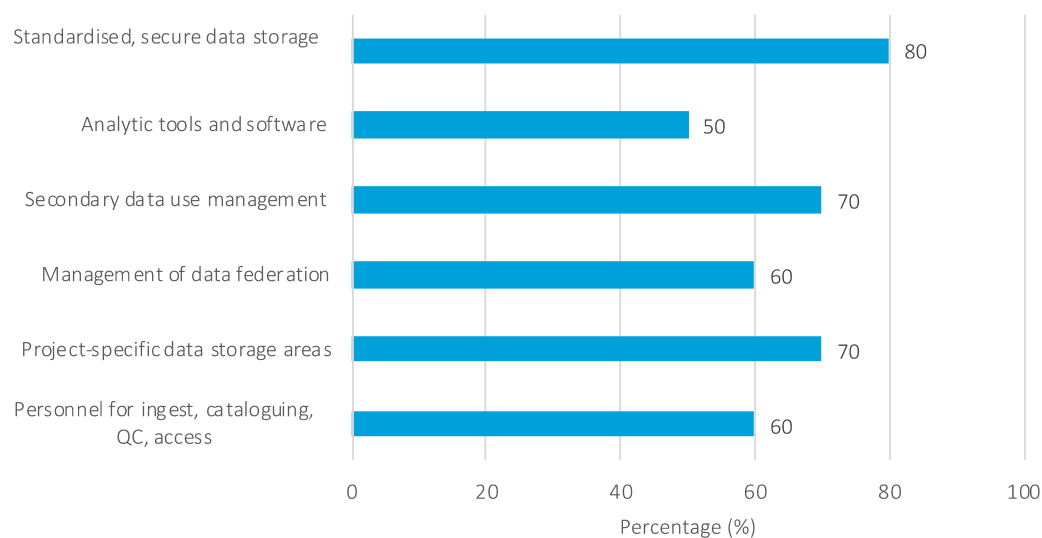
The most frequently noted essential element, for infrastructure managers, was personnel for ingest, cataloguing, QC and access (94%); While fewer (60%) of infrastructure users considered this essential, many indicated project specific data storage areas to be of importance (70%).

A high proportion of both managers and users agreed standardised secure data storage (88% and 80%, respectively), and secondary data use management (88% and 70%, respectively) were essential elements of a future NGIS.

**Table 5.1.1** Essential Components for Managers



**Table 5.1.2** Essential Components for Users



### *Additional Essential Components:*

Infrastructure managers noted additional essential elements for a future NGIS, including:

- Data sharing processes
  - Data sharing agreements for across jurisdictions
  - Clear pathways for sharing clinical data, for clinical and research use
- Data management aspects
  - National guidelines for Australian genomic data, to facilitate federating
  - Processes for data curation and off-boarding (to reduce & remove data)
  - International dataset replication
- Repository characteristics
  - Scalability
  - Working to international standards, internationally interoperable
- Dynamic consent
- Training
- Clear funding mechanism

#### *5.1.3 Infrastructure Users' Perspectives on Essential Components*

##### ***Essential Elements:***

- *Standardised metadata for describing datasets*
- *Guidelines on obtaining consent for data sharing*
- *Capability for granular browsing, cataloguing and selecting of datasets of interest (e.g. by data type, sequencing technology or phenotype), for external data users and dataset owners*

## 5.2 Existing Systems and Software to Incorporate in a Future NGIS

Infrastructure managers and users were asked what existing systems and software should be incorporated in a future NGIS. Responses are summarised below.

**Table 5.2.1** Systems and Software for Inclusion in Future NGIS

	System/Software
Research Infrastructures	Bioplatforms and Australian BioCommons
	National Collaborative Research Infrastructure Strategy (NCRIS) computing and storage facilities
	Australian Research Data Commons (ARDC)
	NeCTAR research cloud
	Major HPC centres, including the National Computational Infrastructure (NCI)
Clinical Infrastructures	National and state health organisations (NSW Health, the Department of Health)
	Clinical genomics organisations and infrastructures (Melbourne Genomics, GenoVic)
	Specialist clinical services (for cancer, Peter MacCallum centre)
Tools and Applications	Commercial cloud, storage and analysis (AWS, Seven Bridges)
	In-house tools and apps for analysis (Garvan workflow, tools)
	Data sharing tools (DaSH)
	Variant curation and sharing systems (VariantGrid, Shariant)
	QC tools (qProfiler)
	Data Management (Graphli)

### Additional feedback:

- Develop tools for national use, to address high costs of commercial tools.
- Leverage or use existing services, platforms and governance frameworks.
- Ensure ease of use.

## 5.3 Willingness to Pilot a National Genomics Infrastructure Service (NGIS)

Of those who responded to this question, most infrastructure managers (14/15) and infrastructure users (7/10) indicated willingness to participate, or support their institute's participation, in a pilot NGIS.

For some managers and users, this was conditional upon:

- Participating through relevant existing partners (e.g. Australian Biocommons)
- Suitability of the infrastructure for their areas of focus (e.g. cancer data)

Reasons for declining included negative past experience archiving data with similar overseas services.

## 5.4 Willingness to Contract Services from an NGIS

Over 60% of responding infrastructure managers (8/13) would contract, or consider contracting, services from a future NGIS. A further 31% (4/13) managers indicated their decision would depend on practical aspects, such as cost and governance agreements.

Of infrastructure users responding to this question, 75% (6/8) of the users would support their institutes contracting an NGIS.

Willingness to contract the NGIS depended on a range of cited factors, summarised in the table below.

**Table 5.4.2** Factors Affecting Willingness

Factor	
Governance Processes	Presence of data sharing and governance agreements
	Ethics
	A Privacy Security Assurance Framework (PSAF)
	Data security
Costing Factors	Cost and model, or organisations own future business model
	Egress charges to move data to a different environment
	Institute-level cost decisions, with individual researchers unable to directly fund access
	The need for it to be coupled with national-level or infrastructural project investment
Repository Elements	Ease of use of the interface
	Capability to customise QC
	Comparable in quality, comprehensiveness and tooling, to existing internal infrastructure
	Ability to keep track of samples and combine with existing data portals, to avoid duplicating

*Additional comments on interest in a potential NGIS, included:*

- As a *secondary* repository and sharing mechanism, or for integrating large genomic datasets, rather than replacing existing local infrastructure (noted by four organisations)
- As a managed system with transparent costs, liabilities, responsibilities and clear legal and ethical framework
- As an opportunity to implement agreements at the institutional level, rather than researcher-level
- As an opportunity to address existing ad hoc systems, with different requirements & high administrative burdens
- As a mechanism for providing essential data sharing

*Reasons for potential declining included:*

- Lack of available funding to do so
- Unwilling to pay for using a national facility, should be available at no charge
- Obligations to use existing infrastructure, with costs already built into procured services
- Lack of obligation to share data


















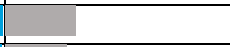




## 6 International Survey Comparison

International genomic infrastructures were also surveyed in parallel for an additional report [International Genomic Data Infrastructures (2020)], with 17 international infrastructures responding. These organisations included: national precision medicine initiatives, cohort infrastructures, access and archiving platforms, and variant databases.

Comparisons between the international and domestic survey responses are summarised below.

### 6.1 Comparisons of Core Infrastructure Elements


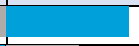






**Table 6.1.3** Comparisons of Core Infrastructure Elements Internationally and Domestically, by Percentage (%)

Element		International	Domestic
Federated Infrastructure			
Cloud			
Hybrid			
FASTQ Stored			
FASTA Stored			
Long Term Storage	Read-level data		
	Reports		
	Results		
Consent & Data Use Permissions Stored			
Standardised Clinical Terminologies			

### 6.2 Comparisons of Infrastructure Processes

International infrastructures, broadly, had more processes in place to support data sharing, including: data sharing by download or data-visiting, storage of consent and data use permissions, availability of publicly discoverable information, submission of workflows and queries by external users.

**Table 6.2.4** Comparisons of Infrastructure Processes Internationally and Domestically, by Percentage (%)

Element	International	Domestic
Harmonised Data Processing Mandated		
External Data Sharing Stored		
Workflow Submissions by External Users Supported		
Querying Repository Contents Supported		

### 6.3 Comparisons of Resourcing and Requirements

*International infrastructures:*

- A greater proportion are larger initiatives (in terms of data usage and human resourcing requirements), including:
  - A greater proportion storing 5,000 TB or more
  - A higher % expecting large future storage increase of 1,000 TB or more per year
  - A higher % requiring 15 FTE or more

- Primarily funded by government, compared to domestically, where funding source for the infrastructures varied by organisation type (research institute infrastructures funded by internal operating costs, clinical infrastructures funded by state health departments, and service-based research platforms operating cost recovery through data owner charges)
- Are not typically charging data depositors for use of the infrastructure, similar to domestically where few incur download or analysis charges, and less than 50% charge for storage.

## 6.4 Comparisons of Evaluations for Current and Future Repository Elements

Comparisons of themes from responses of the international and domestic surveys around current repository elements are provided below in **Table 6.4.1 International and Domestic Infrastructure Survey Response Themes (Part I)**, and **Table 6.4.2 International and Domestic Infrastructure Survey Response Themes (Part II)**.

**Table 6.4.5 International and Domestic Infrastructure Survey Response Themes (Part I)**

Australian Infrastructure Managers and Users		International
Data Ingestion Challenges	<p>Experienced by 53% of managers and 50% of users.</p> <p><i>Types of ingestion challenges</i></p> <ul style="list-style-type: none"> <li>• Data size, format and (missing) requirements *</li> <li>• Connectivity: network speed/bandwidth, firewalls</li> <li>• Time, procedural and resource requirements</li> </ul> <p><i>Those without challenges had:</i></p> <ul style="list-style-type: none"> <li>• Automated processes</li> <li>• Co-located sequencing with storage</li> <li>• High-speed links</li> <li>• Ingest done on their behalf</li> </ul>	<p><i>Governance and Operations</i></p> <ul style="list-style-type: none"> <li>• Requires dedicated personnel</li> </ul> <p><i>Technical and Repository</i></p> <ul style="list-style-type: none"> <li>• Bandwidth requirements</li> <li>• High speed and secure network requirements</li> <li>• Software for data transportation</li> <li>• Encryption and transfer of large datasets</li> </ul> <p><i>Data challenges</i></p> <ul style="list-style-type: none"> <li>• Adherence &amp; availability of standards for metadata</li> <li>• Phenotype quality</li> </ul>
Next Steps	<p>Scale up (2/3 of the infrastructures) e.g. adding cloud</p> <p>Analytic tools, new data types, better data standards</p> <p>Adopting (inter)national frameworks and strategies</p> <p>Data Sharing, AAI and FAIR data</p>	<p>Scaling up infrastructure &amp; New architecture</p> <p>Cloud-based infrastructure</p> <p>Federated models</p> <p>Technical &amp; ELSI* aspects for data sharing</p> <p>*Ethical Legal and Social Implications</p>

Scale Up & Adoption: Limitations	<p><i>Resourcing: Funding, costs, personnel; (77%)</i></p> <p><i>Governance and Operational</i></p> <ul style="list-style-type: none"> <li>Frameworks for data ownership, data sharing, governance</li> <li>Workforce skills (in lab, IT, health department)</li> <li>Managing consent</li> </ul> <p><i>Technical, Repository and Data</i></p> <ul style="list-style-type: none"> <li>No existing solutions for: national AAI, national archiving, front-ends, mature cloud systems</li> <li>Network connectivity</li> <li>Data security</li> </ul>	<p><i>Resourcing: Funding, funding uncertainty, costs (50%)</i></p> <p><i>Governance and Operational</i></p> <ul style="list-style-type: none"> <li>Data interoperability and ELSI issues</li> <li>Different data sharing policies across countries</li> <li>Requirements to retain data locally</li> </ul> <p><i>Technical, Repository and Data</i></p> <ul style="list-style-type: none"> <li>No solutions for scalable warehouse/genomic database</li> <li>Requirements for controlled access infrastructure</li> <li>Compute resourcing</li> </ul>
----------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Table 6.4.6 International and Domestic Infrastructure Survey Response Themes (Part II)**

Australian Infrastructure Managers and Users		International
Best Elements	<p><i>Governance and Operations</i></p> <ul style="list-style-type: none"> <li>Administrated or managed by others (general, data sharing, access control)</li> <li>Local, in-house; Centralised; accessible to many organisations or researchers</li> <li>Privacy and Security compliant</li> <li>Affordability or absence of cost</li> </ul> <p><i>Technical Repository Characteristics</i></p> <ul style="list-style-type: none"> <li>Flexibility, configurability, usability, scalable processes</li> <li>Programmatic accessibility and web-based interfaces</li> <li>Computational capability, speed, performance; co-location with storage</li> <li>Access to computational tools and optimised pipelines</li> </ul> <p><i>Value Creation</i></p> <ul style="list-style-type: none"> <li>Enabling the delivery of new health and biology insights</li> <li>Support x-disciplinary collaboration and innovations</li> </ul>	<p><i>Standards and Interoperability</i></p> <ul style="list-style-type: none"> <li>Adherence to standards – for future harmonising, for submission to access archives</li> </ul> <p><i>Technical Repository Characteristics</i></p> <ul style="list-style-type: none"> <li>Co-located compute and storage</li> <li>Scalable, secure and elastic, through cloud</li> </ul> <p><i>Value and Resource Creation</i></p> <ul style="list-style-type: none"> <li>Enables data retention for future use: knowledge databases, data platforms, resources for healthcare/research</li> <li>Enables data sharing and collaborations; promotes cultural change</li> </ul>

Potential Improvements	<p><i>Governance and Operations</i></p> <ul style="list-style-type: none"> <li>• More (reliable) funding; Sustainable infrastructure;</li> <li>• Roadmap for national services; national/standardised processes for data management</li> <li>• Centralise resources, address fragmentation</li> </ul> <p><i>Data Sharing</i></p> <ul style="list-style-type: none"> <li>• Resolve data sharing and governance challenges for users (clinical, research, non-expert)</li> <li>• Better permission and access control management</li> <li>• Data discoverability, search tools / front ends for querying repository and data</li> </ul> <p><i>Repository and Data Elements</i></p> <ul style="list-style-type: none"> <li>• Expansions – more compute nodes, GPU, cloud, scalability for whole genomes</li> <li>• Improved data quality, phenotype capture (in HPO), metadata</li> <li>• Better integration, programmatic interfaces, back ends for ingress from instruments</li> <li>• Improve data transfer processes, and upload/download challenges to repositories</li> </ul>	<p><i>Data Sharing</i></p> <ul style="list-style-type: none"> <li>• Trans-national data access</li> <li>• Federated structures, including at multi-country or international levels</li> <li>• User interfaces for more granular data access</li> </ul> <p><i>Repository and Data Elements</i></p> <ul style="list-style-type: none"> <li>• Harmonised genomic data, harmonised phenotype data</li> <li>• Standardised ingest</li> <li>• Expansion to other data types e.g. single cell</li> <li>• More efficient and scalable genotype querying</li> <li>• Processing for allele frequencies</li> <li>• Usability and flexibility in secure environment</li> </ul>
------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

## 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.*