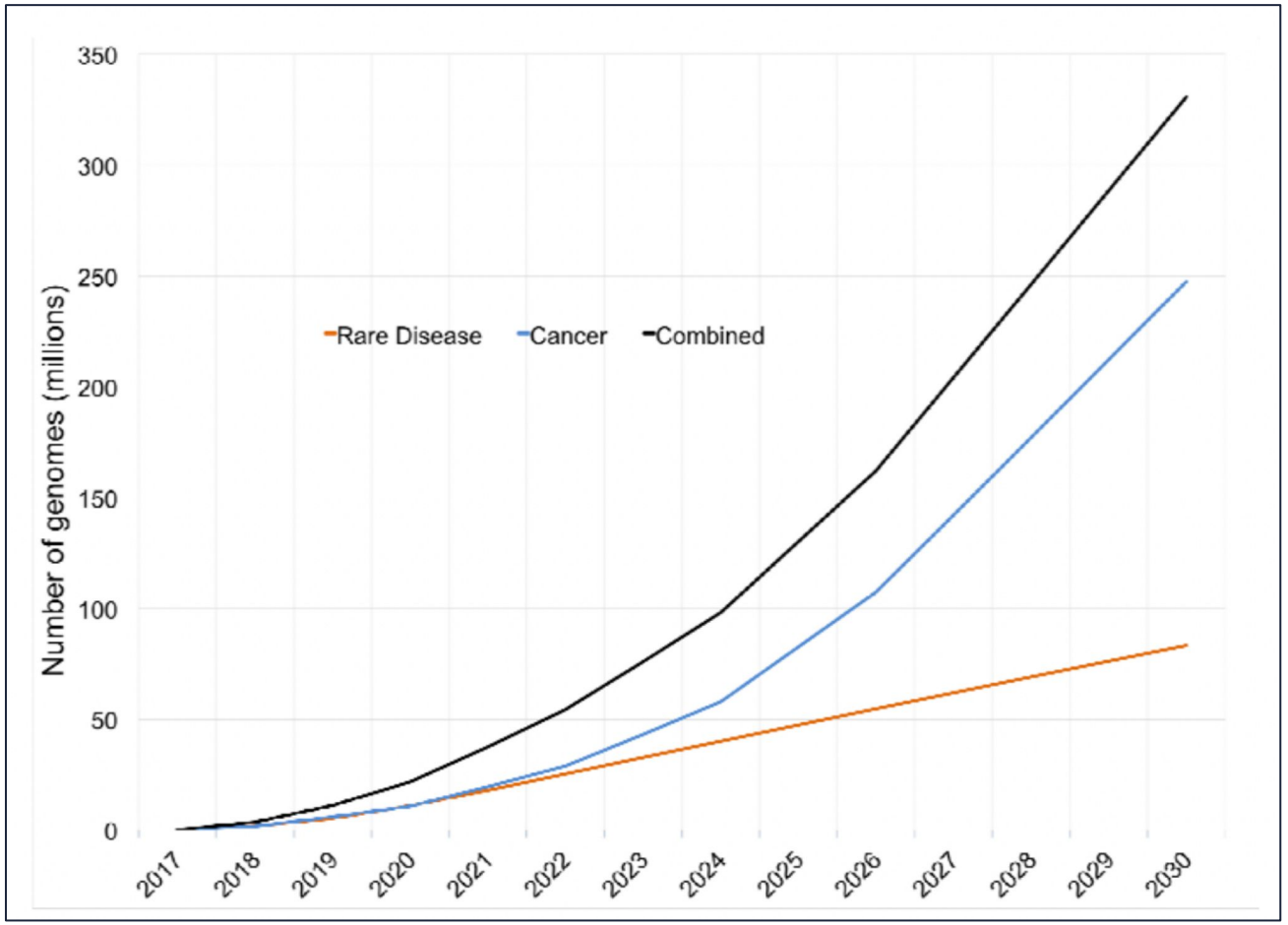
Human genomic data sharing and analysis

NCI-Fujitsu HPC, Cloud and Data Futures Workshop

A/Prof Bernie Pope
Associate Director, Human Genome Informatics
Australian BioCommons
Victorian Health and Medical Research Fellow
Melbourne Bioinformatics
The University of Melbourne
bernie@biocommons.org.au

Australian BioCommons

Predicted global growth of healthcare funded sequenced human genomes



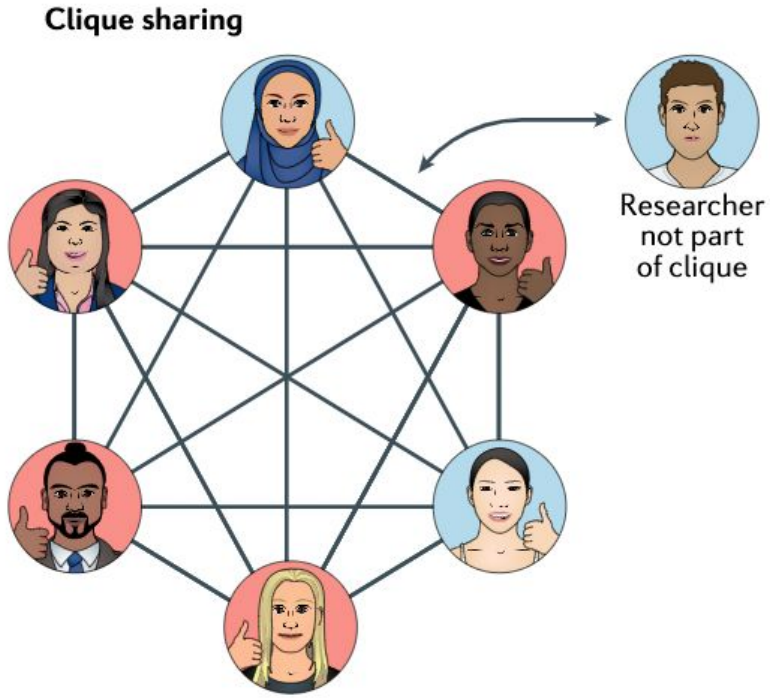
Global storage requirements in 2025 to be exabytes to low zettabytes.

Birney, E., Vamathevan, J., and Goodhand, P. (2017). Genomics in healthcare: GA4GH looks to 2022. bioRxiv

Australian BioCommons

Siloed data

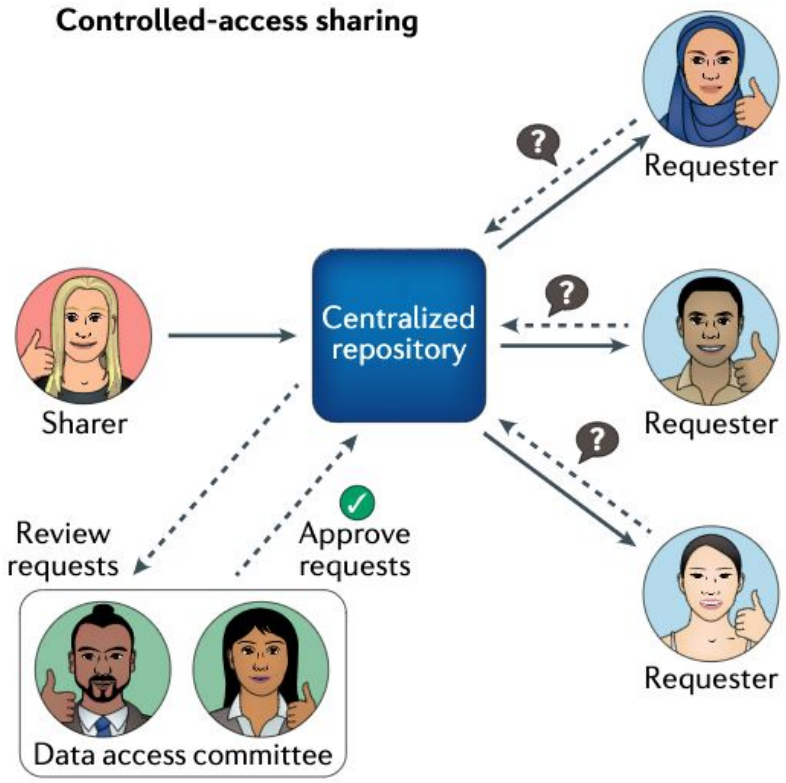
- Human genomics data has often been siloed.
- This limits reuse and reanalysis.
- Public benefit is increased when data is shared.
- Sharing is frequently necessary in human health, especially in rare disease and cancer.
- Large cohorts are needed for statistical power.
- National and international data sharing is highly beneficial but requires considerable collaboration and coordination.



Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X. & Greene, C. S. Responsible, practical genomic data sharing that accelerates research. Nat. Rev. Genet. 21, 615–629 (2020).

Australian BioCommons

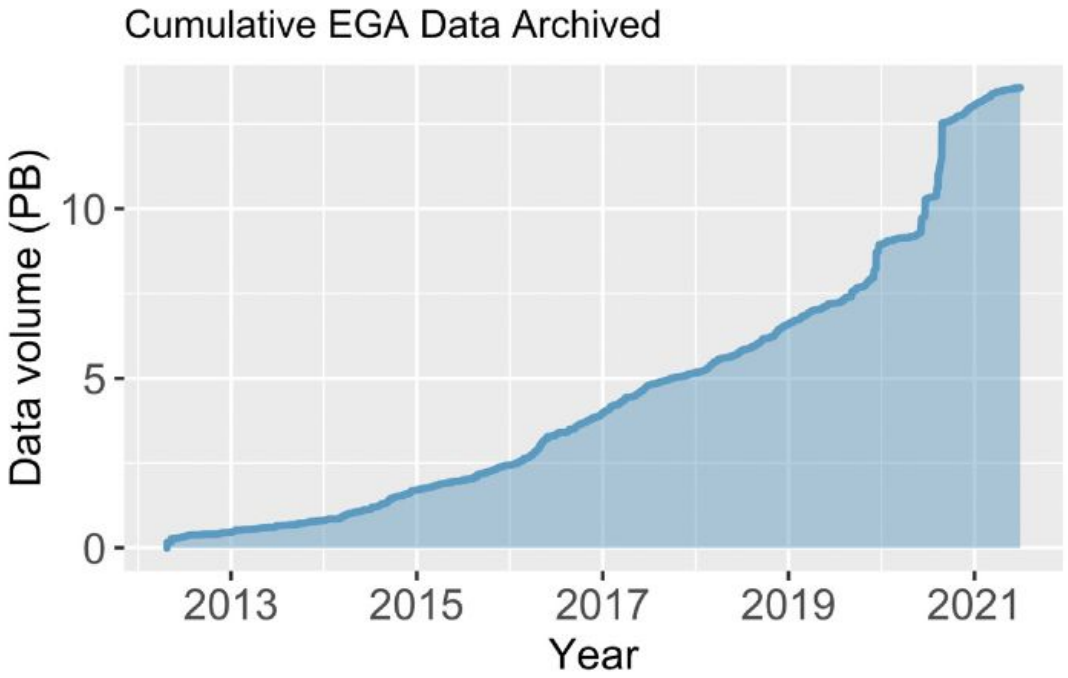
Centralised repositories

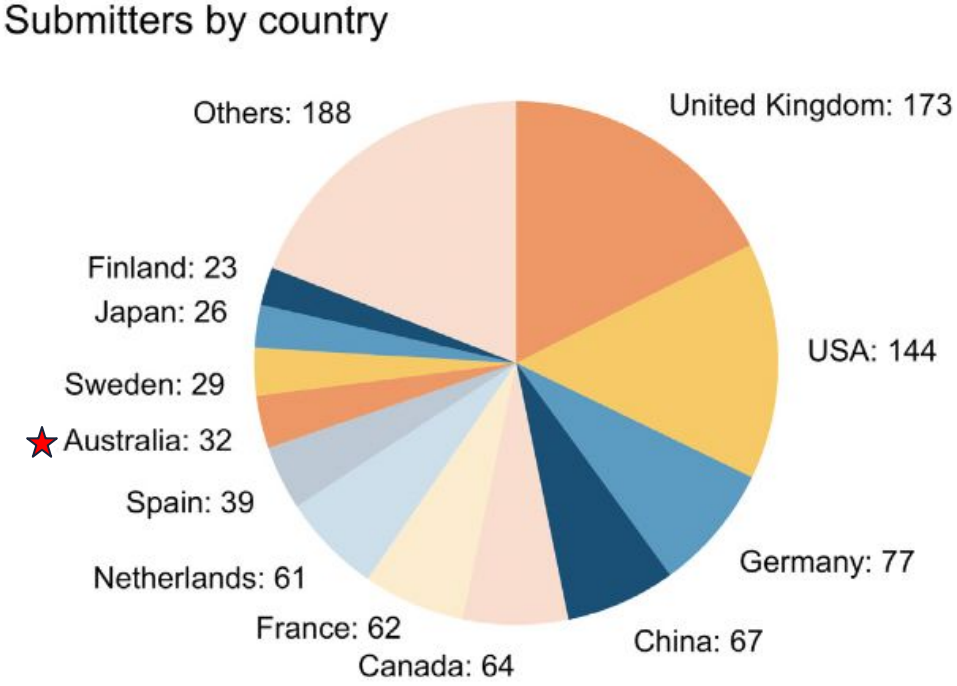


Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X. & Greene, C. S. Responsible, practical genomic data sharing that accelerates research. Nat. Rev. Genet. 21, 615–629 (2020).

- Centralised repositories such as EGA (Europe) and dbGaP (USA) have served as default mechanism for sharing research genomics data.
- Centralisation has limited scalability.
- Access favours European and USA users.
- Focus is on research data, associated clinical data is generally minimal and this limits reuse.
- Connection to analytics platforms has not been a focus of these repositories.

Transition to federated data sharing





The EGA is currently transitioning from a centralised resource to a federated node model.

Freeberg, M. A. et al. The European Genome-phenome Archive in 2021. Nucleic Acids Res. (2021)

Australian BioCommons

Towards global standards for data sharing

- The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework.
- Australian Genomics is a driver project of GA4GH.
- A key outcome is the specification for standard APIs for data sharing technology.
- Recognition that the data life cycle in human genomics is complex and data storage and analysis are parts of a bigger ecosystem.

Rehm, H. L. *et al.* GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics* 1, (2021).

Australian BioCommons

Key challenges and opportunities

Legal, ethical, privacy and security concerns

- jurisdiction boundaries and data sovereignty
- encryption and anonymisation
- approval for research and notification of secondary findings
- authentication and authorization of research users
- data access approval
- consent and appropriate use of data

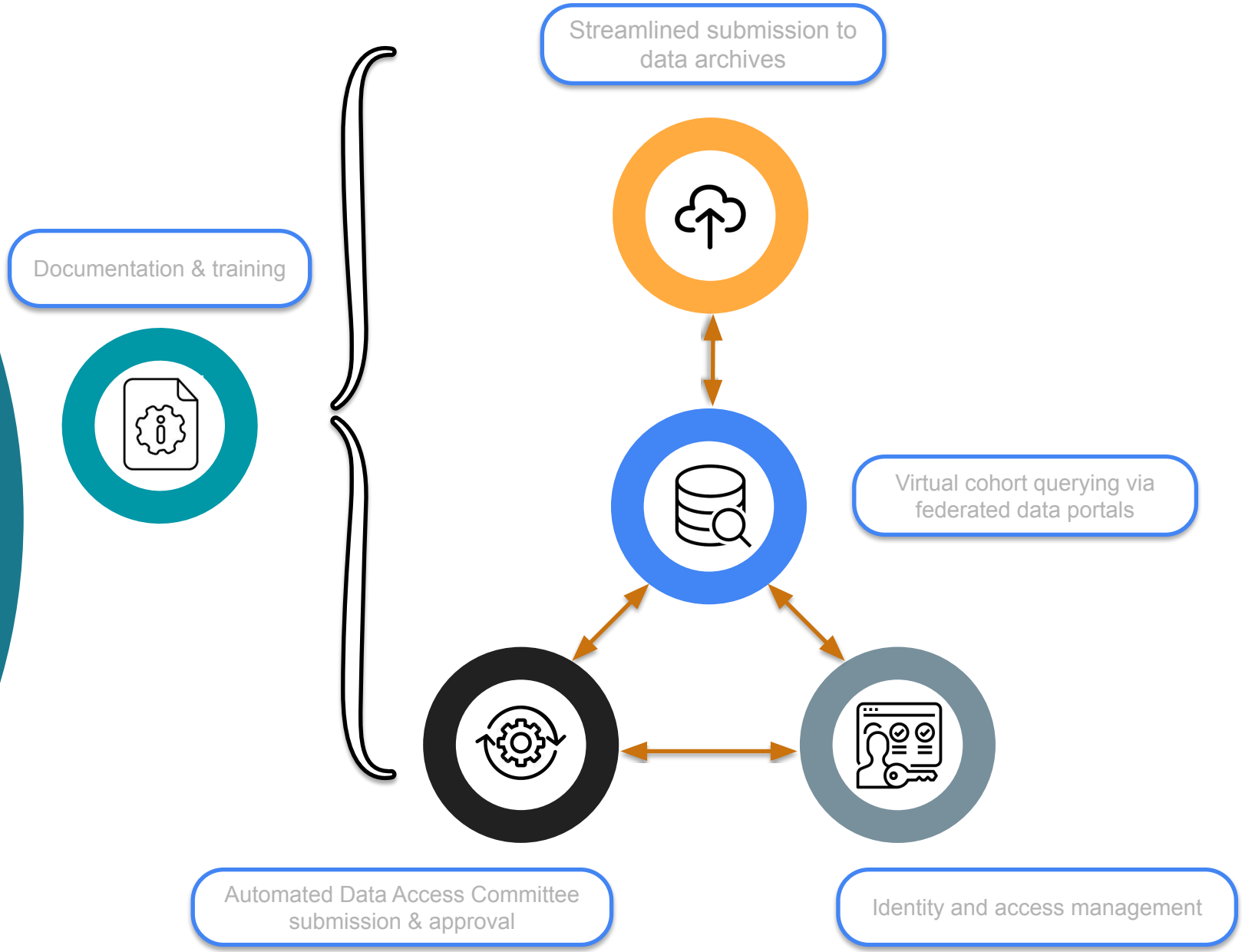
Infrastructure concerns

- data volume
- data aggregation and cohort analyses for statistical power
- harmonised data and standardised/portable analysis workflows
- cost

Human genome data sharing examples in Australia

- Australian Genomics
- Melbourne Genomics
- Queensland Genomics
- Medical Genome Reference Bank (MGRB)
- Centre for Population Genomics, The Garvan Institute of Medical Research
- QIMR Berghofer
- University of Melbourne Centre for Cancer Research
- ZERO Childhood Cancer
- Australian Cardiovascular Alliance
- Human Genomes Platform Project
- many more

HUMAN GENOMES PLATFORM PROJECT

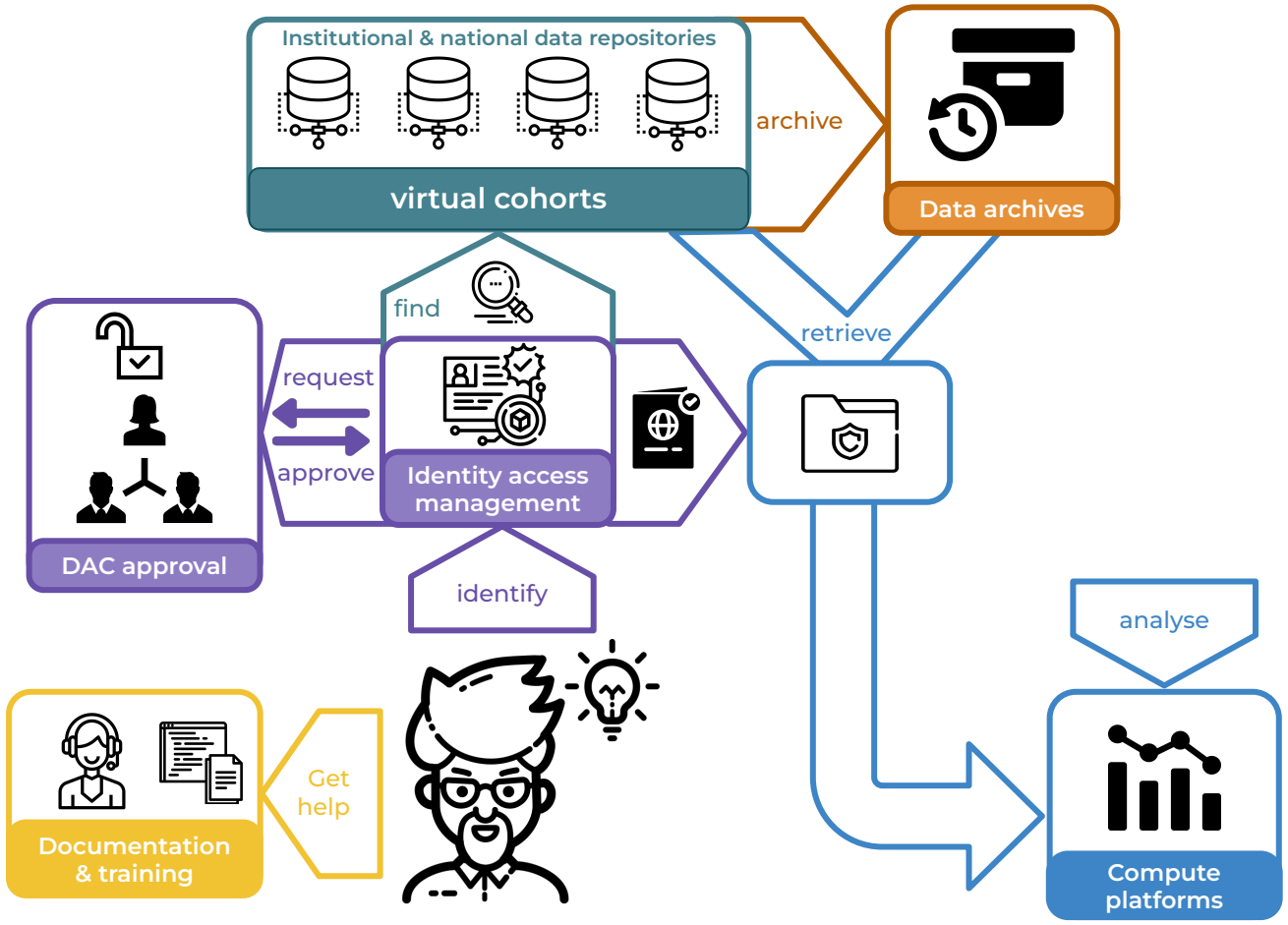


Project Partners

- Australian Access Federation
- Australian BioCommons
- ARDC Australian Research Data Commons
- BioPlatforms Australia
- Garvan Institute of Medical Research
- NCI Australia
- QIMR Berghofer Medical Research Institute
- The University of Melbourne
- zero Childhood Cancer

Jess Holliday and Marion Shadbolt

Infrastructure ecosystem



Example solutions:

- IAM: CILogon, GA4GH passports
- Data commons: Gen3
- DAC approval: REMS
- Analytics: national infrastructure, institutional infrastructure, commercial cloud
- Integrated: Broad Terra + DUOS

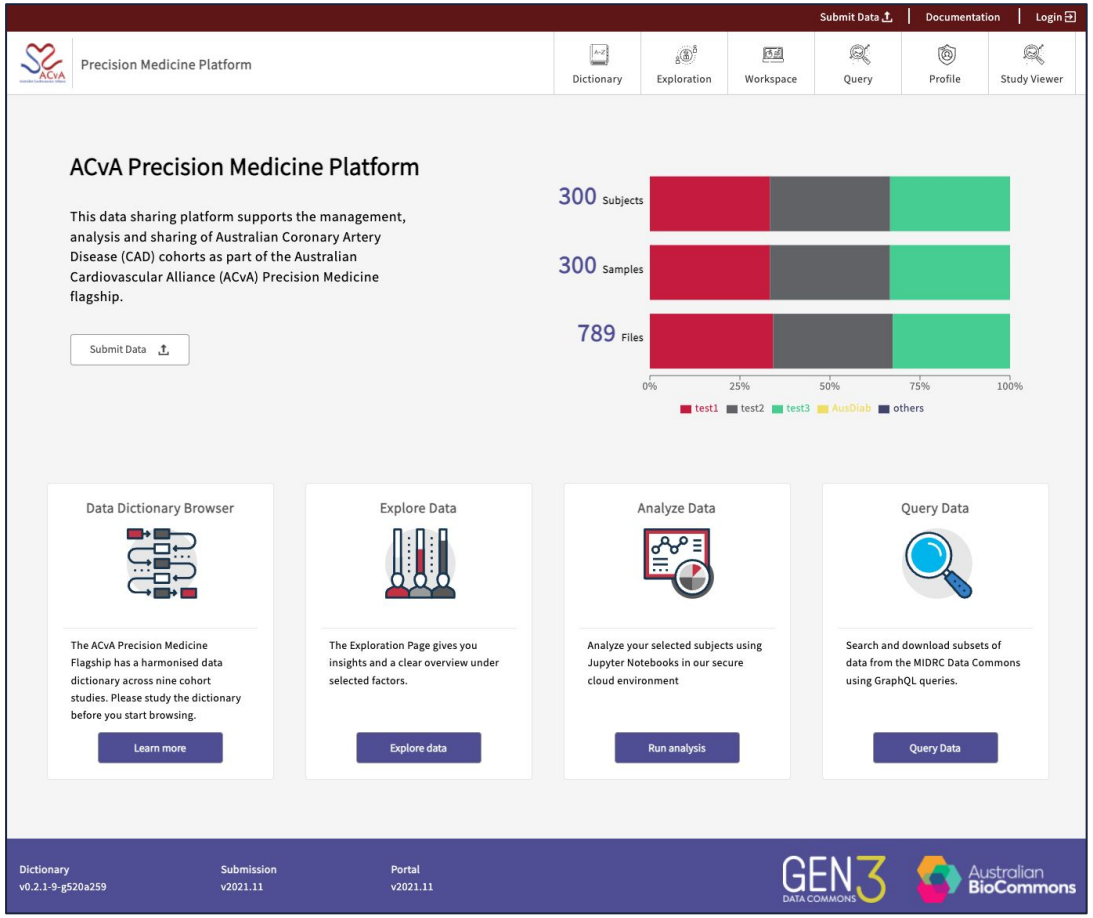
Marion Shadbolt

Australian BioCommons

An instance of this ecosystem

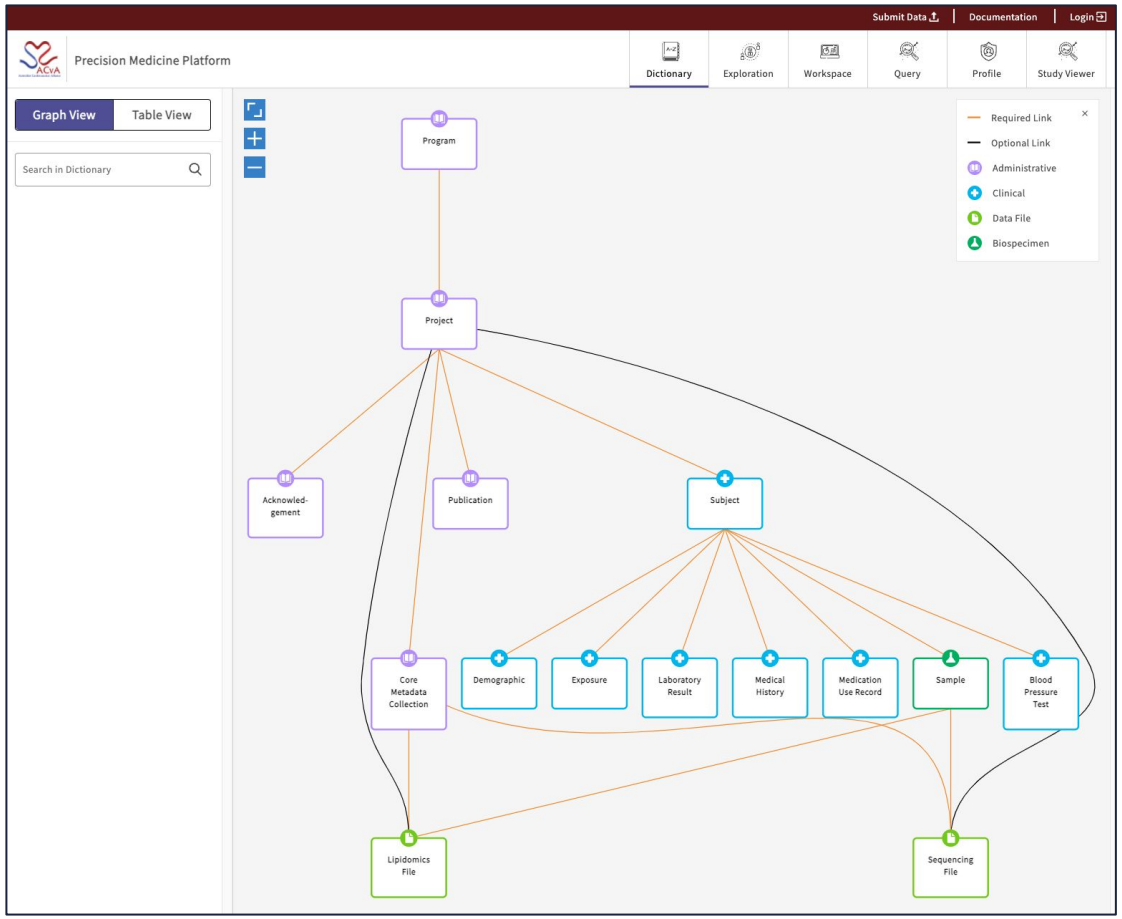
Australian Cardiovascular Alliance (ACvA), precision medicine platform

- Pilot project using Gen3 technology to build a data commons for Coronary Artery Disease (CAD) cohorts



Australian BioCommons

Harmonised data dictionary

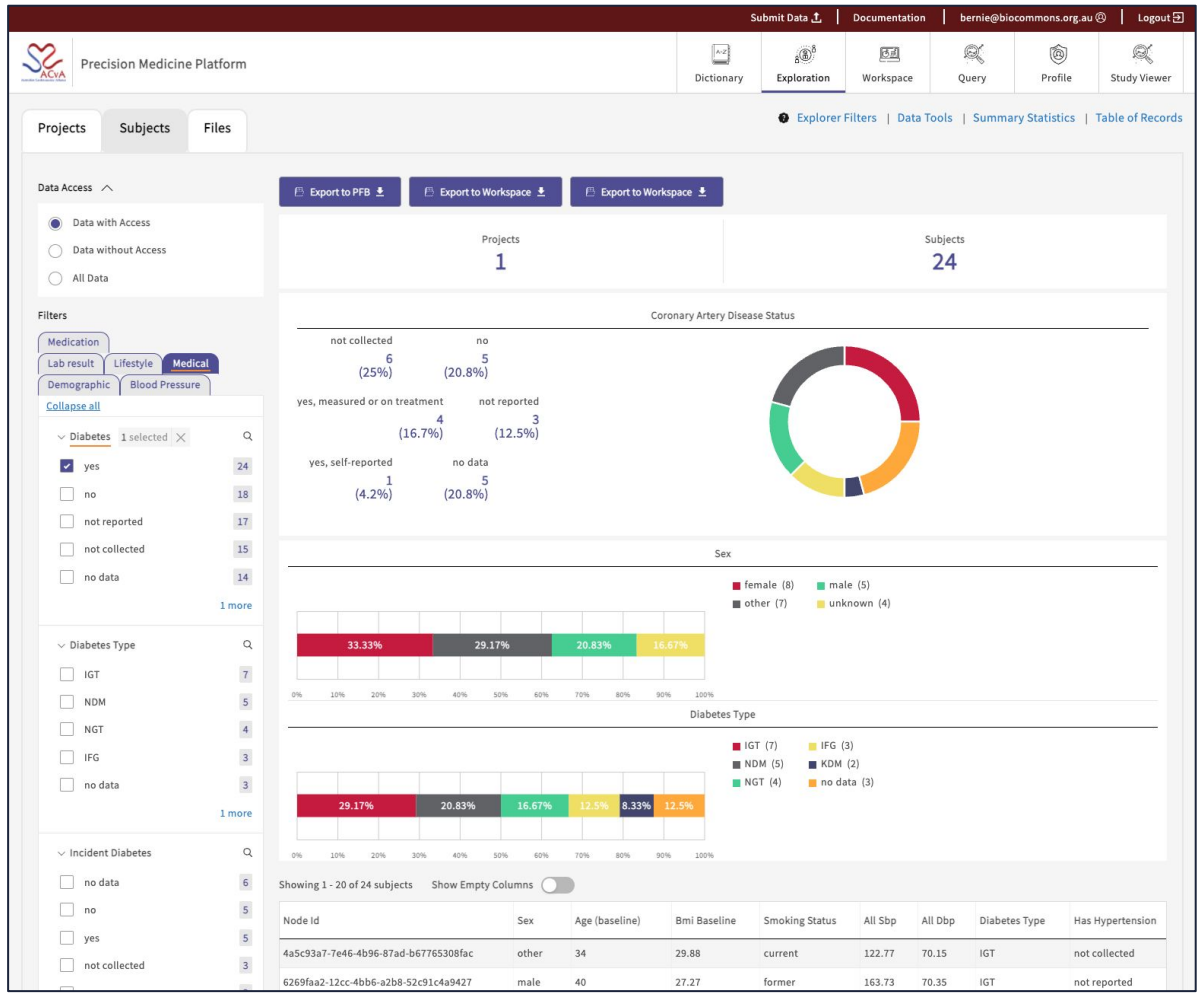


- A data dictionary is a hierarchical description of the dataset
- Variables in the dictionary must be harmonised across cohorts to allow coherent analysis
- Harmonisation will also align with international standards to allow integration with significant international data holdings

Australian BioCommons

Data query and cohort building

- Authorised users can use the Gen3 interface to search for data satisfying constraints
- A study cohort can then be downloaded or sent to an computing platform for analysis *
- Our vision is that standard APIs will ultimately allow data query across multiple data commons, leading to the construction of virtual cohorts nationally and internationally



(*) to be implemented

Australian BioCommons

A national approach to genomics information management (NAGIM)

- The vision for human genomics data sharing in Australia requires considerable coordination and collaboration.
- The NAGIM Blueprint sets out a series of principles to guide decision-making on the responsible collection, storage, use and management of genomic data.
- Australian Genomics is developing recommendations for implementing NAGIM.
- In 2021 Australian Genomics led an implementation prototyping phase in response to NAGIM.
- A panel of external assessors are evaluating prototype submissions presently.

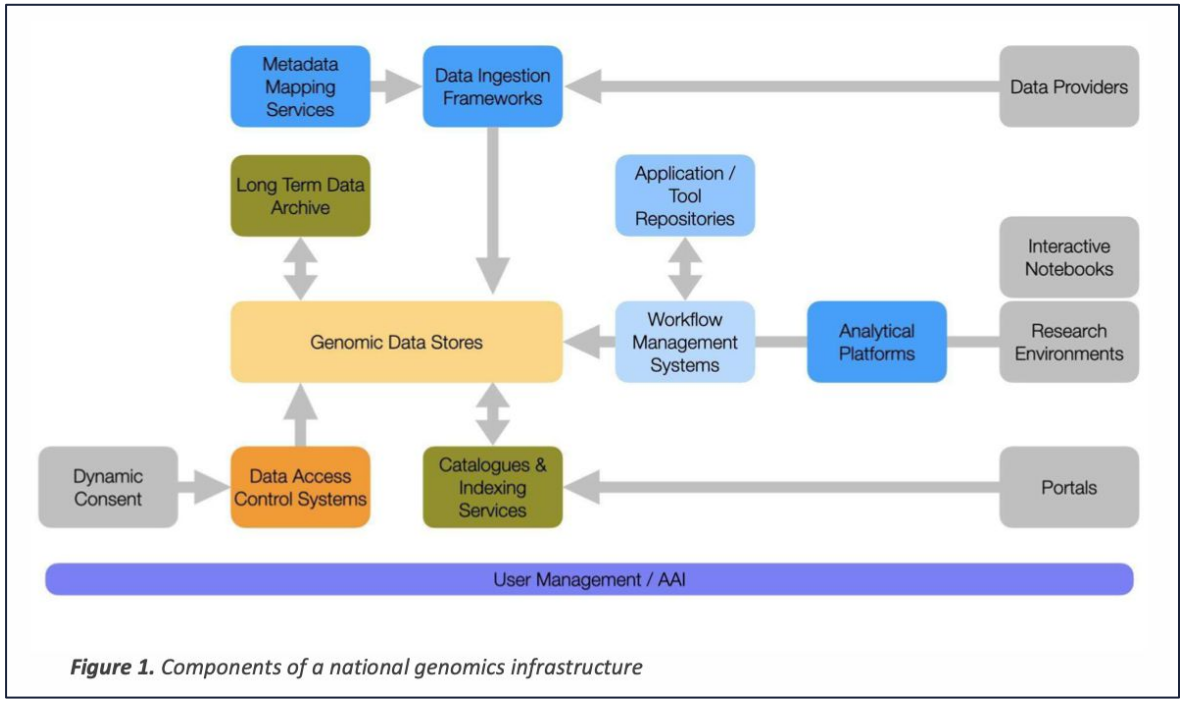


Figure 1. *Components of a national genomics infrastructure*

A National Approach to Genomic Information Management, Australian Genomics Implementation Recommendations Progress Report, November 2021

Australian BioCommons

Integrated high-performance computing, data and cloud systems

- Standardisation and integration are key
- Data is large and will be federated across cloud and institutional storage
  - cloud compatible storage systems are desirable
- Compute will exist within a complex ecosystem of platforms, users and sharing modes
  - streamlined national IAM approach is desirable, such as CILogon
- Harmonised computation across computing platforms is necessary
  - support for containerised tools and portable workflows is needed
- Workloads are more irregular and difficult to manage within traditional merit allocation schemes
  - flexible resource allocation and management

Acknowledgements

- The Australian BioCommons
  - Andrew Lonie
  - Marion Shadbolt
  - Jess Holliday
  - Steven Manos
  - Rhys Francis
  - Jeff Christiansen
- The University of Melbourne, Centre for Cancer Research; Australian Genomics
  - Oliver Hofmann
- NCI
  - Matthew Downton
- Baker Heart and Diabetes Institute
  - Peter Meikle

Australian BioCommons