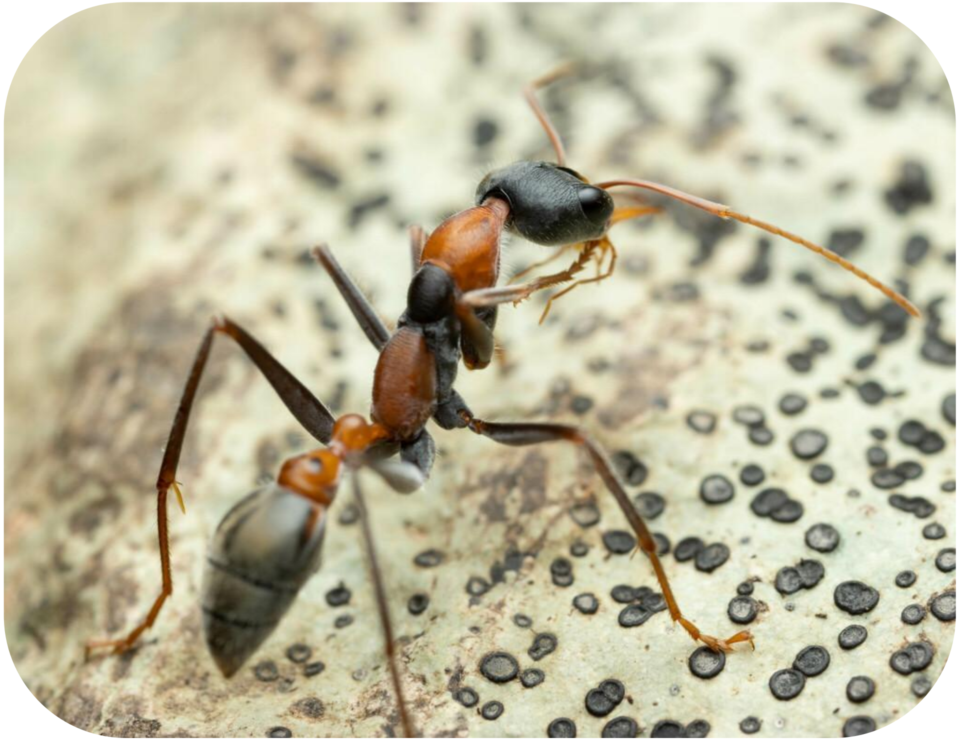
A FAIRer future for genomics research in Australia: creating an Australian Reference Genome Atlas

K. Hall, J. Christiansen, N. dos Remedios, S. Richmond, N. Ward & H. Holewa

eResearch Australasia Conference, 19 October 2022



Myrmecia nigrocincta
Jumping Jack Ant

https://inaturalist.ala.org.au/photos/236637920

Myrmecia nigrocincta
Jumping Jack Ant

© Mark Ayers (CC-BY-NC)
https://inaturalist.ala.org.au/photos/236637920

> "More than a century of research has led to the identification of some key navigational strategies, such as compass navigation, path integration, and route following. Ants have been shown to rely on visual, olfactory, and idiothetic cues for navigational guidance."

Freas, C.A. & Schultheiss, P. (2018) How to Navigate in Different Environments and Situations: Lessons From Ants. *Frontiers in Psychology*, **9**: 841, https://doi.org/10.3389/fpsyg.2018.00841.

Genomic sequencing and analysis have been identified by the Australian Government as being a critical technology for our national prosperity and security.

Key sectors

- Healthcare & Medical
- Agriculture
- Environment
- Defence & Defence Industry
- Energy

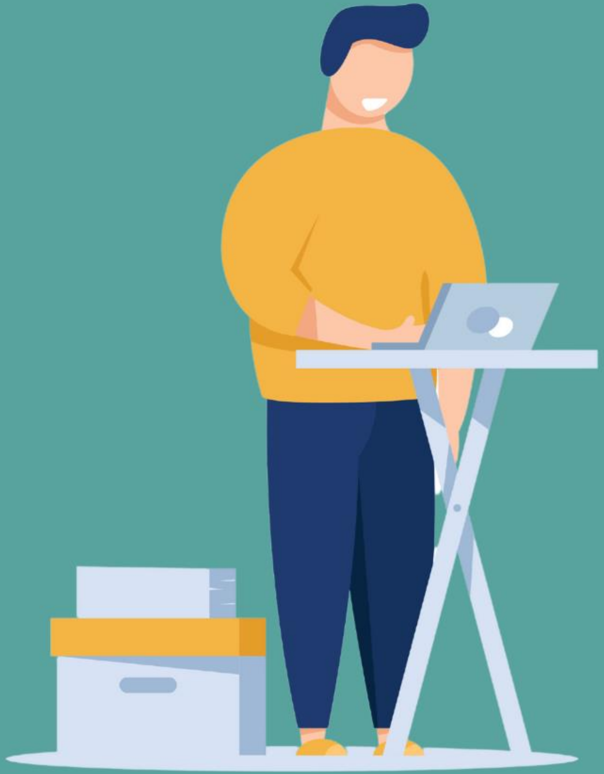
Estimated impact on national interest

- Economic Prosperity - High
- National Security - Med

Source: Department of Industry, Science and Resources, Australian Government, https://www.industry.gov.au/publications/action-plan-critical-technologies/tech-cards/genome-and-genetic-sequencing-and-analysis-next-generation-sequencing

Policy responses to bushfires (and another environmental catastrophe) responses can be proactive, not reactive, when driven by these data.

The Australian BioCommons estimates that there are currently 15,000 life science researchers in Australia currently using genomics data to answer research questions.



Genomics data sources are:

- nested
- hidden
- variable
- complex
- different
- scattered
- embedded
- distributed
- disconnected

GenBank (NCBI)

- created in 1982
- currently holds:

  traditional GenBank records:
  - 240,539,282 sequences
  - 1,562,963,366,851 bases

  set-based (WGS/TSA/TLS) records
  - 2,857,043,692 sequences
  - 18,787,298,109,534 bases
- acceptance standards hinge on genetic annotation, not taxonomic or other isolate metadata accuracy
- not peer-reviewed

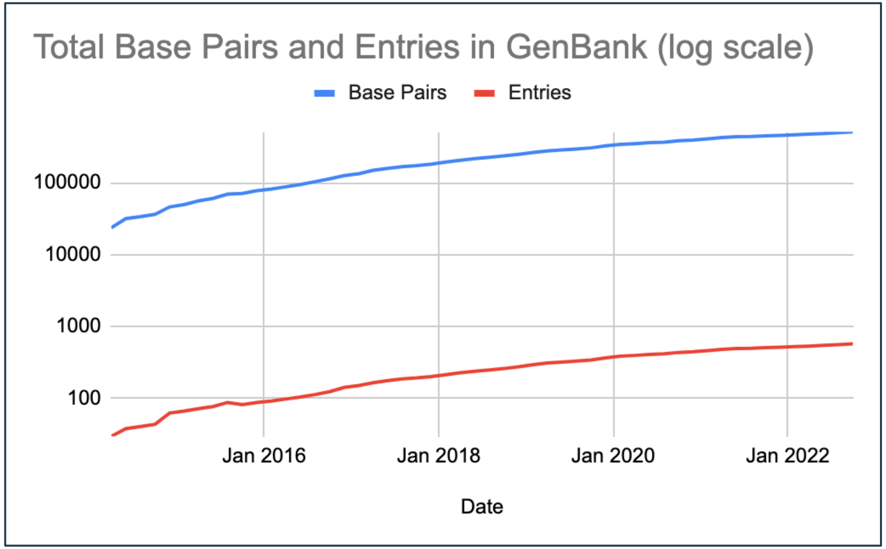
GenBank and EMBL NUCLEOTIDE SEQUENCES 1986/1987 VOLUME I ACADEMIC PRESS
GenBank and EMBL NUCLEOTIDE SEQUENCES 1986/1987 VOLUME II ACADEMIC PRESS
GenBank and EMBL NUCLEOTIDE SEQUENCES 1986/1987 VOLUME III ACADEMIC PRESS
GenBank and EMBL NUCLEOTIDE SEQUENCES 1986/1987 VOLUME IV ACADEMIC PRESS
GenBank and EMBL NUCLEOTIDE SEQUENCES 1986/1987 VOLUME V ACADEMIC PRESS
GenBank and EMBL NUCLEOTIDE SEQUENCES 1986/1987 VOLUME VI ACADEMIC PRESS
GenBank and EMBL NUCLEOTIDE SEQUENCES 1986/1987 VOLUME VIII ACADEMIC PRESS

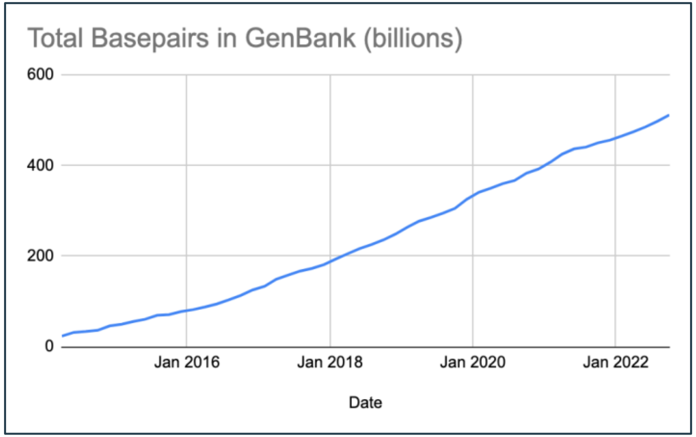
GenBank and EMBL database 1986/1987

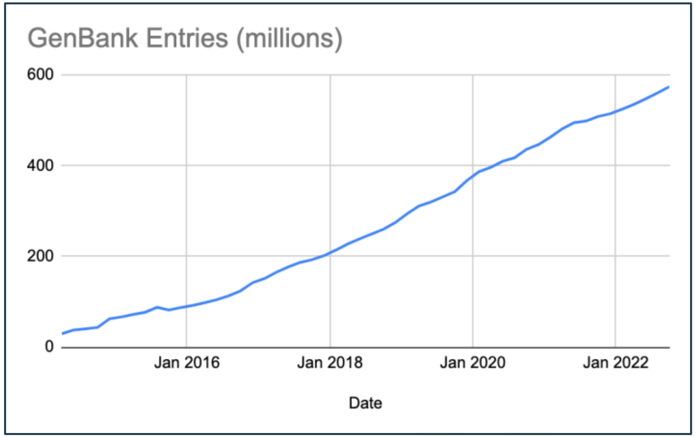
© David Landsman, Bethesda, Maryland (CC-BY)

https://twitter.com/bffo/status/289529886484348928

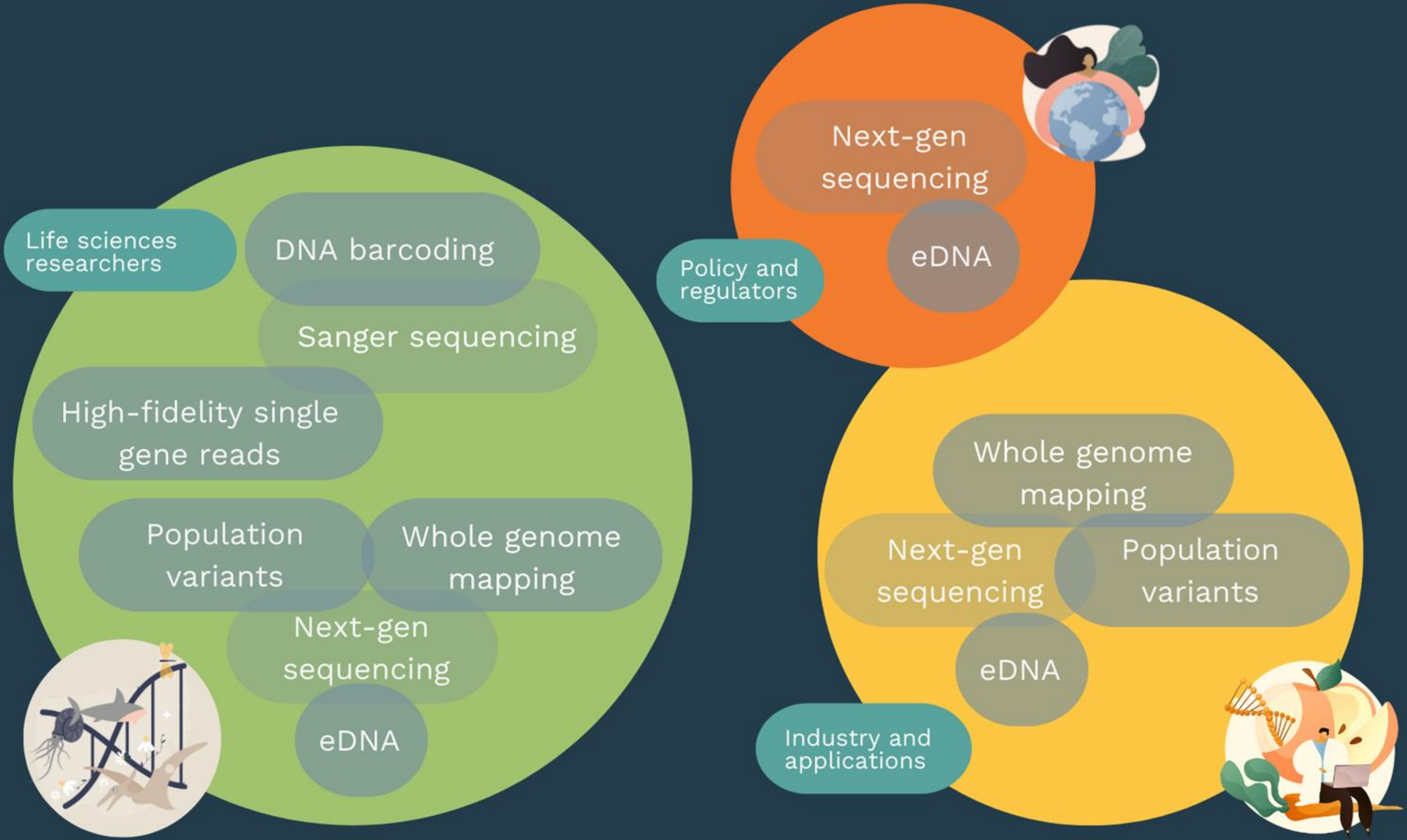
GenBank growth since 2014







Data source: NCBI-GenBank Flat File Release 252.0, Genetic Sequence Data Bank, October 15 2022, https://ftp.ncbi.nih.gov/genbank/gbrel.txt



Life sciences researchers

DNA barcoding

Sanger sequencing

High-fidelity single gene reads

Population variants

Whole genome mapping

Next-gen sequencing

eDNA

Policy and regulators

Next-gen sequencing

eDNA

Industry and applications

Whole genome mapping

Next-gen sequencing

Population variants

eDNA

100% of researchers have told us that they do not trust data they download from GenBank*

Turning to idiothetic cues

* may not actually be true

Current researcher strategies to find data

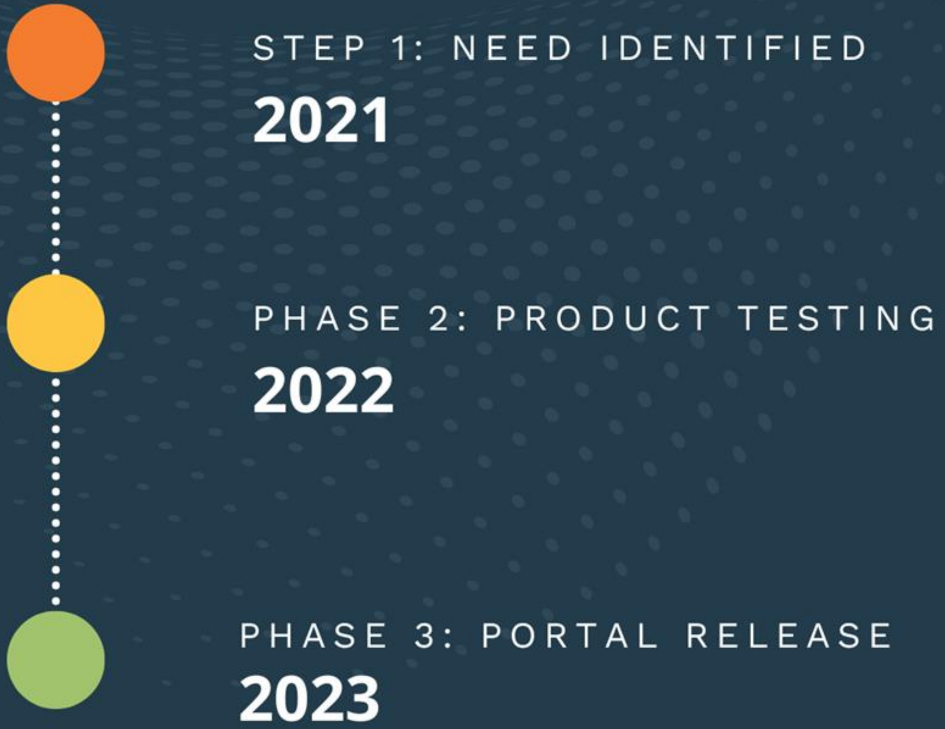
- GenBank
- BioPlatforms data portal
- Barcode of Life Database
- EMBL-ENA
- Project repositories and websites (*e.g.* Apollo, GoaT, Darwin Tree of Life, ReefGenomics)
- Taxon or community databases (*e.g.* WormBase, ANEMONE)
- Read papers and track back to repositories (Dryad, Zenodo, OSF)
- Word-of-mouth
- Own library generation

Current researcher strategies to trust data

- metadata for voucher in recognised and curated collection (*i.e.* a museum or herbarium)
- reputation of researcher generating and depositing data
- metadata about methods

ARGA concept development

Born out of researcher frustration at having to search across multiple data sources to find and access genomics data, ARGA aims to aggregate data from a number of reputed domestic and international sources in a single location.



STEP 1: NEED IDENTIFIED

2021

PHASE 2: PRODUCT TESTING

2022

PHASE 3: PORTAL RELEASE

2023

Acknowledging ARGA partnerships

The Australian Reference Genome Atlas (ARGA) is an NCRIS-enabled platform powered by the Atlas of Living Australia (ALA), in collaboration with Bioplatforms Australia and the Australian BioCommons, with investment from the Australian Research Data Commons (ARDC) (https://doi.org/10.47486/DC011). ARGA integrates data sourced from a number of international repositories, including NCBI GenBank, EMBL-ENA and Bioplatforms Australia.

ARGA
Australian Reference Genome Atlas

Atlas of Living Australia
ala.org.au

Australian BioCommons

BIOPLATFORMS AUSTRALIA

ARDC
Australian Research Data Commons

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

ARGA objectives and vision

The Australian Reference Genome Atlas is an indexing service for discovering, filtering and accessing complex life science data.

For plants, animals, microbiota and other species endemic or relevant to Australia, ARGA will build a platform to locate and aggregate genomic data, including:

- reference genome assemblies
- genome annotations
- population and variant sets
- DNA barcodes
- coding and non-coding DNA sequences

ARGA
Australian Reference Genome Atlas

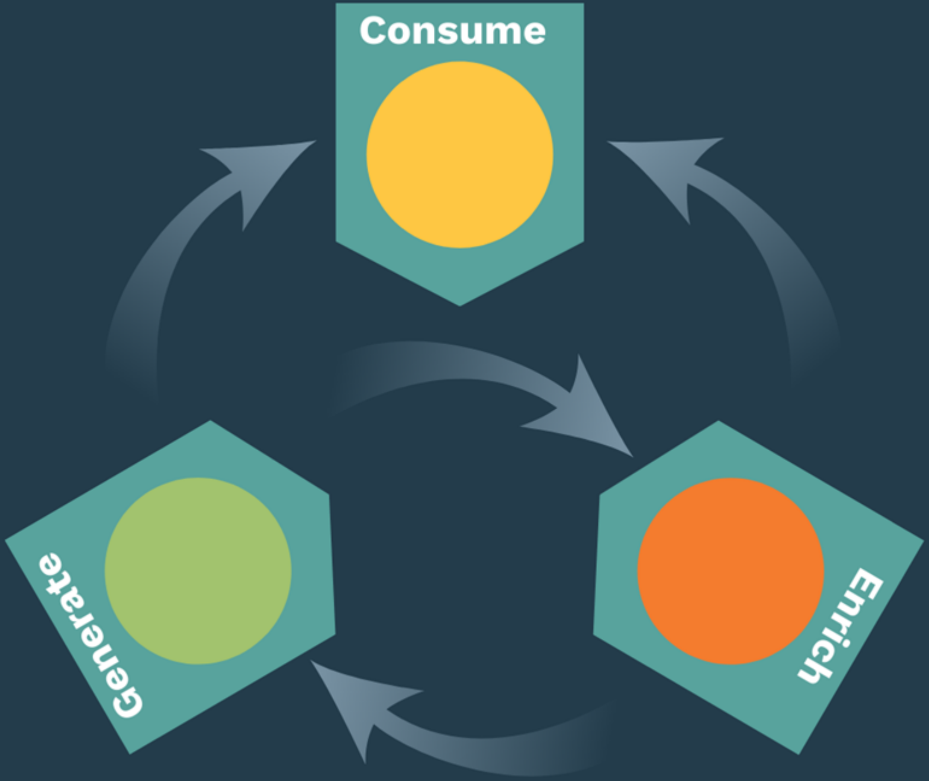
Genomics data cycle

Data that are newly generated by research projects can be consumed by those researchers and also made available for consumption to others.

Genomic data from specimens can also be enriched by intersecting it with other observations, using metadata and processing pipelines.

Enriched data can then be consumed to answer novel questions.

Data enrichment can seed the generation of new data by identifying targets.



Prototype of ARGA Portal

Source code: https://github.com/ARGA-Genomes

- Data from ingested and processed via GBIF pipelines using Darwin Core Archives (DwC-A) metadata standards.
- Working prototype interface built using React.js implementing Solr searching.
- Traits facet filters built to slice results, including: vernacular groups, biomes, data types.

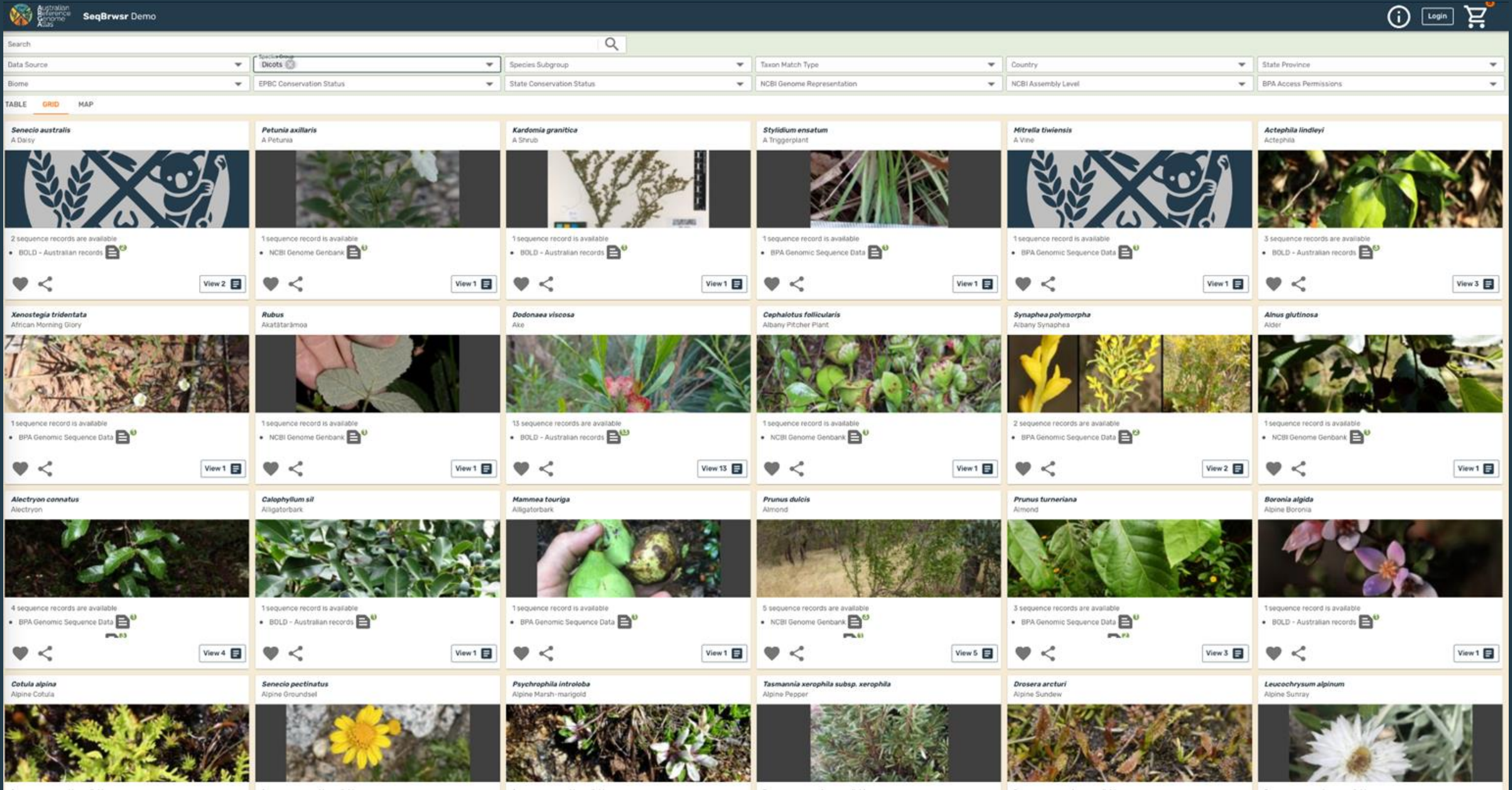
Data sources

Indexed:

- NCBI-GenBank (RefSeq, Genome)
- Bioplatforms Australia
- BOLD systems (Barcode of Life)

Next to be indexed:

- NCBI-GenBank (nucleotides, assembly, SRA)
- The European Nucleotide Archive (ENA)
- Genomes on a Tree (GoaT)
  - ○ Genome Size
  - ○ Plant DNA C-values Database



Senecio australis
A Daisy
2 sequence records are available
- BOLD - Australian records
View 2

Petunia axillaris
A Petunia
1 sequence record is available
- NCBI Genome Genbank
View 1

Kardomia granitica
A Shrub
1 sequence record is available
- BOLD - Australian records
View 1

Stylidium ensatum
A Triggerplant
1 sequence record is available
- BPA Genomic Sequence Data
View 1

Mitrella tiwiensis
A Vine
1 sequence record is available
- BPA Genomic Sequence Data
View 1

Actephila lindleyi
Actephila
3 sequence records are available
- BOLD - Australian records
View 3

Xenostegia tridentata
African Morning Glory
1 sequence record is available
- BPA Genomic Sequence Data
View 1

Rubus
Akatātarāmoa
1 sequence record is available
- NCBI Genome Genbank
View 1

Dodonaea viscosa
Ake
13 sequence records are available
- BOLD - Australian records
View 13

Cephalotus follicularis
Albany Pitcher Plant
1 sequence record is available
- NCBI Genome Genbank
View 1

Synaphea polymorpha
Albany Synaphea
2 sequence records are available
- BPA Genomic Sequence Data
View 2

Alnus glutinosa
Alder
1 sequence record is available
- NCBI Genome Genbank
View 1

Alectryon connatus
Alectryon
4 sequence records are available
- BPA Genomic Sequence Data
View 4

Calophyllum sil
Alligatorbark
1 sequence record is available
- BOLD - Australian records
View 1

Mammea touriga
Alligatorbark
1 sequence record is available
- BPA Genomic Sequence Data
View 1

Prunus dulcis
Almond
5 sequence records are available
- NCBI Genome Genbank
View 5

Prunus turneriana
Almond
3 sequence records are available
- BPA Genomic Sequence Data
View 3

Boronia algida
Alpine Boronia
1 sequence record is available
- BOLD - Australian records
View 1

Cotula alpina
Alpine Cotula

Senecio pectinatus
Alpine Groundsel

Psychrophila introloba
Alpine Marsh-marigold

Tasmannia xerophila subsp. xerophila
Alpine Pepper

Drosera arcturi
Alpine Sundew

Leucochrysum alpinum
Alpine Sunray

Data exploration site: **https://nectar-arga-dev-1.ala.org.au**

A FAIRer future for genomics research

Findable

- data from multiple reputed repositories indexed using recognised metadata standards (Darwin Core and Darwin Core extensions)
- enriched by intersecting with specimen metadata from the Atlas of Living Australia to increase trust and quality assessment
- PIDs, citations and original sources linked

Accessible and Interoperable

- multiple data formats, user-selectable
- new and flexible search strategies by taxon, phenotype and genomic annotations

Reusable

- DOIs of searches
- integration with BioCommons platforms

ARGA Development Team

Nick dos Remedios	Atlas of Living Australia	Lead Systems Engineer
Christopher Mangion	Australian BioCommons	Data Engineer
Matt Andrews	Atlas of Living Australia	Systems Support
Mok	Australian BioCommons	UX/UI Designer
Keeva Connolly	Australian BioCommons	Scientific Business Analyst
Kathryn Hall	Atlas of Living Australia	Project Manager