

# Curating species lists: Aggregating data to enhance context

eResearch Conference, 18 October 2023

Keeva Connolly, Kathryn Hall



**ARGA**  
Australian Reference Genome Atlas

# 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



Australian  
**BioCommons**



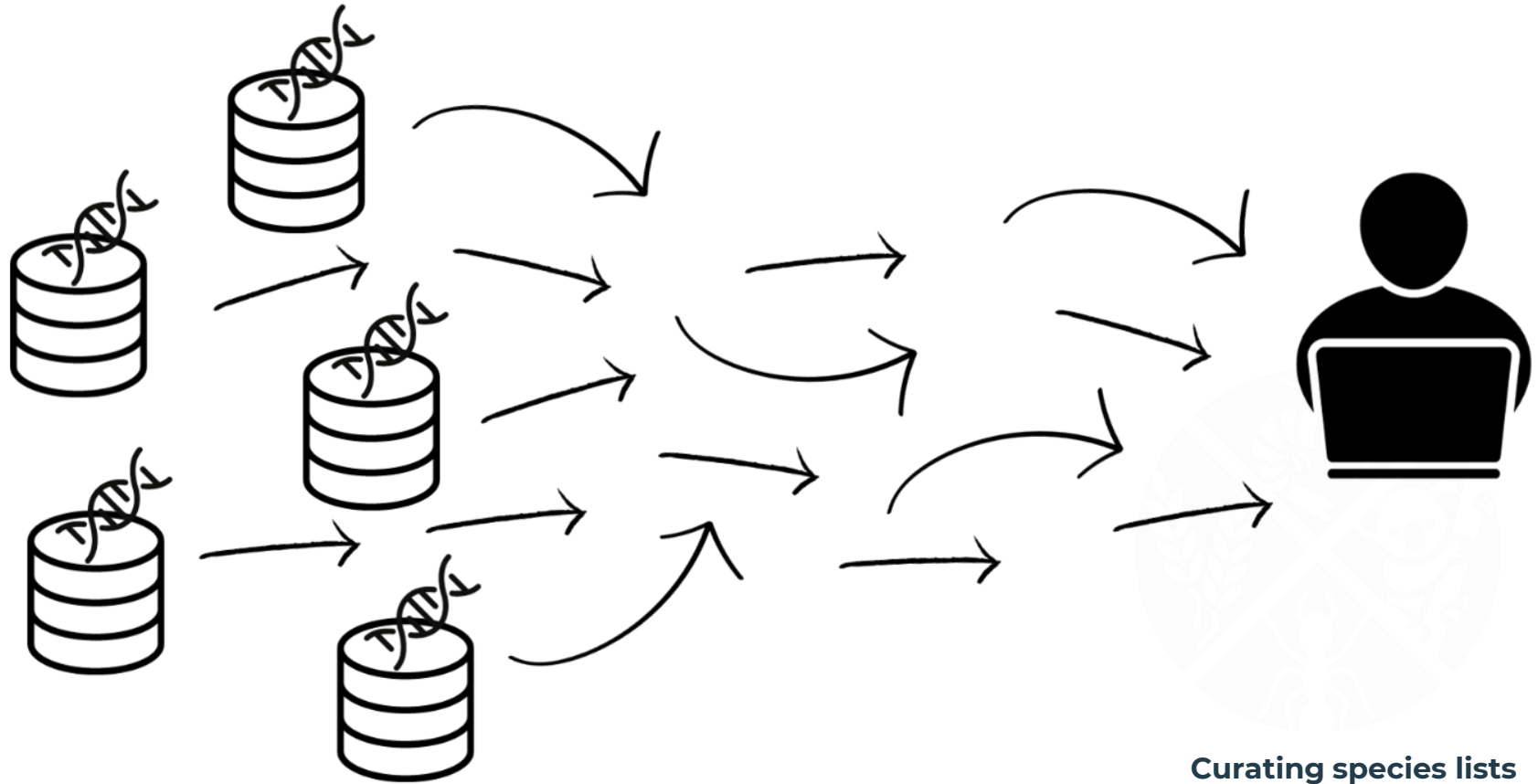
**BIOPLATFORMS**  
**AUSTRALIA**



Australian Research Data Commons

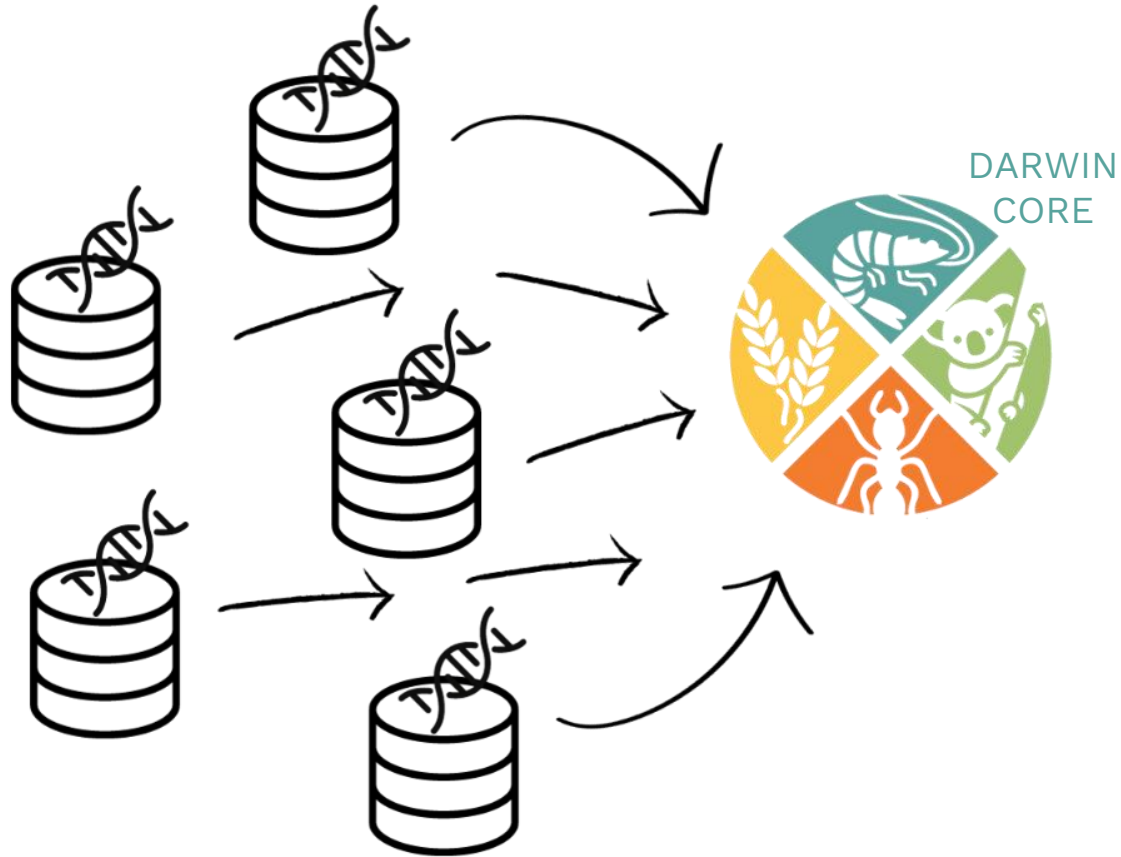


# ARGA is an indexing platform



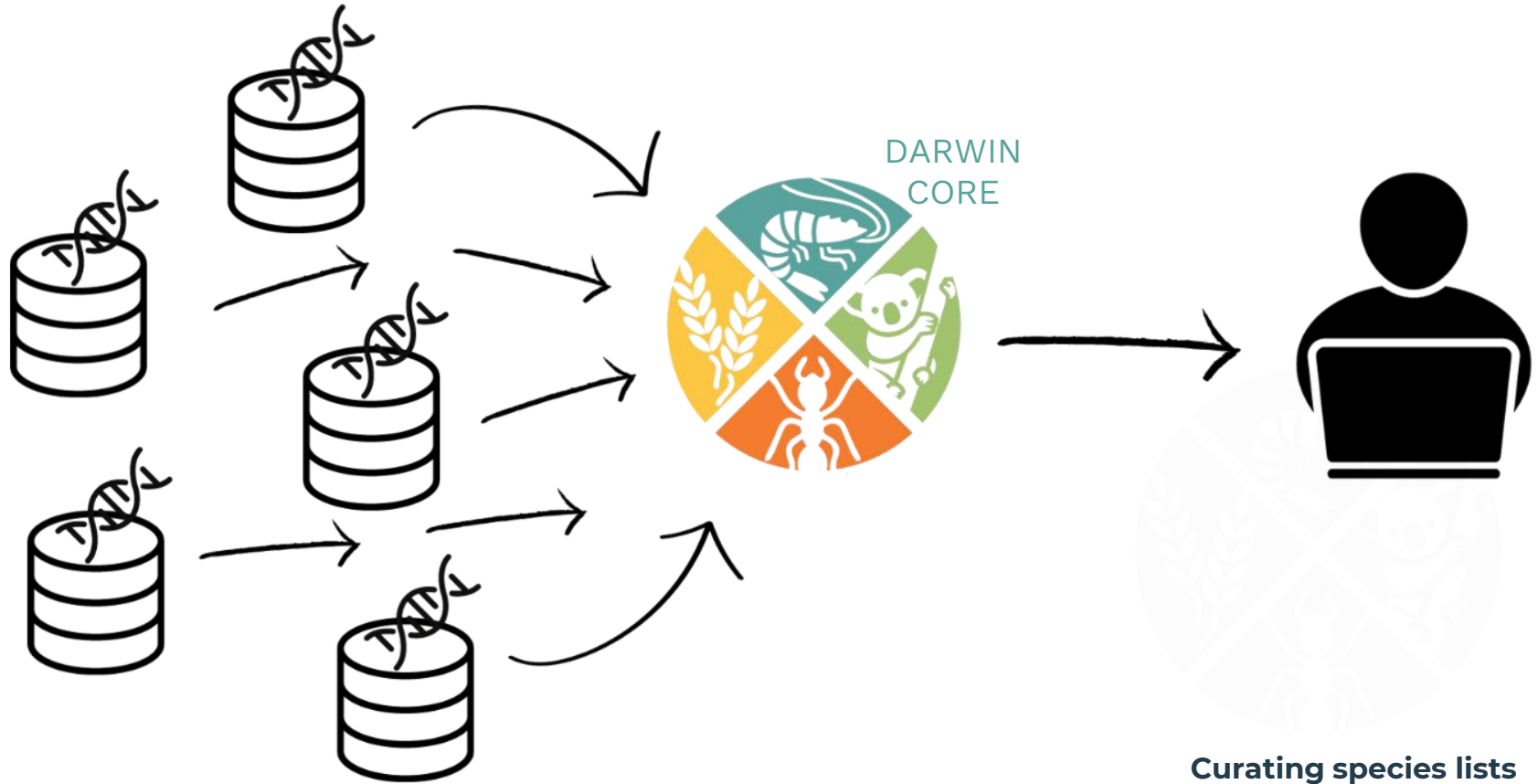
Curating species lists

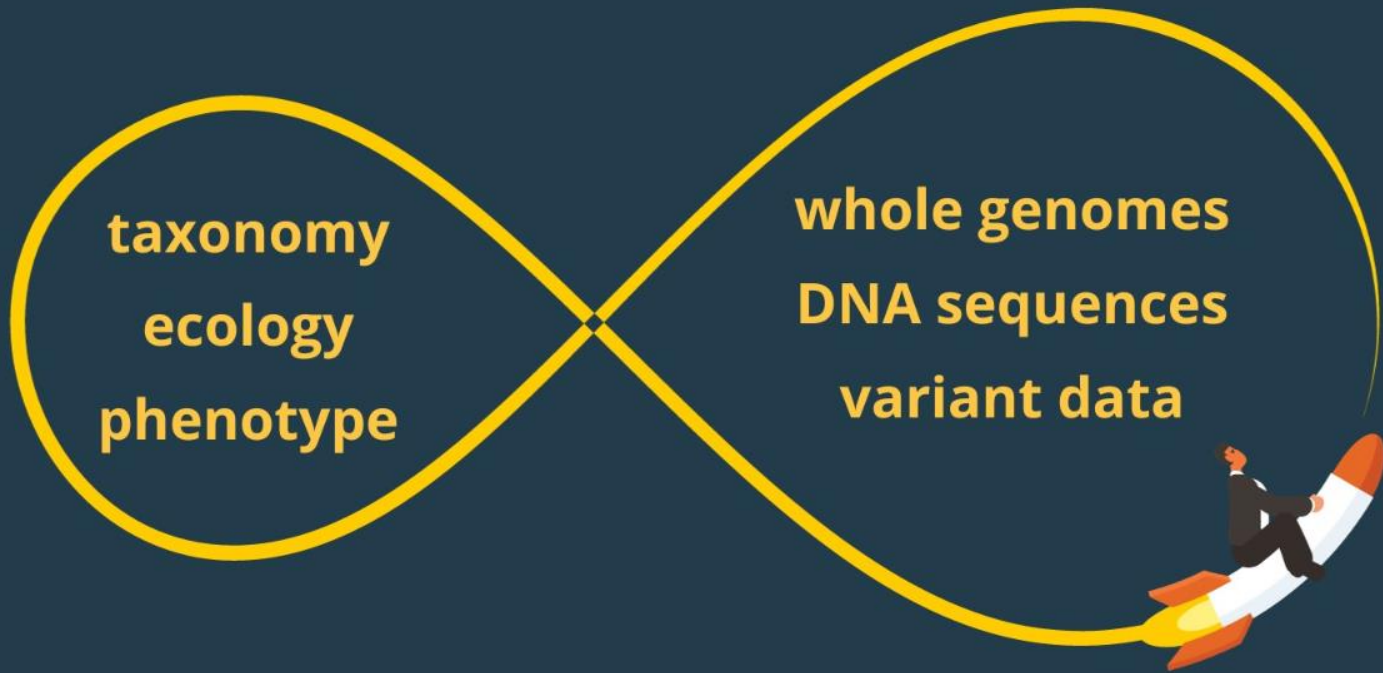
# ARGA is an indexing platform



Curating species lists

# ARGA is an indexing platform

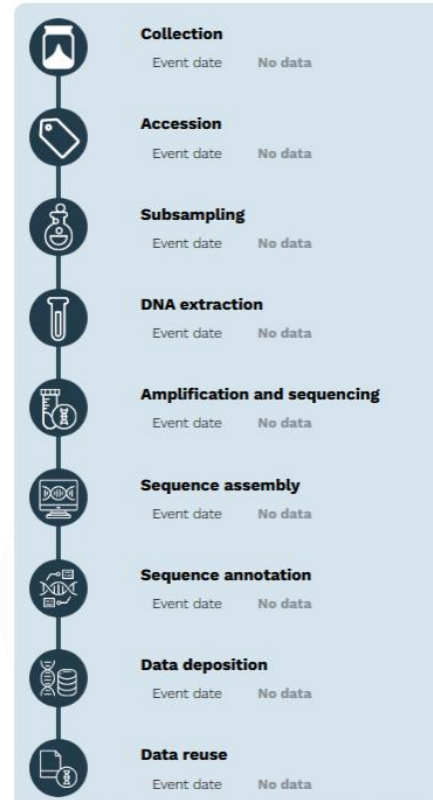




**ARGA Project is building an indexing service for discovering, filtering and accessing complex life science data within biological contexts.**

# Context at the specimen and sequence level

- Sequence data deposited without descriptive metadata impedes trust and re-use
- ARGAs aim to expose context to help users make informed quality judgements
- Aggregates and clarifies metadata at each step of the event chain



**Curating species lists**

# Context at the species level

- Sequence data are rarely deposited with related ecological data
- Collating species lists from trusted sources to connect sequence data to relevant, species-level ecological, biological and legal data
- Functions of species lists:
  - Provide information from species profile pages
  - Enable searching, filtering and browsing according to traits
  - Summarise genomic data availability for groups of species



# Species list themes

- Widely applicable
- Relevant to genomic data
- Australian-centric



**Curating species lists**

# Species list themes

- Widely applicable
- Relevant to genomic data
- Australian-centric



Threatened species



**Curating species lists**

# Species list themes

- Widely applicable
- Relevant to genomic data
- Australian-centric



Threatened species



CITES-listed species

**Curating species lists**

# Species list themes

- Widely applicable
- Relevant to genomic data
- Australian-centric



Threatened species



Invasives and pests



CITES-listed species

**Curating species lists**

# Species list themes

- Widely applicable
- Relevant to genomic data
- Australian-centric



Threatened species



Crop wild relatives



Invasives and pests



CITES-listed species

**Curating species lists**

# Species list themes

- Widely applicable
- Relevant to genomic data
- Australian-centric



Threatened species



Crop wild relatives



Invasives and pests



Migratory species



CITES-listed species

**Curating species lists**

# Species list themes

- Widely applicable
- Relevant to genomic data
- Australian-centric



Migratory species



Invasives and pests



Threatened species



Crop wild relatives



Species vulnerable to fire



CITES-listed species

**Curating species lists**

# Species list themes

- Widely applicable
- Relevant to genomic data
- Australian-centric



Threatened species



Crop wild relatives



Species vulnerable to fire



Invasives and pests



Migratory species



Venomous and  
poisonous species



CITES-listed species

**Curating species lists**

# Sourcing lists

- Many lists and data sources are available online, but they vary widely in scope, in how well evidenced they are, and in currency
- We prioritised lists published by:
  - Federal and state/territory governments
  - Multilateral organisations (e.g. UN Environment Programme)
  - Established databases and infrastructures (e.g. US National Plant Germplasm System - GRIN-Global)
  - Australian institutions and scientists (e.g. Kamilaroi, Noongar Boodjar and South East Arnhem Land encyclopaedias of plants and animals published via ALA)

# Collating lists - standardising to one schema

- Record data fields
  - verbatimName, vernacularName
- Source metadata fields
  - datasetName, datasetID, datasetCitation, dataCurrency
- ARGGA record metadata fields
  - dateCreated, dateModified



# Collating lists - standardising to one schema

- Icons
  - E.g. EPBC Act Category: Endangered, EPBC Act Category: Vulnerable
- External IDs
  - E.g. GRIN-Global nomen number, CAAB (Codes for Australian Aquatic Biota) code
- Filters
  - E.g. jurisdiction, trait type





# After ingestion

- Follow a refresh schedule to check whether composite lists have been updated and update ARGAs lists accordingly
- Gather user feedback on utility and interactive-ness
- Functions of species lists:
  - Provide information from species profile pages
  - Enable searching, filtering and browsing according to traits
  - Summarise genomic data availability for groups of species





Data Distribution

Taxonomy

Whole Genomes

Markers

Other Genetic Data

Specimen

Indexed data distribution



[Hide](#)

**Indexed data**

Whole genomes 0/1

Loci 0/0

Other data 0/0

Specimens 1/94



**Distribution**

Arnhem Coast, Arnhem Plateau, Brigalow Belt North, Brigalow Belt South, Cape York Peninsula, Central Arnhem, Central Kimberley, Central Mackay Coast, Daly Basin, Dampierland, Darwin Coastal, Desert Uplands, Einasleigh Uplands, Gulf Coastal, Gulf Fall and Uplands, Gulf Plains, NSW North Coast, Nandewar, New England Tablelands, Northern Kimberley, Ord Victoria Plain, Pine Creek, South Eastern Queensland, Tiwi Cobourg, Victoria Bonaparte, Wet Tropics

Species ***Platyplectrum ornatum***



IEK:  
SEAL

Reference Genome

[Data Distribution](#)[Taxonomy](#)[Whole Genomes](#)[Markers](#)[Other Genetic Data](#)[Specimen](#)

Indigenous Ecological Knowledge: South East Arnhem Land

Name: Senfrog

Food use

Medicinal use

Cultural connection

### Indexed data distribution



[Hide](#)

### Indexed data

Whole genomes 0/1

Loci 0/0

Other data 0/0

Specimens 1/94

### Distribution

Arnhem Coast, Arnhem Plateau, Brigalow Belt North, Brigalow Belt South, Cape York Peninsula, Central Arnhem, Central Kimberley, Central Mackay Coast, Daly Basin, Dampierland, Darwin Coastal, Desert Uplands, Einasleigh Uplands, Gulf Coastal, Gulf Fall and Uplands, Gulf Plains, NSW North Coast, Nandewar, New England Tablelands, Northern Kimberley, Ord Victoria Plain, Pine Creek, South Eastern Queensland, Tiwi Cobourg, Victoria Bonaparte, Wet Tropics

## Search for data

Q e.g. sequence accession, taxon identifier, genus name

## Browse by data type



**Whole genomes**  
8.23k records



**Markers**  
230.68k records



**Specimen**  
0 records

## Browse by taxonomic group



**Animals**  
9 records



**Plants**  
0 records



**Fungi**  
0 records



**Bacteria**  
0 records



**Protista**  
0 records

## Browse by functional or ecological group



**Agriculture**  
0 records



**Aquaculture**  
0 records



**Terrestrial**  
0 records



**Marine**  
0 records



**Threatened**  
0 records

**view  
all**

**All species**  
84.58k records



ORDER **Anura**

### Taxonomy

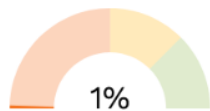
Scientific name	No data
Status	No data
Source	No data

### Higher classification

Kingdom	Phylum	Class	Order	Family	Genus
<a href="#">Animalia</a>	<a href="#">Chordata</a>	<a href="#">Amphibia</a>	<a href="#">Anura</a>	<a href="#">Bufonidae</a>	<a href="#">Rhinella</a>

### Data summary

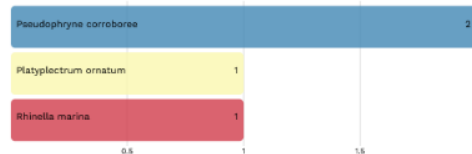
Percentage of species with genomes



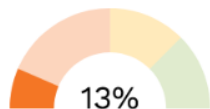
Families with genomes



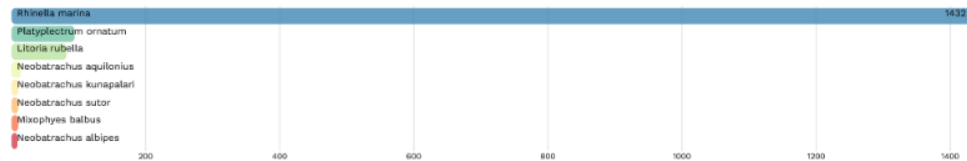
Species with genomes



Percentage of species with any genetic data



Species with any genetic data



### Taxonomic breakdown

Number of families	<b>6</b>
Number of species/OTUs	<b>245</b>
Families with genomes	<b>3</b>
Species with genomes	<b>3</b>
Families with data	<b>5</b>
Species with data	<b>32</b>

Species with most genomes

[Pseudophryne corroboree](#)

Species with most data

[Rhinella marina](#)

# ARGA is launching in November

ARGA is launching on Friday the 3rd of November - you can find the link to register for our launch webinar here:

<https://www.biocommons.org.au/events/arga-launch>

Email:

[keeva.connolly@qcif.edu.au](mailto:keeva.connolly@qcif.edu.au)



**Curating species lists**

## ARGA Development Team

Goran Sterjov

Atlas of Living Australia

Software Engineer

Christopher Mangion

Australian BioCommons

Data Engineer

Vikas Nagaraju

Atlas of Living Australia

Software 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



ARGA  
Australian Reference Genome Atlas