The importance of being indexed:

introducing Genome Tracker by the Australian Reference Genome Atlas

Kathryn Hall, Matt Andrews, Peter Brenton, Jack Brinkman, Keeva Connolly,
Christopher Mangion, Winnie Mok, Sarah Richmond, Goran Sterjov, Andrew Turley & Nigel Ward

Atlas of Living Australia | *Australian BioCommons* | *Bioplatforms Australia*

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

taxonomy
ecology
phenotype

genomic data

The Australian Reference Genome Atlas (ARGA) is an indexing service for discovering, filtering and accessing complex life science data within biological contexts.

The ARGA application

ARGA's data can support research in:

Conservation

ARGA contains 5,382 species listed by international and domestic agencies

Biosecurity

ARGA contains 2,823 notifiable, pest and invasive species from authoritative watchlists

Biodiscovery

ARGA contains 1,244 crop wild relatives and 876 species known to be venomous or poisonous

Industry

ARGA contains 717 species utilised for agriculture, aquaculture and forestry

Biodiversity

ARGA contains 43,937 plant and animal species known to be endemic to Australia

Data types indexed by ARGA

Genomic	Genome assemblies Genetic locus sequences Sequence read files
Taxonomic	Classification systems Taxonomic histories Taxonomic synonyms
Specimen	Voucher information Collection location Institutional provenance
Species-level traits	Ecological groupings Legal and regulatory status Commercial applications

Data sources processed

9	genomic data resources
5	taxonomic data resources
2	specimen data resources
126	trait data resources
142	**total data resources**

Data sources ingested

3	genomic data resources
5	taxonomic data resources
2	specimen data resources
87	trait data resources
98	**total data resources**

ARGA is about data in context to solve problems

Conservation Biology

Comparing 2 genomes of the corroboree frog, conservation biologists can look for disease resistance or heat stress genes to help make populations resilient to fungus or climate stress.

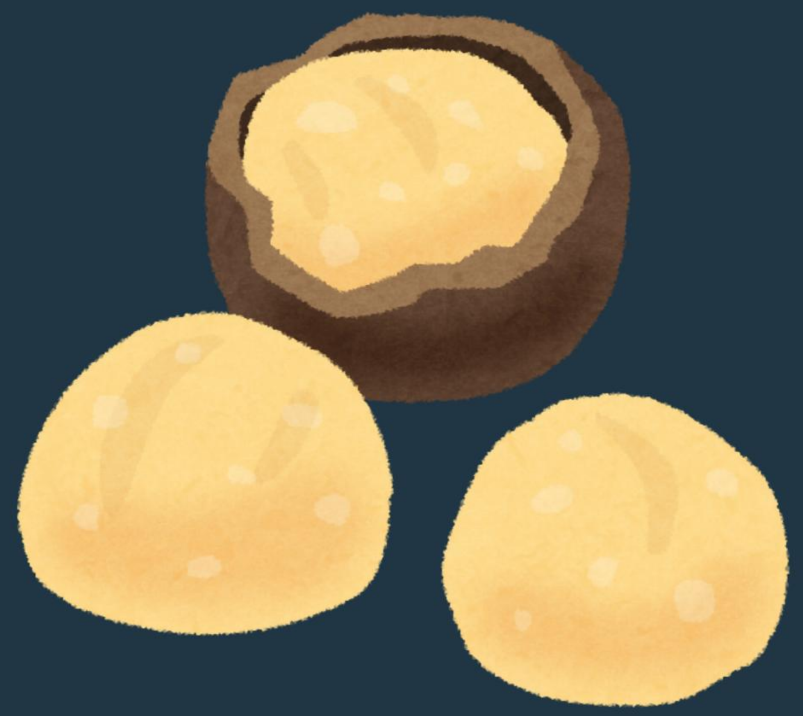
Australian Reference Genome Atlas. 2025. Pseudophryne corroboree Moore, 1953, Whole genomes, The Australian Reference Genome Atlas. Accessed at: https://app.arga.org.au/species/Pseudophryne_corroboree/whole_genomes on 23-9-2025.

Biosecurity Services

Biosecurity researchers can download any of 19 genome assemblies and 777 single gene sequences to build a probe to detect invasive carp using eDNA in Australian river systems.

Australian Reference Genome Atlas. 2025. Cyprinus carpio Linnaeus, 1758, Summary, The Australian Reference Genome Atlas. Accessed at: https://app.arga.org.au/species/Cyprinus_carpio/summary on 23-9-2025.

Agricultural Science



Commerically valuable macadamia nuts are endemic to Australia. Scientists can look at the genomes from all 4 described species to enhance pest resistance and climate tolerance, including from the critically endangered Bulburin Nut Tree.

Australian Reference Genome Atlas. 2025. Macadamia F.Muell., Macadamia, The Australian Reference Genome Atlas. Accessed at: https://app.arga.org.au/genus/Macadamia on 29-9-2025.

Evolution and Ecology

Researchers can look at gene expression in tissues to look for adaptations to Australia's unique environments, such as identifying taste receptors in endangered numbats that allowed them to exploit termites as a food-source.

Australian Reference Genome Atlas. 2025. Myrmecobius fasciatus Waterhouse, 1836, Summary, The Australian Reference Genome Atlas. Accessed at: https://app.arga.org.au/species/Myrmecobius%20fasciatus/summary on 29-9-2025.

Traits contexts

LIST GROUP

Australia's 110 Priority Threatened Species

ARGA Team

© Australian Reference Genome Atlas (ARGA) Project for the Atlas of Living Australia and Bioplatforms Australia

View full list on ALA Lists

Dataset highlights

Latest genome release: *Unknown*

Most genomes: *Phascolarctos cinereus*

Most loci: *Carcharias taurus*

Most data: *Carcharias taurus*

Data summary

Taxonomic diversity of this dataset

This graph shows the aggregated total of species for which a whole genome has been sequenced and made available. The first instance of a whole genome sequence for an individual species has been plotted as an accumulated total. Statistics based on records indexed within ARGA.

Taxonomic breakdown

Species with genomes	16
Species with loci	34
Species with data	37

Percentage of species with genomes

Species with genomes

- Phascolarctos cinereus
- Chelonia mydas
- Dasyurus hallucatus
- Macrotis lagotis
- Dryococelus australis
- Lathamus discolor
- Myrmecobius fasciatus
- Setonix brachyurus

Percentage of species with loci

Species with loci

- Carcharias taurus
- Chelonia mydas
- Hippocampus whitei
- Astacopsis gouldi
- Pristis pristis
- Neophema (Neonanodes)
- Phascolarctos cinereus
- Myrmecobius fasciatus

Percentage of species with any genetic data

Species with any genetic data

- Carcharias taurus
- Chelonia mydas
- Hippocampus whitei
- Astacopsis gouldi
- Pristis pristis
- Phascolarctos cinereus
- Dasyurus hallucatus
- Macrotis lagotis

Note:

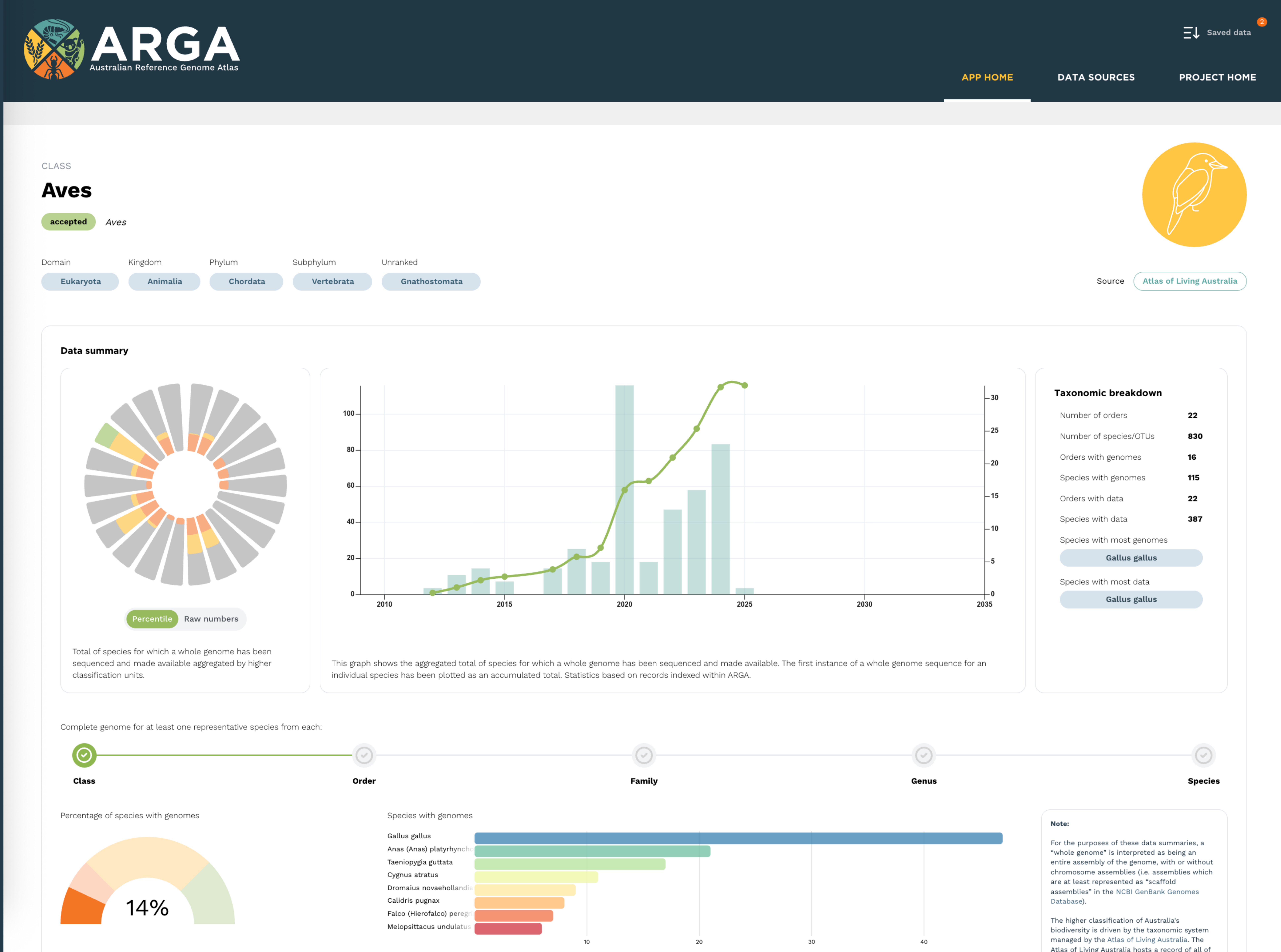
For [illegible] of these data summaries, a "whole [illegible]" is interpreted as being an entire [illegible] of the genome, with or without chromosome assemblies (i.e. assemblies which are at least represented as "scaffold assemblies" in the NCBI GenBank Genomes Database).

The higher classification of Australia's biodiversity is driven by the taxonomic system managed by the [illegible] Australia. The Atlas of Living Australia hosts a record of all of the species that appear on the Australian National Species List, and services nationally agreed nomenclature for these species.

The data used to generate the page statistics and graphics are accurate to 28 September 2025. Data and graphics on this page may be shared under a CC BY 4.0 licence.

- header includes
  - source
  - data sharing and licence conditions
  - counts of species
- data highlights block, showing
  - latest genome release
  - species with most genomes
  - species with most genetic locus sequences
  - species with most data overall
- charts of
  - taxonomic breakdown of species
  - accumulation plot of complete genomes
  - tachometers
  - leaderboards

Taxonomic contexts



Aves

accepted *Aves*

Domain	Kingdom	Phylum	Subphylum	Unranked
Eukaryota	Animalia	Chordata	Vertebrata	Gnathostomata

Source Atlas of Living Australia

Data summary

Total of species for which a whole genome has been sequenced and made available aggregated by higher classification units.

This graph shows the aggregated total of species for which a whole genome has been sequenced and made available. The first instance of a whole genome sequence for an individual species has been plotted as an accumulated total. Statistics based on records indexed within ARGA.

Taxonomic breakdown

Number of orders	22
Number of species/OTUs	830
Orders with genomes	16
Species with genomes	115
Orders with data	22
Species with data	387
Species with most genomes	Gallus gallus
Species with most data	Gallus gallus

Complete genome for at least one representative species from each:

Percentage of species with genomes

Species with genomes

Note:

For the purposes of these data summaries, a "whole genome" is interpreted as being an entire assembly of the genome, with or without chromosome assemblies (i.e. assemblies which are at least represented as "scaffold assemblies" in the NCBI GenBank Genomes Database).

The higher classification of Australia's biodiversity is driven by the taxonomic system managed by the Atlas of Living Australia. The Atlas of Living Australia hosts a record of all of

Do we have genomes for every order?

Data deep dive

- we need to establish:
  - which taxa **have genomes**
  - which do **not**
  - which taxa are within which rank

- we need to know:
  - baseline of species
  - classification of species: how taxa nest within higher ranks
  - **genomes** for each species

Building a data ROV

- navigate via complex Diesel and SQL queries
- use taxonomy DAG to:
  - count species with at least 1 genome
  - calculate genera with at least 1...
  - iterate up tree

Code:

```
SELECT taxa.rank,
       COUNT(*),
       SUM(CASE WHEN (taxa_tree_stats.total_full_genomes_coverage > 0) THEN 1 ELSE 0
       END),
       (sum(total_full_genomes_coverage::float4 / case when children = 0 then 1 else
       children end) / count(*))::float4
FROM (taxa_tree_stats INNER JOIN taxa ON (taxa_tree_stats.id = taxa.id))
WHERE ((taxa_tree_stats.taxon_id = $2) AND (taxa.rank = ANY ($3)))
GROUP BY taxa.rank -- binds: [936f045e-74d1-4cde-a216-306682b4a098, [Domain,
Kingdom, Phylum]]
```

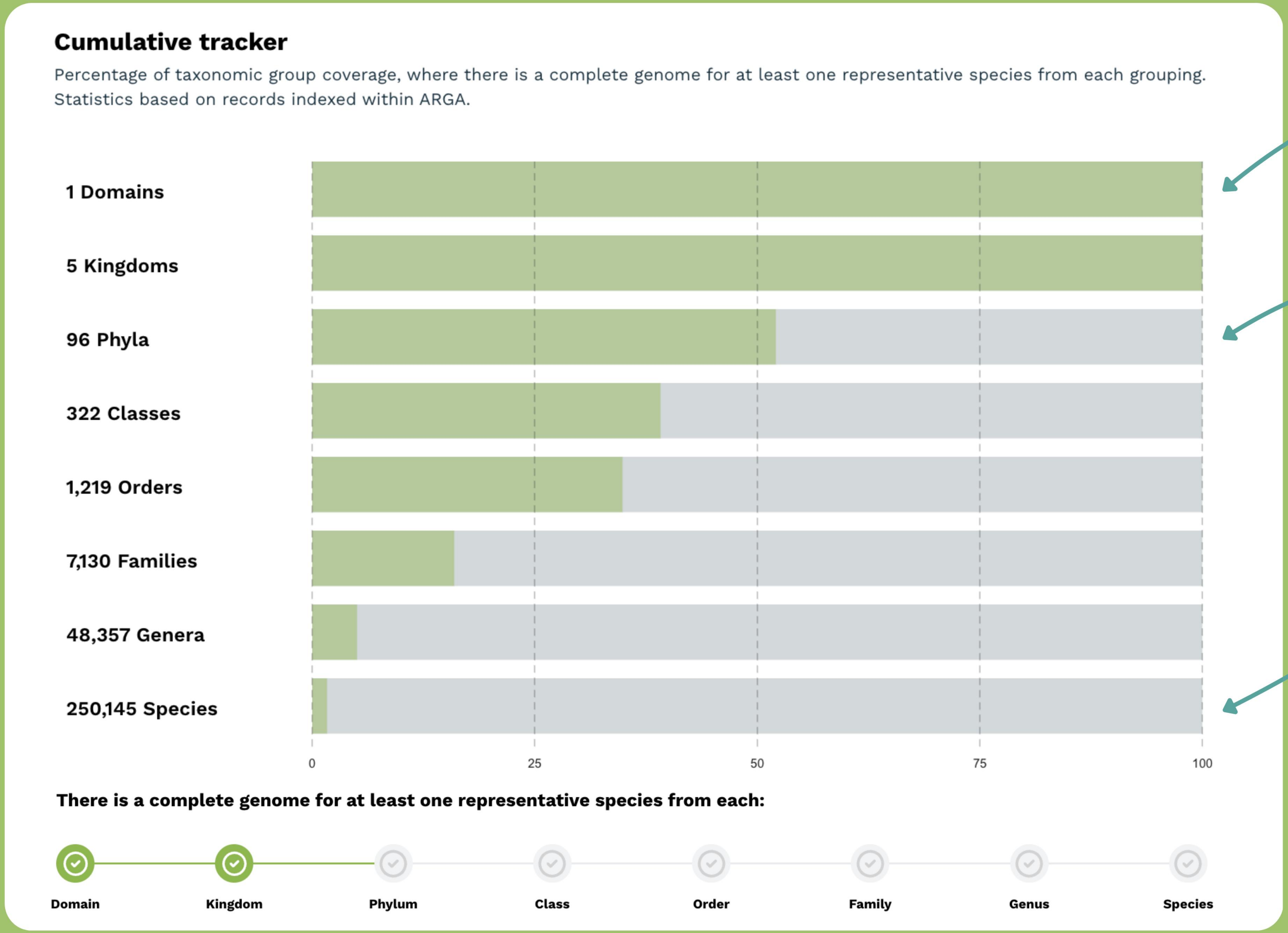
Repositories:

https://github.com/ARGA-Genomes/arga-backend/blob/main/server/src/database/stats.rs#L153
https://github.com/ARGA-Genomes/arga-backend/blob/main/core/schema.sql#L1554

Genome Tracker in action

Genome Tracker

progress, transparency, and Australia’s contribution to global genomic discovery



Cumulative tracker

Percentage of taxonomic group coverage, where there is a complete genome for at least one representative species from each grouping. Statistics based on records indexed within ARGA.

There is a complete genome for at least one representative species from each:

- **all kingdoms** in Australia’s biodiversity have at least **1 complete genome**
- **> 50%** of **phyla**
- **~ 2%** of **species**

Taxonomy as time travel

- splits between ranks, like kingdoms and phyla show ancient divergence events:
    - eukaryotes arose between 1.1 and 2.3 Ga
    - mitochondrial symbiosis occurred ~1.8 Ga
    - animals and plants split ~1,524 ± 53 Ma
    - fungi and animals split ~1,492 ± 46 Ma
    - vertebrates emerged ~964 ± 132 Ma

References:

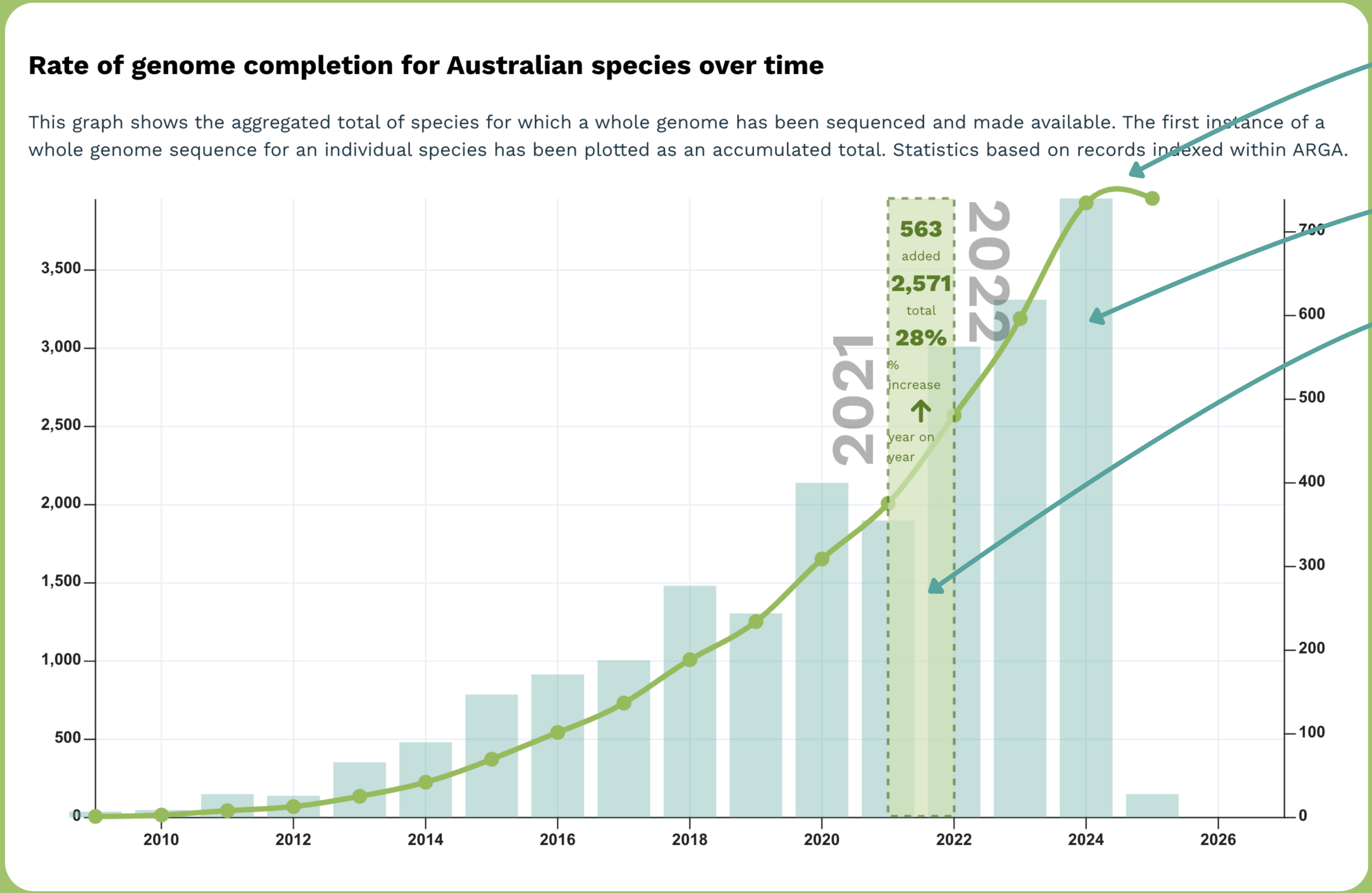
Chernikova, D., Motamedi, S., Csürös, M. et al. A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. Biol Direct 6, 26 (2011). https://doi.org/10.1186/1745-6150-6-26.

Hedges SB, Chen H, Kumar S, Wang DY, Thompson AS, Watanabe H. A genomic timescale for the origin of eukaryotes. BMC Evol Biol. 1:4 (2001). https://doi.org/10.1186/1471-2148-1-4.

Hedges SB, Blair JE, Venturi ML, Shoe JL. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol. 4:2 (2004). https://doi.org/10.1186/1471-2148-4-2.

Genome Tracker

progress, transparency, and Australia’s contribution to global genomic discovery



Rate of genome completion for Australian species over time

This graph shows the aggregated total of species for which a whole genome has been sequenced and made available. The first instance of a whole genome sequence for an individual species has been plotted as an accumulated total. Statistics based on records indexed within ARGA.

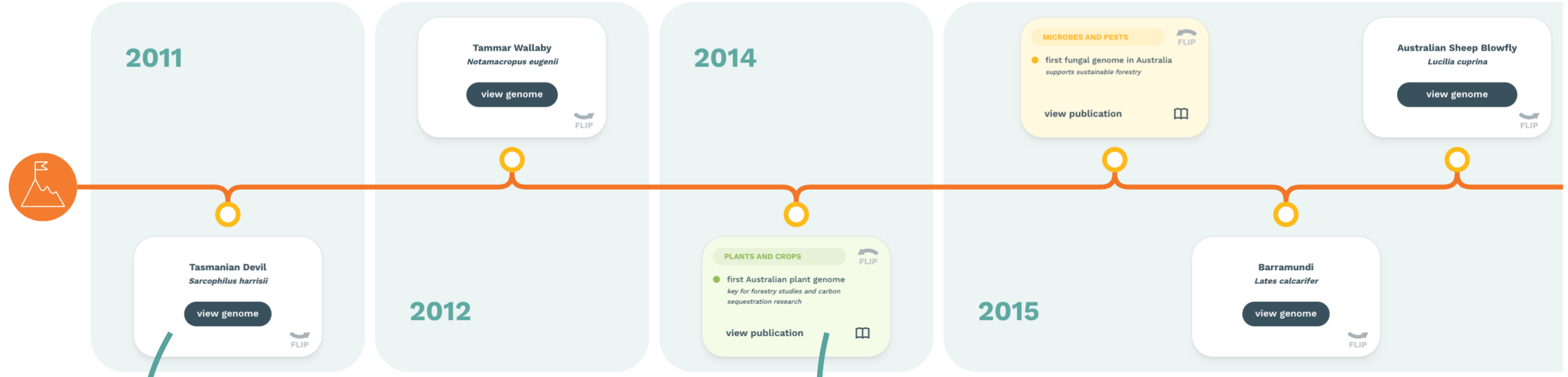
- plot **accumulation curve** of running aggregate
- show **total per year**
- calculate **difference in rate** of genomic sequencing

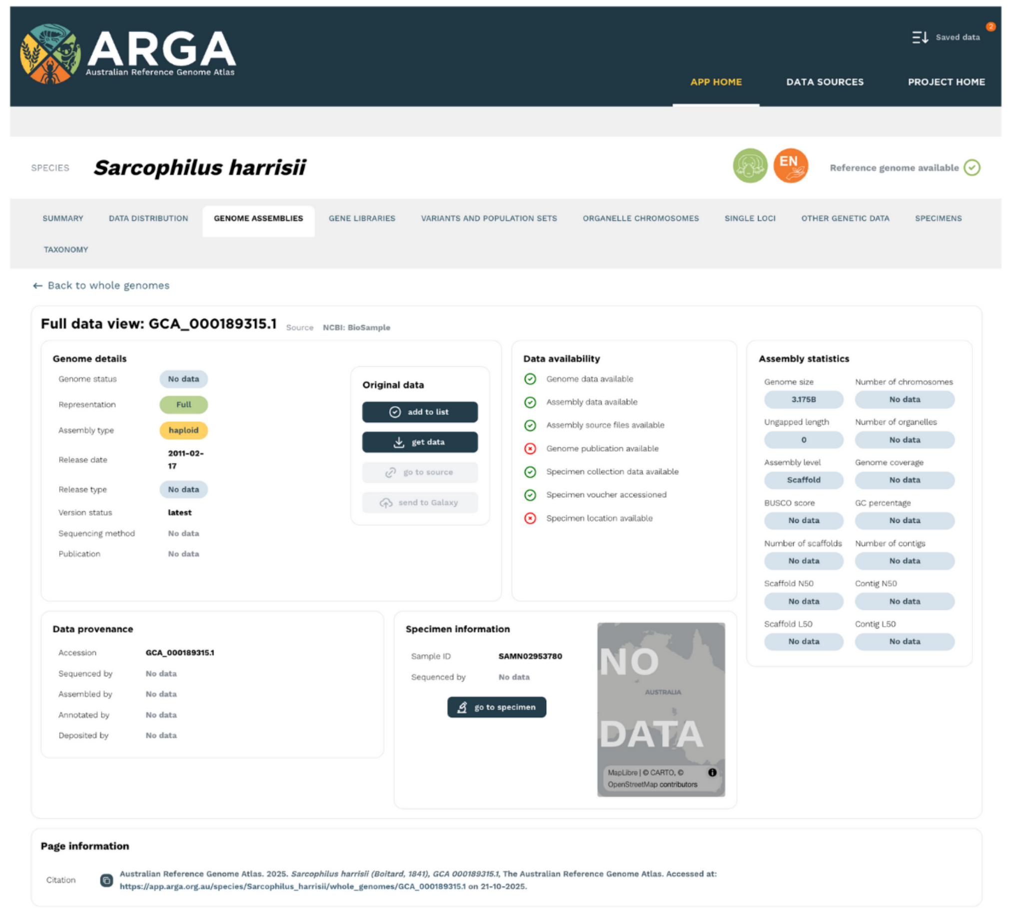
impact of funding

Genome Tracker

progress, transparency, and Australia's contribution to global genomic discovery

Significant milestones in the genome sequencing of Australia's biodiversity





SPECIES *Sarcophilus harrisii*

Reference genome available ✓

Full data view: GCA_000189315.1 Source NCBI: BioSample

Genome details

Genome status: No data
Representation: Full
Assembly type: haploid
Release date: 2011-02-17
Release type: No data
Version status: latest
Sequencing method: No data
Publication: No data

Original data

Data availability

✓ Genome data available
✓ Assembly data available
✓ Assembly source files available
✗ Genome publication available
✓ Specimen collection data available
✓ Specimen voucher accessioned
✗ Specimen location available

Assembly statistics

Genome size 3.175B	Number of chromosomes No data
Ungapped length 0	Number of organelles No data
Assembly level Scaffold	Genome coverage No data
BUSCO score No data	GC percentage No data
Number of scaffolds No data	Number of contigs No data
Scaffold N50 No data	Contig N50 No data
Scaffold L50 No data	Contig L50 No data

Data provenance

Accession: GCA_000189315.1
Sequenced by: No data
Assembled by: No data
Annotated by: No data
Deposited by: No data

Specimen information

Sample ID: SAMN02953780
Sequenced by: No data

Page information

Citation: Australian Reference Genome Atlas. 2025. *Sarcophilus harrisii (Boitard, 1841), GCA 000189315.1*, The Australian Reference Genome Atlas. Accessed at: https://app.arga.org.au/species/Sarcophilus_harrisii/whole_genomes/GCA_000189315.1 on 21-10-2025.

Article | Open access | Published: 11 June 2014

The genome of *Eucalyptus grandis*

Alexander A. Myburg, Dario Grattapaglia, Gerald A. Tuskan, Uffe Hellsten, Richard D. Hayes, Jane Grimwood, Jerry Jenkins, Erika Lindquist, Hope Tice, Diane Bauer, David M. Goodstein, Inna Dubchak, Alexandre Poliakov, Eshchar Mizrachi, Anand R. K. Kullan, Steven G. Hussey, Desre Pinard, Karen van der Merwe, Pooja Singh, Ida van Jaarsveld, Orzenil B. Silva-Junior, Roberto C. Togawa, Marilia R. Pappas, Danielle A. Faria, … Jeremy Schmutz + Show authors

Nature **510**, 356–362 (2014) | Cite this article

80k Accesses | **779** Citations | **342** Altmetric | Metrics

Abstract

Eucalypts are the world's most widely planted hardwood trees. Their outstanding diversity, adaptability and growth have made them a global renewable resource of fibre and energy. We sequenced and assembled >94% of the 640-megabase genome of *Eucalyptus grandis*. Of 36,376 predicted protein-coding genes, 34% occur in tandem duplications, the largest proportion thus far in plant genomes. *Eucalyptus* also shows the highest diversity of genes for specialized metabolites such as terpenes that act as chemical defence and provide unique pharmaceutical oils. Genome sequencing of the *E. grandis* sister species *E. globulus* and a set of inbred *E. grandis* tree genomes reveals dynamic genome evolution and hotspots of inbreeding depression. The *E. grandis* genome is the first reference for the eudicot order Myrtales and is placed here sister to the eurosids. This resource expands our understanding of the unique biology of large woody perennials and provides a powerful tool to accelerate comparative biology, breeding and biotechnology.

Editorial Summary

***Eucalyptus* genome sequenced**

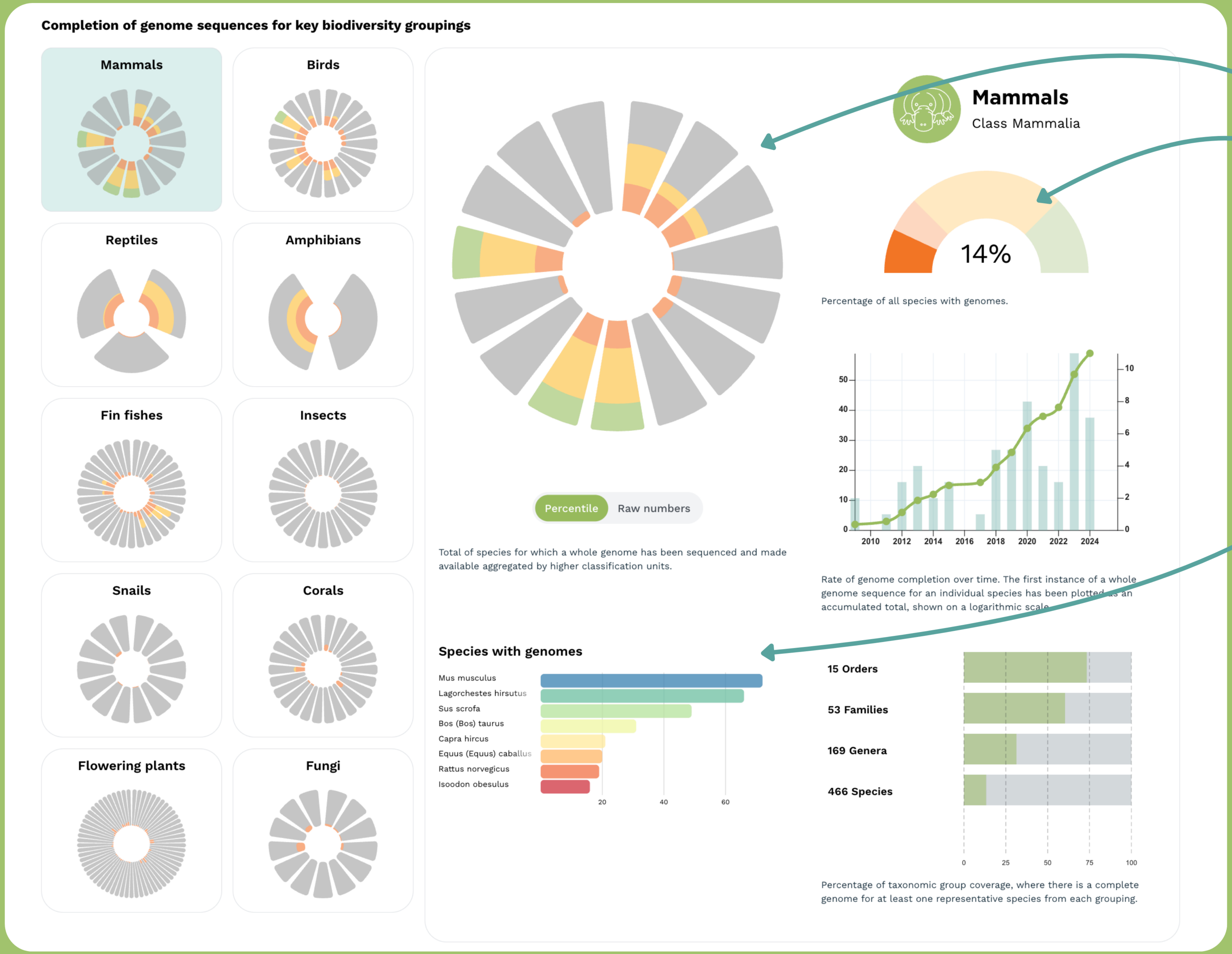
Fast-growing *Eucalyptus* trees form the basis of an international pulp, paper and chemical cellulose industry and they are also seen as potential biomass feedstocks for bioenergy and biomaterials. The genome of *Eucalyptus grandis* show all

- interactive timeline feature showing **key milestones**
- learn and explore
- view **genome**
- navigate to **source** publication

recognition and engagement

Genome Tracker

progress, transparency, and Australia's contribution to global genomic discovery



Completion of genome sequences for key biodiversity groupings

Percentage of all species with genomes.

Total of species for which a whole genome has been sequenced and made available aggregated by higher classification units.

Rate of genome completion over time. The first instance of a whole genome sequence for an individual species has been plotted as an accumulated total, shown on a logarithmic scale.

Species with genomes

- Mus musculus
- Lagorchestes hirsutus
- Sus scrofa
- Bos (Bos) taurus
- Capra hircus
- Equus (Equus) caballus
- Rattus norvegicus
- Isoodon obesulus

15 Orders

53 Families

169 Genera

466 Species

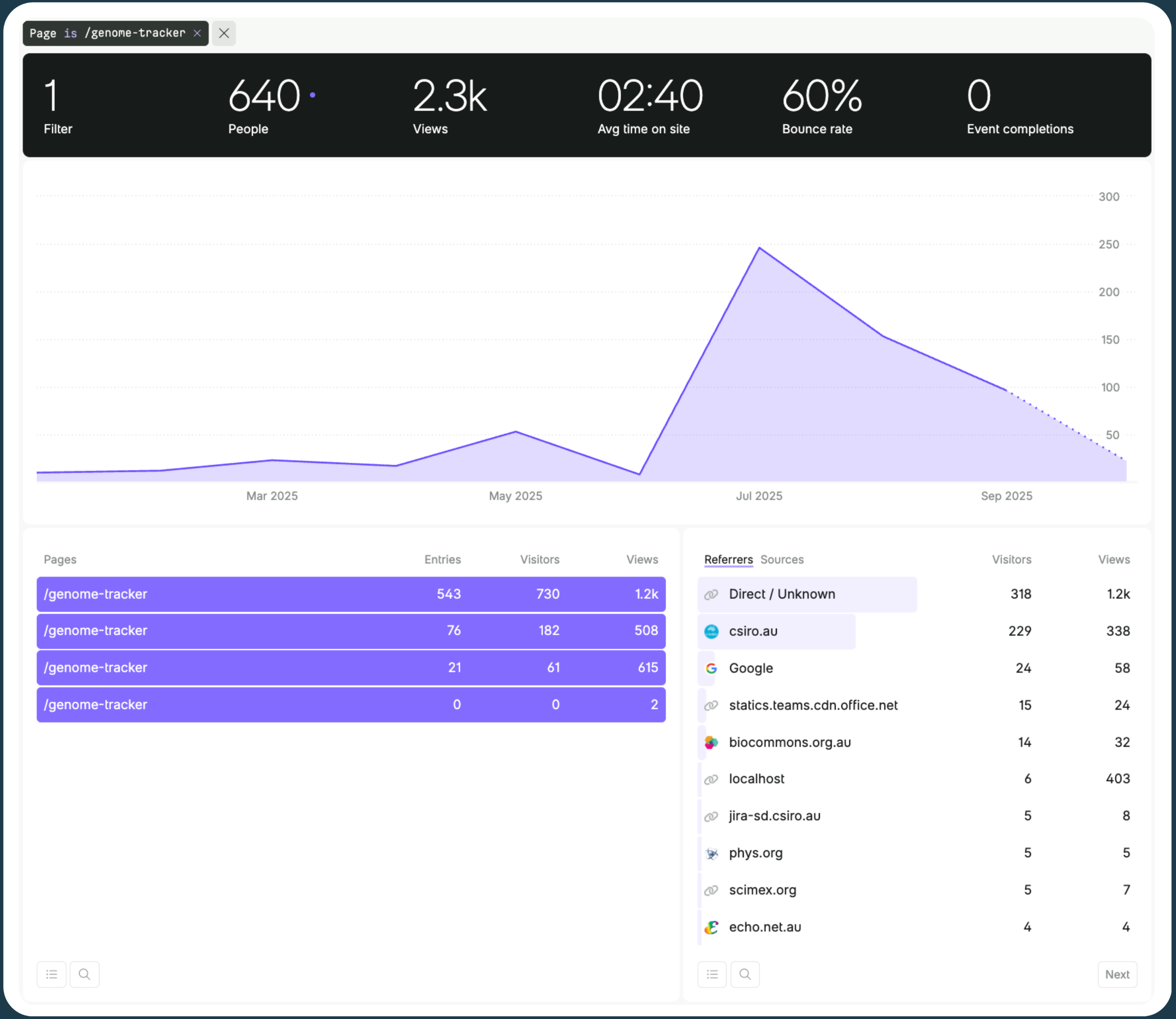
Percentage of taxonomic group coverage, where there is a complete genome for at least one representative species from each grouping.

- detailed **taxon views** for lower ranks
- pinwheel to show internal coverage
- **tachometer** to track percentage with genomes
- species **leaderboard**

research planning and gap-finding

Genome Tracker

progress, transparency, and Australia’s contribution to global genomic discovery



1 Filter
640 People
2.3k Views
02:40 Avg time on site
60% Bounce rate
0 Event completions

Pages	Entries	Visitors	Views
/genome-tracker	543	730	1.2k
/genome-tracker	76	182	508
/genome-tracker	21	61	615
/genome-tracker	0	0	2

Referrers Sources	Visitors	Views
Direct / Unknown	318	1.2k
csiro.au	229	338
Google	24	58
statics.teams.cdn.office.net	15	24
biocommons.org.au	14	32
localhost	6	403
jira-sd.csiro.au	5	8
phys.org	5	5
scimex.org	5	7
echo.net.au	4	4

- **640 unique users** (2.3k views)
- Average **2 mins 40 seconds** on site shows purposeful **interaction**
- **CSIRO** (229 visitors) and **BioCommons** referrals prove **ecosystem** recognition

↓

meaningful engagement

Source:
Fathom analytics. https://usefathom.com/

Insights from the data sea floor

- real time updating makes Genome Tracker a living ecosystem for **data exploration**
- gap-finding capability moves ARGA from static data index to **discovery acceleration engine**

ARGA Team

ARGA Core Project Team

Kathryn Hall: Scientific Product Manager
Jack Brinkman: Software Engineer
Keeva Connolly: Scientific Business Analyst
Christopher Mangion: Data Engineer
Goran Sterjov: Senior Software Engineer

ARGA Project Management

Sarah Richmond
Andrew Turley
Nigel Ward

Other contributors

Matt Andrews
Peter Brenton
Simon Checksfield
Nick dos Remedios
Yasima Kankanamge
Winnie Mok
Vikas Nagaraju
Lars Nauheimer
Caitlin Ramsay

Key contacts

https://arga.org.au

https://app.arga.org.au

https://app.arga.org.au/genome-tracker

kathryn.hall@csiro.au

ARGA
Australian Reference Genome Atlas