Contribution ID: **89** Type: **Presentation**

# The Australian Reference Genome Atlas: supercharged exploratory infrastructure for national-scale genomic data discovery

*Monday 13 October 2025 14:30 (11 minutes)*

The Australian Reference Genome Atlas (ARGA) is a next-generation platform designed to index, connect and expose genomic data for Australia's mega-biodiversity. It sits at the interface between the twin problems of genomic data discovery in an age where data are rapidly proliferating, and the crisis in documenting and understanding Australia's vulnerable biodiversity. More than 80% of Australia's estimated 500,000 species are endemic, including diverse lineages of marsupials, reptiles, flowering plants, fungi, and marine invertebrates. This exceptional biodiversity presents both an opportunity and a challenge for genomic science: the need to coordinate, contextualise, and make accessible a growing body of data across highly diverse taxa and ecosystems.

Despite the rapid proliferation of genomics datasets, researchers face persistent obstacles to discovery and reuse. Genomic data are scattered across disconnected repositories, stored under inconsistent taxonomies, and often lack sufficient provenance metadata to support informed reuse. No single source captures the full range of sequence types, methods, and specimen contexts relevant to a given taxon. This fragmentation significantly hampers the biosciences community's ability to conduct comprehensive, comparative, or ecologically contextualised research.

ARGA provides a concrete response to these challenges, combining rigorous data stewardship with practical infrastructure for researchers. For example, a conservation biologist studying threatened plants can use ARGA to locate and compare available genome assemblies for Critically Endangered plant species, trace sample provenance from herbarium vouchers through to public sequence repositories, and identify key taxonomic gaps where genomic data are still lacking. As a national infrastructure platform, ARGA has been purpose-built to bring FAIR and TRUST principles to life, making genomic data for Australian biodiversity taxa not only findable, accessible, interoperable, and reusable, but also transparent, contextualised, and trusted.

With the foundations now in place following an ambitious two-year development pilot, ARGA is ready to be commended to the scientific community for integration into research workflows. It represents the culmination of technical innovation, collaborative platform design, and principled data stewardship. Developed by the Atlas of Living Australia, Bioplatforms Australia, and the Australian BioCommons, with investment from the Australian Research Data Commons, ARGA offers researchers a unified platform (https://app.arga.org.au) for exploring genome assemblies, annotations, barcodes and marker sequences, and linked specimen metadata, contextualised through taxonomic, geographic, and traits filters.

ARGA's architecture harmonises Darwin Core standards with MIxS checklists (Genomic Standards Consortium) via a custom event model that traces the provenance of genomic data derived from biological samples. At the core of ARGA is a belief in transparent infrastructure: every datum indexed in ARGA is traceable to its source. A specimen-to-sequence timeline allows researchers to interrogate data quality, completeness, and methodology.

ARGA's technical architecture is purposefully lightweight and independent. A React-based frontend supports intuitive exploration of taxonomically indexed data, while a GraphQL layer provides fine-grained control over queries. Underneath, PostgreSQL serves as the backbone for structured metadata, supported by a custom Rust-based resolver layer optimised for speed and stability. Harmonisation of data across external sources (including NCBI GenBank, Barcode of Life Data Systems, and Bioplatforms Australia Data Portal) is achieved through ingest pipelines that map records to a shared, extensible event model aligned with Darwin Core concepts. These architectural choices are deliberate: to ensure flexibility, and to support an Open Source, Open

Science ecosystem. All code is maintained in public repositories under a copyleft licence, and the platform is structured to support community reuse, extension, and review.

ARGA was co-designed with users to prioritise clarity over complexity. Researchers can navigate by systematic groupings, explore ecological traits, and expose under-sequenced lineages. Key features of the ARGA platform include:
- rich metadata and visualisations of **genomic data** for species, with integrated download functionality and evidenced taxonomic histories;
- **taxon dashboards** showing genome coverage, gaps, and sequencing progress by systematic rank;
- **specimen-to-sequence timelines** that visualise provenance from original collection to data reuse;
- **trait-based filtering** for ecological and management attributes (*e.g.* bushfire vulnerability, invasive species);
- **curated species lists** drawn from authoritative sources to guide strategic data use;
- **linked specimen metadata** from museums, herbaria, and biobanks;
- **persistent identifiers and transparent mappings** to support reproducibility and trust.

FAIR and TRUST principles are foundational. From transparent mappings of openly available vocabularies to fully citable and reproducible data downloads, ARGA is engineered to be not just functional, but credible. It is a place where the absence of data is as visible as its presence —where researchers can engage critically with the structure, lineage, and limitations of the data they use.

SciDataCon 2025 at International Data Week marks the full product launch of ARGA. Here we demonstrate the platform's functionality, share technical and governance lessons, and discuss future product direction and planned integrations. We will showcase key product features, including Genome Tracker, a newly developed visual tool to assess genomic coverage across Australia's biota, which we see as having utility as a strategic planning aid and gap analysis tool for both research and policy sectors. Tools like Genome Tracker have been made possible through key data architecture decisions made early in the conceptualisation of ARGA, and demonstrate the breadth of data insights and dividends that can be actualised from a core commitment to data provenance principles.

**Primary authors:** Dr HALL, Kathryn (Atlas of Living Australia, CSIRO); Mr BRINKMAN, Jack (Atlas of Living Australia, CSIRO); Ms CONNOLLY, Keeva (Australian BioCommons); Mr MANGION, Christopher (Atlas of Living Australia, CSIRO); Ms MOK, Winnie; Mr STERJOV, Goran (Atlas of Living Australia, CSIRO)

**Co-authors:** Mr ANDREWS, Matt (Atlas of Living Australia, CSIRO); Mr BRENTON, Peter (Atlas of Living Australia, CSIRO); Mr CHECKSFIELD, Simon (Atlas of Living Australia, CSIRO); Dr CHRISTIANSEN, Jeff (Australian BioCommons); Dr DOS REMEDIOS, Nick (Atlas of Living Australia, CSIRO); Mr HOLEWA, Hamish (Australian Research Data Commons (ARDC)); Ms KANKANAMGE, Yasmina (Atlas of Living Australia, CSIRO); Mr NAGARAJU, Vikas (Atlas of Living Australia, CSIRO); Dr NAUHEIMER, Lars (Atlas of Living Australia, CSIRO); Mr RAMSAY, Caitlin (Atlas of Living Australia, CSIRO); Ms RICHMOND, Sarah (Bioplatforms Australia); Dr WARD, Nigel (Australian BioCommons)

**Presenter:** Dr HALL, Kathryn (Atlas of Living Australia, CSIRO)

**Session Classification:** Presentations Session 2: Data and Research & Data Science and Data Analysis

**Track Classification:** SciDataCon Persistent Themes: Data and Research