Building an Australian Reference Genome Atlas

Nicholas J dos Remedios[‡], Sarah Richmond[§], Jeff Christiansen[|], Nigel Ward[|], Hamish Holewa[‡], Kathryn A Hall[¶]

‡ Atlas of Living Australia, Canberra, Australia
§ Bioplatforms Australia, Macquarie University, Australia
| Australian BioCommons, Brisbane, Australia
¶ Atlas of Living Australia, Dutton Park, Australia

Corresponding author: Kathryn A Hall (kathryn.hall@csiro.au)

Abstract

Currently, genomics data for living species are stored in public and private repositories online. These repositories remain largely disconnected and only partially findable. The Australian Reference Genome Atlas (ARGA) Project is solving the problem of genomics data obscurity by creating an online platform where life sciences researchers can comprehensively and confidently search for data for taxa relevant to Australian research. At its most basic, ARGA is a tool for aggregating and indexing publicly available genomics (and genetics) data. We aim to improve the experience of discovering and accessing this data by building search functionality, based on features such as phenotypic traits and predicted and observed species distributions, and supporting data packaging and transfer to analysis environments. ARGA will index GenBank (National Institutes of Health (NIH), USA), the European Nucleotide Archive (EMBL-ENA), the database of Bioplatforms Australia, and selected DNA repositories in Australian faunal collections and herbaria. We will integrate these records with the occurrence records and taxonomic framework of the Atlas of Living Australia (ALA) to enrich the data and make it searchable using taxonomy, location, ecological characteristics and selected phenotypic data.

The chief aims and outputs for the project are to:

1. create a system to enable contextual metadata about a species to be used as a pointer to a variety of genomic data associated with that species;
2. add functionality to that system to enable additional contextual information groupings, and community curation of these created groupings;
3. create a user-facing web-accessible interface for the system; and
4. devise a mechanism that allows the researchers searching the multiple genomic repositories, via ARGA, to select files for subsequent analysis and export them to other cloud-based analysis infrastructure.

Our approach to ARGA incorporates:

- ingesting species metadata from multiple sequence repositories into a consistent data format using Darwin Core Archive (DwC-A);
- processing metadata using the Pipelines system developed by the Global Biodiversity Information Facility (GBIF), and as implemented in the ALA and other Living Atlases.
- indexing metadata using a Solr search engine; and
- providing a front-end web interface for users to find, select and export sequence files to a number of cloud-based analysis platforms.

Here we will present an overview of the ARGA infrastructure and demonstrate an early prototype of the platform. We will show how ARGA can be used to interrogate DNA sequence records for taxa relevant to Australian research questions, realising a vision where genomics-based solutions to biological questions in conservation, ecology, agriculture and biosecurity can be manifested.

Keywords

database, index, genomics, bioinformatics, DNA sequence data, reference genomes, Darwin Core, data mobilisation

Presenting author

Kathryn Hall

Presented at

TDWG 2022

Acknowledgements

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 .

Funding program

ARDC Bushfire Data Challenge (https://doi.org/10.47486/DC011).

Grant title

Establishing an Australian Reference Genome Atlas (ARGA) and a leadership application in bushfire data.

Hosting institution

Atlas of Living Australia (ALA)

Conflicts of interest

none reported