The Australian Reference Genome Atlas (ARGA): Finding, sharing and reusing Australian genomics data in an occurrence-driven context

Kathryn Hall[‡], Matt Andrews[§], Keeva Connolly[|], Yasima Kankanamge[§], Christopher Mangion[|], Winnie Mok[¶], Lars Nauheimer[#], Goran Sterjov[¤], Nigel Ward[|], Peter Brenton[§]

‡ Altas of Living Australia, Dutton Park, Australia
§ Atlas of Living Australia, Canberra, Australia
| Australian BioCommons, Brisbane, Australia
¶ Australian BioCommons, Melbourne, Australia
# Atlas of Living Australia, Cairns, Australia
¤ Atlas of Living Australia, Melbourne, Australia

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

Abstract

Fundamental to the capacity of Australia's 15,000 biosciences researchers to answer questions in taxonomy, phylogeny, evolution, conservation, and applied fields like crop improvement and biosecurity, is access to trusted genomics (and genetics) datasets. Historically, researchers turned to single points of origin, like GenBank (part of the United States' National Center for Biotechnology Information), to find the reference or comparative data they needed, but the rapidity of data generation using next-gen methods, and the enormous size and diversity of datasets derived from next-gen sequencing methods, mean that single databases no longer contain all data of a specific class, which may be attributable to individual taxa, nor the full breadth of data types relevant for that taxon. Comprehensively searching for taxonomically relevant data, and indeed, data of types germane to the research question, is a significant challenge for researchers. Data are openly available online, but the data may be stored under synonyms or indexed via unconventional taxonomies. Data repositories are largely disconnected and researchers must visit multiple sites to have confidence that their searches have been exhaustive. Databases may focus on single data types and not store or reference other data assets, though they may be relevant for the taxon of interest. Additionally, our survey of the genomics community indicated that researchers are less likely to trust data with inadequately evidenced provenance metadata. This means that genomics data are hard to find and are often untrusted. Moreover, even once found, the data are in formats that do not interoperate with occurrence and ecological datasets, such as those housed in the Atlas of Living Australia.

We built the Australian Reference Genome Atlas (ARGA) to overcome the barriers faced by researchers in finding and collating genomics data for Australia's species, and we

have built it so that researchers can search for data within taxonomically accepted contexts and defined intersections and conjunctions with verified and expert ecological datasets. Using a series of ingestion scripts, the ARGA data team has implemented new and customised data mappings that effectively integrate genomics data, ecological traits, and occurrence data within an extended Darwin Core Event framework (GBIF 2018). Here, we will demonstrate how the architecture we derived for ARGA application works, and how it can be extended as new data sources emerge. We then demonstrate how our flexible model can be used to:

- locate genomics data for taxa of interest;
- explore data within an ecological context; and
- calculate metrics for data availability for provincial bioregions.

Keywords

Darwin Core mapping

Presenting author

Kathryn Hall

Presented at

TDWG 2023

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.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- GBIF (2018) Best Practices in Publishing Sampling-event data, version 2.2. Copenhagen: GBIF Secretariat. URL: https://ipt.gbif.org/manual/en/ipt/2.5/best-practices-sampling-event-data