Genome Annotation Infrastructure Roadmap for Australia

V4.0
14 July 2020

Tiffanie Nelson, Philippa Griffin and Jeffrey H Christiansen

doi: 10.5281/zenodo.3942716

Contents

1. Executive Summary

Genome annotation is the process of identifying and labelling features on a genome assembly. The higher the quality of the annotation, the higher the utility of the genome in comparative, functional, evolutionary and ecological genomics.

In Australia, species targeted for genome annotation span the whole tree of life, often with a focus on endemic Australian species, and many of the genome annotations generated in this country are of great interest to the rest of the world as representative key taxa for investigating evolutionary relationships and understanding gene and genome function across a variety of species.

This document includes:

- a general summary of genome annotation tools and methodologies,
- how the Australian community currently undertakes this work and common data-, software- or compute-related infrastructure challenges faced by that community when using current approaches (information obtained through consultation with a 'Special Interest Group' (SIG) of researchers undertaking genome annotation across Australia), and
- a high level description of four key components of a shared infrastructure for Australia, which, when implemented, would enable Australian researchers from a wide range of institutions to perform genome annotation work who would otherwise be unable to do so because of the reported data-, software- or compute-related roadblocks. i.e:

> **D1. A platform for performing genome annotation:** *to provide Australian researchers with access to a shared platform with selected tools and workflows for genome annotation, underpinned by sufficient compute resources, and easily connected to a variety of data storage locations and key datasets from public repositories.*
>
> **D2 and D3. Systems to enable sharing and improvement of draft genome annotations, and submission of those from Australia to appropriate global repositories:** *to make it easier for any Australian researcher to share their genome annotation files publicly, and globally, in accordance with best-practice open science guidelines.*
>
> **D4. A collaborative connection to the world-leading EMBL-EBI and Ensembl genome annotation teams:** *to enable Australian researchers to collaborate directly with the Ensembl team at EMBL-EBI - through training and upskilling activities; through access to the GeneBuild annotation pipeline-installed at EMBL-EBI via a merit-based scheme; and to have input on standards, taxon priorities and improving ease of submission to international repositories.*

This document has incorporated feedback on earlier versions received from the Genome Annotation SIG, other Australian researchers undertaking genome annotation, the Ensembl team at EMBL-EBI, and Australian research IT infrastructure partners. Implementation of the infrastructure outlined in this roadmap document will begin in 2020 according to the timelines indicated in Section 4.4.

2. Background and Context

In late 2018, the EMBL-Australia Bioinformatics Resource (EMBL-ABR)[1] convened a “Genome Annotation Special Interest Group (SIG)” and invited participation from over 50 researchers across Australia with either experience in, or interest in, genome annotation[2].

That work produced a first version of a **Genome Annotation Infrastructure Roadmap for Australia**[3] which outlined the genome annotation methods and tools used across the SIG, as well as actual or expected data-, software- or compute-related infrastructure roadblocks and challenges described by members of the SIG. That document also described the broad features/requirements for shared, national infrastructure solution options that could help address the challenges described by the SIG and forms the basis of a Genome Annotation Infrastructure Roadmap for Australia.

Since Q1 2019, work towards production of a final Genome Annotation Infrastructure Roadmap for Australia has been carried out through the Bioplatforms Australia[4], ARDC[5] and AARNet[6] funded **Australian BioCommons Pathfinder Project**[7], where the document is an output of the “Communities and Infrastructure Services Identified for Genome Annotation” activity[8].

Community input is welcomed at all times, as is nomination of additional members of the SIG, by emailing communities@biocommons.org.au .

This document (V4.0) has incorporated feedback on earlier versions[3] which were received from the SIG and other Australian researchers undertaking genome annotation, the Ensembl genome annotation resources and analysis infrastructure team[9] at EMBL’s European Bioinformatics Institute (EMBL-EBI), and numerous Australian research IT infrastructure partners[10].

Implementation of the infrastructure outlined in this roadmap document will begin in 2020 (according to the timelines indicated in Section 4.4).

[1] https://doi.org/10.1093/bib/bbx071
[2] see Sections 3.2 and 3.3 for methodology employed for formation of the group and membership
[3] See https://docs.google.com/document/d/1iLHOLbvFHm85bpPVnQbby3ukhz_SyJ1FQ9Q6r1Noq1s, https://docs.google.com/document/d/16goXrxbM4nuQsGoOekou16u77nrwawYYrG_IYAjvJ9M, https://docs.google.com/document/d/1JndHXJegGr1HJmU2vdJ4-GGkIk-d1E5zpltTo3numb8
[4] https://www.bioplatforms.com/
[5] https://ardc.edu.au
[6] https://www.aarnet.edu.au
[7] http://biocommons.org.au/pathfinding
[8] https://www.biocommons.org.au/pathfinder-assembly-annotation
[9] https://www.ebi.ac.uk/about/people/paul-flicek
[10] AAF, AARNet, ARDC, NCI, Pawsey, QCIF, SIH.

3. Genome Annotation - methods and community

3.1 What is genome annotation and how is it done?

Genome annotation is the process of identifying features of interest in a genomic sequence (e.g. genes, repetitive elements, promoters) and associating them with putative functional information. One predicted gene can have multiple annotated features: for example, location, coding status, name, evidence used for prediction, and protein ID are all annotations. An unannotated genome assembly is of limited use, as it consists of vast stretches of unlabelled DNA sequence, whereas the higher the quality of the annotation, the more is known about the genome features, and higher the utility of the (annotated) genome in comparative genomics, functional genomics, and evolutionary and ecological genomics.

The general process of genome annotation requires a genome assembly as input[11], and can be broken down into 5 high-level steps (Aken *et al* 2016):

- genome preparation (repeat identification and masking),
- structural annotation (identifying the location and structure of genes and other genomic features),
- functional annotation (predicting function of the genomic features),
- manual correction (i.e. curation of automatically-produced genomic features), and
- visualisation (typically using a genome browser displaying multiple data tracks).

The genome preparation and structural and functional annotation steps rely on external evidence as well as the primary, input genome sequence. For example, RNA sequence data is commonly used as a source of structural evidence (Dominguez Del Angel *et al* 2017). In this approach, long (e.g. Iso-Seq) or short (e.g. Illumina HiSeq) RNAseq reads are processed, aligned to the input genome assembly, and used to assign gene locations. Protein sequence, ChIPseq, exon capture and other kinds of data can be used in a similar way. External data from high-confidence, well-curated protein databases (e.g. UniRef) can also be used to provide a source of known coding genes that can be computationally searched for and identified in the species of interest. Other databases (e.g. Pfam, a database of protein families with shared structural domains) may also be used to annotate putative function of genomic features (Mudge & Harrow 2016). Most methods combine the use of additional evidence with prediction of genomic features based on the genome sequence alone (Dominguez Del Angel *et al* 2017).

The highest-quality annotations often involve expert human examination and correction/curation of the many (and often independently assigned) automatically-generated features (Aken *et al* 2016). This usually requires visualising the genome assembly with associated automated

[11] Note that methods and tools for genome assembly are not included in this plan, which exclusively focuses on the challenge of genome annotation. See the accompanying Genome Assembly Infrastructure Roadmap for Australia.

annotation(s) and other data (typically as ‘tracks’ using a genome browser), and a capacity to manually edit the automated annotations.

As well as annotating new genome assemblies, there is a real need for ongoing improvement and curation of existing genome annotations (Aken *et al* 2016) to achieve accurate and reliable gene models and other elements. Community efforts may be required to coordinate this and it is important to provide a software infrastructure to facilitate the updating of genomic data (Dominguez Del Angel *et al* 2017). Visualisation tools are also crucial for collaborative work and data sharing. Choices about how to perform manual curation will differ for different species and require extensive community coordination and ongoing communication (Dominguez Del Angel *et al* 2017). It is worth noting that manual annotation is not possible with the number of genomes and genome versions being produced. Automated annotations do not suffer human bias and so can be more readily compared. Both approaches are valuable and necessary.

3.2 Who in Australia is annotating genomes, and which species are they tackling?

The critical importance of genomics as a key methodology to help to address challenges of strategic importance to Australia is outlined in several decadal plans for science[12]: Biodiversity[13], Agricultural Science[14], Marine Science[15] and Ecoscience[16]. Annotating genomes in whole or in part from a wide and diverse range of organisms will be a key process that must be undertaken to fully realise the application of genomics within this vision.

To date, the number of annotated genomes that have been generated and reported in the literature by Australian researchers is fairly modest (See Figure 1). However, the advent of affordable genome sequencing is enabling whole genome assembly as a routine method for groups working on a variety of non-model organisms, and many groups and consortia across Australia are actively working on producing high-quality genome assemblies, including (but not limited to): native mammals[17], plants[18]; agricultural crops[19],[20],[21] livestock[22], endangered species[23] , or other representative organisms[24]. Many or all of the genome assemblies produced in these efforts will need to be annotated.

[12] 10-year strategic plans for science disciplines, developed by the Australian Academy of Science’s National Committees for Science.
[13] https://www.science.org.au/support/analysis/decadal-plans-science/discovering-biodiversity-decadal-plan-taxonomy
[14] https://www.science.org.au/support/analysis/decadal-plans-science/decadal-plan-agricultural-sciences-2017-2026
[15] https://www.science.org.au/support/analysis/reports/national-marine-science-plan
[16] https://www.science.org.au/support/analysis/reports/foundations-future-long-term-plan-australian-ecosystem-science
[17] https://ozmammalsgenomics.com/whole-genomes/
[18] https://www.genomicsforaustralianplants.com
[19] https://stories.uq.edu.au/qaafi/innovation-digital-agriculture/index.html#group-genomics-and-genetics-JwvYS6ADsl
[20] http://appliedbioinformatics.com.au/index.php/Main_Page
[21] https://www.scu.edu.au/southern-cross-plant-science/research/breeding-genetics-and-genomics/
[22] https://qaafi.uq.edu.au/centre-for-animal-science
[23] https://crowdfunding.sydney.edu.au/project/15030
[24] https://www.dnazoo.org/post/dna-zoo-australia-going-live

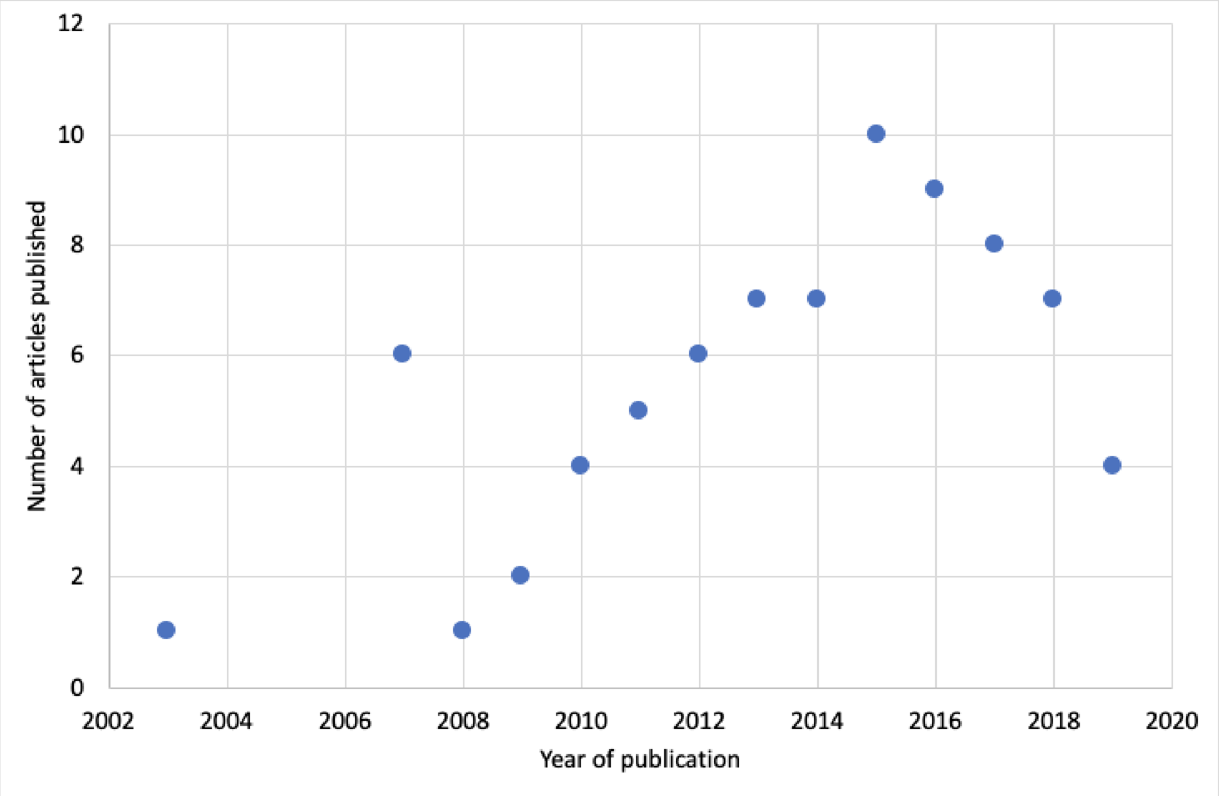

Figure 1: Estimates of the numbers of genomes being annotated Australia.
In order to gain an estimate of the numbers of genomes that have been historically annotated in Australia, a search was conducted of the Scopus database for articles with: ‘genome annotation’ or ‘annotated genome’ in the title, abstract or keyword; and ‘Australia’ in the ‘affiliation’. Articles retrieved from the search were manually reviewed to include only those whose focus included the production of an annotated genome (and exclude others whose focus was on developing or evaluating genome annotation methods or tools). The search was conducted on 01/07/2019 therefore the number shown for 2019 is not for the entire calendar year. The complete list of citations including abstracts can be found here.

In 2018, the EMBL Australia Bioinformatics Resource (EMBL-ABR) invited over 50 researchers across Australia to participate in a Special Interest Group (SIG). These researchers were identified through their involvement in Bioplatforms Australia-facilitated framework dataset projects (e.g. Oz Mammals Genomics Initiative, Genomics of Australian Plants Initiative), or via EMBL-ABR Heads of Nodes suggestions as researchers having either experience, or interest in, genome annotation.

EMBL-ABR sought information from the SIG about each participant’s level of expertise, current (and desired) practices via an online survey; and also held an open video conference follow-up[25] to gather further information.

Within the SIG, species targeted for genome annotation span the whole tree of life, from bacteria through to plants and animals, often with a focus on endemic Australian species. This makes the genome annotations generated in Australia of great interest to the rest of the world,

[25] On 4th October 2018 - Meeting agenda and minutes

as these are often key taxa for investigating evolutionary relationships and understanding gene and genome function across a variety of species.

3.3 How is genome annotation being done in Australia (and why)?

3.3.1 Tools

Based on information received from the SIG through the online survey or during the subsequent open videoconference, a number of tools are used by Australian researchers for various aspects of genome annotation.

These, and others used internationally, are described in the ‘Tool-level table’ tab in this spreadsheet.

All annotation workflows reported as being used by SIG members (see Survey results) require multiple software tools or pipelines; and typically multiple tools are available for each step of a workflow (see background information spreadsheet).

Researchers in the SIG reported choosing a specific tool to use for a particular step of the process based on: (a) what was already accessible to them or their collaborators, (b) ease of installation, (c) ease of use, and (d) best/accepted practice. It is also worth noting that researchers who seldom perform genome annotation are likely to choose tools based on availability, but in cases where annotation of tens or hundreds of genomes is undertaken, factors such as throughput and efficiency are likely to come into play.

A few researchers reported that while certain tools or workflows would be preferred, difficulties they had encountered with resourcing, installing and/or administering these tools within their own group (e.g. TEdenovo and the Ensembl Genebuild pipeline) precluded their use. Some researchers may prefer or require particular tools: for example, some tools may have been designed for use in one taxon but remain unsuitable for use with other taxa. Very large genomes like wheat were also reported to be *potentially* problematic (for example, the genomes are so large they cannot be indexed using standard approaches)[26].

3.3.2 Data

New datasets

All researchers surveyed are continuously generating new datasets to inform their genome annotation work (see Survey results and Survey summary). The most common datatypes used to inform the annotations are gene expression and reference transcriptomes (RNAseq),

[26] Note however that the international wheat community have developed a workaround for this unusually large genome (D. Edwards, UWA - pers comm).

generated through both short-read and long-read sequencing. Most researchers also had access to genetic variation data and some were generating ChIP-Seq datasets.

Existing datasets / databases

Most researchers also reported using data sourced from multiple external databases to inform the genome annotation process. These included Repbase (a database of repetitive DNA elements), UniProt (a protein sequence database combining automatically and manually curated records), UniRef (a non-redundant protein sequence database from UniProt), UniGene (a non-redundant database of clustered coding sequences), RefSeq (a partially curated database of genomes with their transcripts and proteins), and MiRBase (a database of published microRNA sequences and annotations), as well as smaller datasets like the BUSCO (and its precursor CEGMA) sets of universal core genes.

In addition to these pan-species resources, researchers also make use of species-specific datasets that correspond either to their study species or to other species that are related (ranging from closely to distantly) to the study species. These may be publicly available or ‘private’ datasets where access is mediated through some mechanism. Despite most researchers wanting to make use of as much data as is available, a number of researchers did report access roadblocks or quality issues with some private/mediated-access datasets (see Survey results and Survey summary).

3.3.3 Compute infrastructure

Researchers in the SIG currently use a variety of compute infrastructure for genome annotation and may mix-and-match depending on what is available to them via their institutional affiliations. High-performance computing at their (and/or collaborators’) host institution/university, State and National (Tier 1) high-performance computing infrastructure, national NCRIS-funded cloud computing (NeCTAR cloud) and local desktop resources were all used and most researchers used more than one of the above types of compute infrastructure (more investigation would be needed to ascertain exactly when each resource is used and why). Only one of the surveyed researchers reported that they were currently using a commercial cloud provider.

3.4 Challenges being faced

3.4.1 What technical challenges are researchers encountering?

A variety of limitations/challenges with current computational infrastructures that have been identified by the SIG, including:

- Standard walltime limits imposed by compute infrastructure providers are not sufficient to complete jobs (however, this may result from a lack of technical IT expertise to run the jobs efficiently on various infrastructures, for example as batches).
- Difficulties installing some software tools

- Data format issues (using MAKER pipeline with custom repeat annotation)
- Data size
    - limits on genome size accepted by indexing tools
    - Data sharing challenges (at least 2 groups regularly shipping hard drives due to data volumes)
- Problems meeting international repository file submission requirements, transferring files to/from these repositories, and obtaining assistance in doing so.
- For researchers based in government research organisations (e.g. those situated in museums, herbaria or departments such as primary industries) compute limitations due to general IT rules of the government department (e.g. firewalls, bandwidth limits, data transfer limits).
- Assessing the quality of an assembled genome is tightly linked to annotating the genome. The current methods for assessing genome assemblies may assess connectivity but not correctness and this is an ongoing challenge with complex data.
- The time it takes to produce a high-quality genome assembly can take many years and thus many genomes may remain in a draft format and not made available through international repositories or publications. As there is a benefit to having access to all available genomes, draft or complete, access to these will be worthwhile for practitioners.

A clear majority of respondents will either need expanded/replacement compute infrastructure for their genome annotation work, or are unsure what their requirements will be, both for 2 years from now (10/15 respondents) and 5 years from now (13/15 respondents).

Half of the surveyed researchers are not currently making their genome annotations available in the standard genomic data repositories (GenBank or ENA) because of data transfer issues, problems complying with the required format, or lack of expertise with the process itself. Even those that do so, reported challenges complying with repository format requirements (e.g. see 3.4.2 below).

3.4.2 Standards and Interoperability challenges

Several members of the SIG expressed opinions that it would be very important that any nationally available shared data resource should comply with (and utilise wherever possible) community-endorsed (i.e. ‘best practice’) formats, standards, metadata schemas and controlled vocabularies/ontologies, as well as any other standards required by international repositories.

Although GFF3 is considered the best-practice format for annotation files[27], several researchers reported encountering data format incompatibility issues in the past when attempting to submit

[27] The GFF3 format is better described and allows for a richer annotation than GTF https://www.ncbi.nlm.nih.gov/genbank/genomes_gff/

GFF3 annotations to public repositories (GenBank)[28]. It was felt it would be valuable to have a more direct national community voice to GenBank and/or ENA and be able to contribute to changes required by the research community.

This area needs continued scoping in the context of genome annotation in Australia. Understanding format requirements (both input and output) of software tools and workflows will be an important factor in decisions about software suitability. Metadata standards for annotation files and the associated ontologies should also be further investigated and described.

3.4.3 Upskilling challenges

There is a vast array of software tools available to perform various steps of the annotation process and it can be extremely difficult or daunting as an inexperienced user to assess the suitability and compatibility of each tool.

Whilst most researchers get their information from existing discussion forums and numerous publications, more than half of survey respondents indicated it was crucial or important to access in-person training (how to use the platform or tools/pipelines), and nearly all stated that it was important to maintain access to numerous existing discussion forums to share expertise with other users (e.g. to answer questions such as: whether tools are appropriate for their study species; which alternative tool for a pipeline step might be most suitable for their input data; which infrastructures have been used to successfully deploy and run certain tools/pipelines).

3.5 Is a shared national solution palatable to the research community?

While some researchers surveyed have strong preferences for specific software tools for genome annotation, others are more tool-agnostic and would be prepared to make use of good-quality annotation workflows if provided for them, especially if tool maintenance was taken care of externally.

All surveyed researchers agreed that a shared data collaboration / analysis platform would overcome many of the identified challenges, and said they would use such a platform provided it was well designed and supported.

[28] GenBank does allow GFF3 submission but through a beta process https://www.ncbi.nlm.nih.gov/genbank/genomes_gff/

4. Meeting the needs of Australian researchers for high-quality, accessible genome annotation infrastructure

4.1 Goal/Aim

To develop a ‘Genome Annotation Infrastructure Roadmap for Australia’ that describes collaborative infrastructure, which, when implemented (from 2020 onwards), will enable Australian researchers from a wide range of institutions to perform high-quality genome annotation work who would otherwise be unable to do so because of data-, expertise-, software- or compute-related infrastructure roadblocks.

Four versions of the Roadmap document are planned, each to incorporate content and feedback from different groups. Planned dates for development of the Roadmap are as follows:

- v1 - content based on SIG survey results and input from SIG meeting - Feb 2019
- v2 - content modified to incorporate feedback from SIG, other researchers undertaking genome annotation and international partners - Oct 2019
- v3 - content modified to incorporate feedback from various national computational infrastructure providers - Nov 2019
- v4 - content modified to incorporate final feedback from SIG - Jan 2020

4.2 Objectives

The high level objectives of deploying the proposed infrastructure and associated services are:

A. to provide Australian researchers with access to a platform with:
    a) a selection of tools and workflows that will allow genome annotation to be performed,
    b) underpinned by sufficient computational infrastructure and resources, and
    c) easily connected to a variety of data storage locations and key datasets from public repositories.

B. to make it easier for Australian researchers to share and improve their draft annotations and to also publish their polished genome annotation files in accordance with best-practice open science guidelines.

C. to enable Australian researchers to collaboratively access expertise and infrastructure within the Ensembl genome annotation team at EMBL-EBI - i.e.
    a. to access training and advice around annotation approaches, quality and support when utilising locally-installed annotation software tools and workflows;

b. to enable merit-based collaborative access to the EMBL-EBI-hosted Ensembl GeneBuild pipeline for appropriate datasets; and

c. to enable the Australian community to have input on standards, taxon priorities and improving ease of submission to international repositories.

4.3 Outputs

To address the objectives, four broad outputs/ infrastructure components are proposed for implementation:

- **D1 - A platform to enable collaborative genome annotation;**
- **D2 - A system to enable sharing and improvement of draft genome annotations;**
- **D3 - A system to enable submission of publication ready genome annotations from Australia to appropriate global repositories;**
- **D4 - A collaborative connection to EMBL-EBI and the Ensembl genome annotation team.**

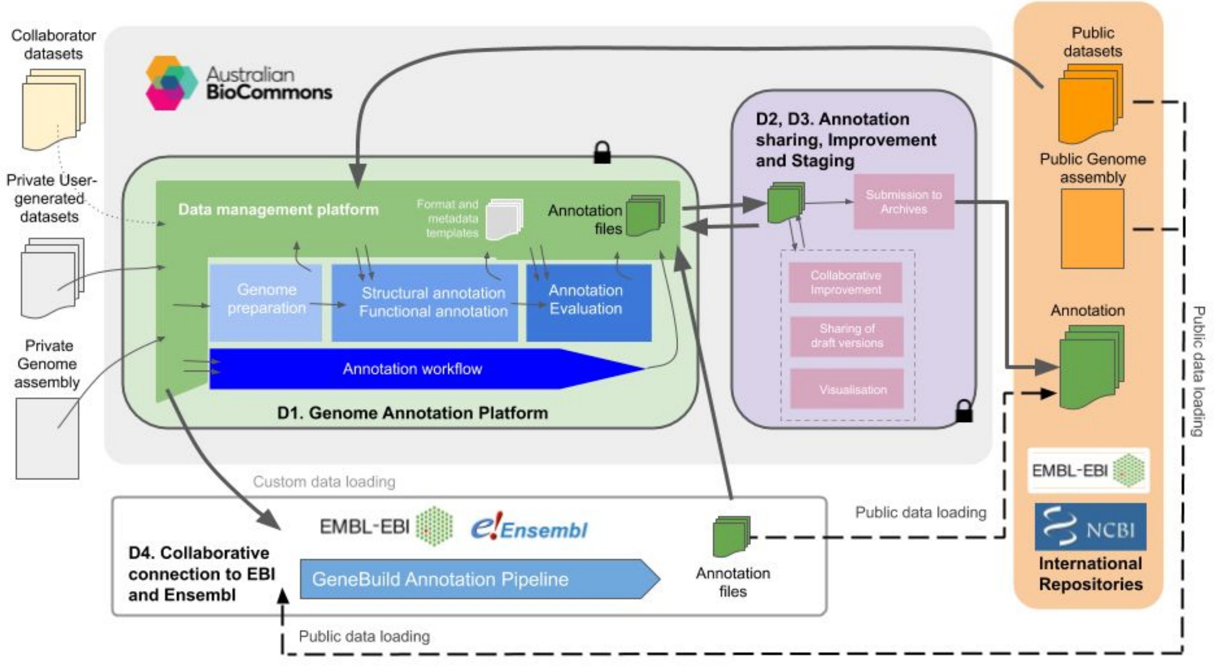

Figure 2. Schematic showing data flow through the proposed Genome Annotation Platform (D1), Data Staging Post (D2, D3) and International Repositories. Direct links to Ensembl Genebuild infrastructure and genome annotation expertise housed at EMBL-EBI is also shown (D4). Arrows indicate the flow of data. Thick arrows show the rapid data transfer capability. Dashed lines show the scenario when public data that is already housed in archives (e.g. ENA) is loaded into the GeneBuild pipeline. Blue shapes indicate stages of the genome annotation process. See Appendix 1 for a list of tools/pipelines it is proposed may be included. Higher resolution image.

D1. A platform to enable collaborative genome annotation

To address objective 1 (i.e. providing Australian researchers with access to a selection of tools, workflows and reference datasets that will allow genome annotation to be performed), it is proposed to implement a platform in Australia[29], that:

A. Includes a set of key tools for data preparation, genome annotation and evaluation (selected from those listed in Appendix 1) installed (plus all other dependencies).
B. Is easily connected to a variety of data storage locations (e.g. institutional or other data storage), and with the ability to upload/mount user generated datasets that are required as inputs for an annotation pipeline.
C. Appropriate user authorisation and sharing mechanisms to allow for data sharing, solely at the discretion of a data owner/custodian
D. Includes high-speed access to well managed, synchronised copies of key reference datasets from international repositories (and the ability to make subsets of these where relevant) when these are required as inputs for an annotation pipeline.
E. Is tightly associated with a data management component that contains shared metadata templates that include all elements required to enable seamless submission of annotation files to international repositories (when required).
F. Is underpinned by Australian BioCommons associated computational resources[30].
G. Has support available from experts for installation of extra software tools and maintenance, and download/upload of key datasets as needed.
H. Includes documentation (including a knowledgebase with community contributed content) and
I. Includes training for all of the above.

D2. A system to enable collaborative improvement of draft genome annotations

To address objective 2 (i.e. to make it easier for Australian researchers to share and improve their draft annotations) it is proposed to implement:

A. A system to enable sharing of draft genome annotations (including the visualisation of these through a genome browser).
B. with resources/tools available for collaborative curation / improvement of the annotation quality of draft genome annotations
C. with appropriate user authorisation and sharing mechanisms to allow for data sharing, solely at the discretion of a data owner/custodian.

[29] Subject to the results of a platform functionality comparison / gap analysis, scoping of compute requirements, agreement with various computational providers about hosting, and outcomes of further consultation with end users

[30] See https://www.biocommons.org.au/pathfinder-biocloud

D. documentation on how to use the system (including a knowledgebase with community contributed content)
E. training.

D3. A system to enable submission of publication ready genome annotations from Australia to appropriate global repositories

To address objective 2 (i.e. to make it easier for Australian researchers to submit publication ready genome annotations from Australia to appropriate global repositories) it is proposed to implement[31]:

A. A temporary data storage ‘staging post’ in Australia for annotation files ready for public international release.
B. Has support available from experts in formatting data and curating metadata to comply with ENA repository format requirements[32].
C. Includes a rapid data transfer from the ‘staging post’ to the ENA archive.
D. documentation on how to use the system (including a knowledgebase with community contributed content)

D4. A collaborative connection to EMBL-EBI and the Ensembl genome annotation team

To address objective 3 (i.e. to enable Australian researchers to access training and advice around annotation approaches, quality and support when utilising locally-installed annotation software tools and workflows; enable merit-based collaborative access to the EMBL-EBI-hosted Ensembl GeneBuild pipeline for appropriate datasets; and to enable the Australian community to have input on standards, taxon priorities and improving ease of submission to international repositories), it is proposed to implement[33]:

A. An Australian BioCommons-funded long-stay travel program to enable meritorious Australian researchers to visit EMBL-EBI and exploit the EMBL-EBI-hosted Ensembl GeneBuild annotation capability through direct collaboration with the Ensembl genome annotation group. Triaging of appropriate datasets for annotation using this method would be based on multiple criteria, to be set by the Ensembl genome annotation group

[31] Subject to the outcomes of further consultation with EMBL-EBI and end users
[32] potentially building on the previous data submission service which was offered by the EMBL-ABR: QCIF node
[33] Subject to agreement with Ensembl and EMBL-EBI

in collaboration with the Australian BioCommons[34]. Proposals must include a clear plan for effective communication back in Australia about the outcomes of the visit.

B. Webinar training events led by Ensembl genome annotation experts, for Australian researchers undertaking genome annotation, including potential users of the merit-based collaborative access scheme to the EMBL-EBI-hosted Ensembl GeneBuild pipeline.
C. Australian BioCommons-funded short-stay travel grants for Ensembl team members to visit Australia. This visit should incorporate at least 1 seminar/webinar and 1 workshop (either BioCommons "hybrid" training or face-to-face training) centred around the GeneBuild annotation pipeline, as well as other Ensembl resources. Preference would be given to proposals that incorporate attendance at key Australian meetings, and a clear plan for ongoing collaboration after the visit.
D. Ongoing representation of the Australian genome annotation community at key GenBank, Ensembl and EMBL-EBI meetings. This should involve regular community consultation, remote attendance at meetings, and reporting back to the community. This could be one aspect of a broader outreach role that may cover multiple projects and BioCommons deliverables. Alternatively it could be a paid contribution to a researcher with an existing role.

4.4 Implementation Timelines

For proposed implementation timelines, please see the separate (living) document: Genome Annotation Infrastructure: Proposed Implementation Timeline (Dec 2019).

[34] There are two primary modes of sourcing input data for the GeneBuild pipeline, and it is envisaged that both options should be supported in the proposed Australian BioCommons infrastructure framework: (a) a preferred 'public data loading' option where the input genome assembly, RNAseq and other data are sourced from the public ENA and/or other international archives. When this option is used, it is expected that the resultant annotation will be made publicly available via Ensembl in the subsequent release cycle. (b) a 'custom data loading' option where the input genome assembly, RNAseq and other data are not necessarily public and can be uploaded from anywhere (and in the proposed framework, from the BioCommons supported data management platform). Where possible, the public data loading option is the option preferred by the EMBL-EBI Ensembl operators of the GeneBuild pipeline. For whichever option is utilised, it is critical that appropriate metadata related to the sample/species etc is associated with all input data to ensure the GeneBuild pipeline can link data where appropriate.

References

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T. and Howe, K. 2016. The Ensembl gene annotation system. Database, 2016. https://doi.org/10.1093/database/baw093

Dominguez Del Angel V, Hjerde E, Sterck L et al. Ten steps to get started in Genome Assembly and Annotation [version 1; referees: 2 approved]. F1000Research 2018, 7(ELIXIR):148 https://f1000research.com/articles/7-148/v1

Mudge, J. and Harrow, J. (2016), The state of play in higher eukaryote genome annotation. Nature Reviews Genetics, 17: 758-772. https://www.nature.com/articles/nrg.2016.119

Appendix 1

Genome annotation tools for consideration for inclusion in a shared analysis environment. Note that the Ensembl Genebuild workflow incorporates a number of other tools not listed as line items here. More detail can be found in https://docs.google.com/spreadsheets/d/1-QhdwvhIy7fNAGVSA8XvalzOXZaUL4yO2zPhUUqd6AQ/edit#gid=0

Workflow Step	High-level component	Component	Subcomponent	Tool(s)	Download URL	Info/manual URL	Publication DOI
1 to 3	Workflow	Workflow	Workflow	**Ensembl Genebuild Workflow (including all component tools)**	https://github.com/Ensembl		https://doi.org/10.1093/database/baw093
1.1	Genome preparation	Repeat Identification and Masking	Repeat Identification and Masking	**RepeatRunner**	http://www.yandell-lab.org/downloads/repeat_runner.tar.gz	http://www.yandell-lab.org/software/repeatrunner_docs.html	https://doi.org/10.1016/j.gene.2006.09.011
1.1	Genome preparation	Repeat Identification and Masking	Repeat Identification and Masking	**REPET**	https://urgi.versailles.inra.fr/download/repet/REPET_linux-x64-2.5.tar.gz	https://urgi.versailles.inra.fr/Tools/REPET	https://doi.org/10.1371/journal.pone.0016526 , https://doi.org/10.1371/journal.pcbi.0010022
1.1	Genome preparation	Repeat Identification and Masking	Repeat Identification and Masking	**WindowMasker**	ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools++/CURRENT/		10.1093/bioinformatics/bti774
1.1.1	Genome preparation	Repeat Identification and Masking	Repeat Identification	**PILER-DF**	https://www.drive5.com/piler/	https://www.drive5.com/piler/piler_userguide.html	https://doi.org/10.1093/bioinformatics/bti1003
2	Structural genome annotation	Structural genome annotation	Combined evidence method.	**JAMg**	https://github.com/genomecuration/JAMg	http://jamg.sourceforge.net/	
2	Structural genome annotation	Structural genome annotation	Combined evidence method (can also be used ab initio).	**Augustus**	https://github.com/Gaius-Augustus/Augustus	https://github.com/Gaius-Augustus/Augustus/blob/master/README.md	doi: 10.1093/bioinformatics/btw494
2	Structural genome annotation	Structural genome annotation	Combined evidence method.	**MAKER (current version MAKER2)**	http://yandell.topaz.genetics.utah.edu/cgi-bin/maker_license.cgi	http://weatherby.genetics.utah.edu/MAKER/wiki/index.php	10.1186/1471-2105-12-491
2	Structural genome annotation	Structural genome annotation	Combined evidence method.	**MAKER-P**	via http://www.yandell-lab.org/software/maker.html	http://weatherby.genetics.utah.edu/MAKER/wiki/index.php	https://doi.org/10.1104/pp.113.230144

2	Structural genome annotation	Structural genome annotation	Combined evidence method	**funannotate**	https://funannotate.readthedocs.io/en/latest/install.html	https://funannotate.readthedocs.io/en/latest/	
2.1.1	Structural genome annotation	Structural genome annotation from assembly (ab initio methods)	Gene prediction from assembly	**Genemark**	http://topaz.gatech.edu/GeneMark/license_download.cgi	http://exon.gatech.edu/GeneMark/background.html	https://doi.org/10.1093/nar/gki937
2.1.1	Structural genome annotation	Structural genome annotation from assembly (ab initio methods)	Gene prediction from assembly	**SNAP**	https://github.com/KorfLab/SNAP	https://github.com/KorfLab/SNAP/blob/master/README.md	https://doi.org/10.1186/1471-2105-5-59
2	Structural genome annotation	Structural genome annotation	Combined evidence method	**BRAKER2**	https://github.com/Gaius-Augustus/BRAKER	https://github.com/Gaius-Augustus/BRAKER/blob/master/docs/userguide.pdf	10.1093/bioinformatics/btv661
2.2.2	Structural genome annotation	Structural genome annotation from assembly + same-individual or same-taxon genomic data	Gene prediction (and/or improvement) from RNAseq data	**PASA**	https://github.com/PASApipeline/PASApipeline	https://github.com/PASApipeline/PASApipeline/wiki	https://dx.doi.org/10.1093%2Fnar%2Fgkg770
2.2.4	Structural genome annotation	Structural genome annotation from assembly + same-individual or same-taxon genomic data	Gene prediction from long-read RNA data	**Iso-Seq Tofu, PacBio ICE, GMAP in pipeline**			
2.3.1.3	Structural genome annotation	Structural genome annotation from assembly + external databases	Short non-coding RNA prediction (specifically 2'-O-methylation guide snoRNAs) from assembly	**snoscan**	http://eddylab.org/software/snoscan/snoscan.tar.gz	http://eddylab.org/software/snoscan/snoscan.README	10.1126/science.283.5405.1168
2.3.1.5	Structural genome annotation	Structural genome annotation from assembly + external databases	Special prediction for miRNAs	**Splign**	ftp://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/splign/linux-i64/	https://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi?textpage=documentation	10.1186/1745-6150-3-20
2.3.1.6	Structural genome annotation	Structural genome annotation from assembly + external databases	Special prediction for rRNAs, snoRNAs and snRNAs	**Infernal**	http://eddylab.org/infernal/	http://eddylab.org/infernal/Userguide.pdf	10.1093/bioinformatics/btt509
2.3.4	Structural genome annotation	Structural genome annotation from assembly + external databases	Transfer of annotations using conserved synteny (positional information)	**RATT**	http://ratt.sourceforge.net/	http://ratt.sourceforge.net/documentation.html	10.1093/nar/gkq1268

2.4	Structural genome annotation	Structural genome annotation quality control and improvement	Gene model filtering to produce final protein-coding gene set	**MUSCLE**	https://www.drive5.com/muscle/downloads.htm	https://www.drive5.com/muscle/manual/	10.1093/nar/gkh340
2.4	Structural genome annotation	Structural genome annotation quality control and improvement	Gene model filtering to produce final protein-coding gene set	**EVidenceModeler (EVM)**	https://github.com/EVidenceModeler/EVidenceModeler/releases	https://evidencemodeler.github.io/	10.1186/gb-2008-9-1-r7
2.5.1	Structural genome annotation	Structural genome annotation using final transcript set (+ external datasets or pipelines in some cases)	Protein-coding potential prediction from transcript set	**Coding Potential Assessment Tool**	https://sourceforge.net/projects/rna-cpat/files/	http://rna-cpat.sourceforge.net/	10.1093/nar/gkt006
2.5.1	Structural genome annotation	Structural genome annotation using final transcript set (+ external datasets or pipelines in some cases)	Protein-coding potential prediction using external dataset	**Coding Potential Calculator**	http://cpc.cbi.pku.edu.cn/download/	http://cpc.cbi.pku.edu.cn/docs/	10.1093/nar/gkm391
4	Visualisation	Visualisation	Genome browser	**JBrowse**	https://github.com/GMOD/jbrowse	https://jbrowse.org/docs/installation.html	10.1101/gr.094607.109
4	Visualisation	Visualisation	Genome browser	**Ensembl browser**	https://asia.ensembl.org/info/docs/webcode/mirror/install/index.html		
4	Visualisation	Visualisation	Genome browser	**Artemis**	https://github.com/sanger-pathogens/Artemis	https://www.sanger.ac.uk/science/tools/artemis	10.1093/bioinformatics/btr703
5	Data sharing and curation	Collaborative, ongoing manual curation of structural and functional annotations	Collaborative, ongoing manual curation of annotations	**Apollo**	https://github.com/GMOD/Apollo	https://genomearchitect.readthedocs.iofrnaseq/en/latest/	10.1186/gb-2013-14-8-r93 10.5281/zenodo.268535
6	Genome Evaluation	Genome Assembly and Annotation evaluation		**BUSCO**	https://busco.ezlab.org/		
6	Genome Evaluation	Genome Assembly evaluation		**QUAST**	http://bioinf.spbau.ru/quast		
6	Genome Evaluation	Genome Assembly evaluation		**gEVAL**	https://geval.sanger.ac.uk/index.html		
	Genome Evaluation	Genome Assembly evaluation		**Hawkeye**	http://amos.sourceforge.net/wiki/index.php?title=Hawkeye		

Document Control

VERSION	DATE	AUTHOR(S)	DESCRIPTION
v1.0	2019-02-28	Philippa Griffin (EMBL-ABR); Jeff Christiansen (Australian BioCommons)	First draft - intended for community comment by Genome Annotation SIG members and others (please add comments directly into document). Subsequent iterations of the document will be produced following consultation with other stakeholder groups (i.e. international entities operating genome annotation infrastructure elsewhere and Australian research IT infrastructure partners)
v1.1	2019-06-20	Tiff Nelson (Aus BioCommons) Jeff Christiansen (Aus BioCommons)	Modifications made based on community feedback received from v1.0
v2.0	2019-09-25	Jeff Christiansen (Aus BioCommons)	Further modifications made based on further community feedback received from v1.1, and initial verbal discussions between the Ensembl and GeneBuild teams at EBI and Australian BioCommons leadership.
v2.1	2019-10-03	Jeff Christiansen (Aus BioCommons)	Section 4.3 and Fig 2 adjusted to better reflect the conceptual division between D1 (data management and annotation tools/pipelines) and D2 (draft sharing and publishing of polished annotations)
v2.2	2019-10-06	Jeff Christiansen (Aus BioCommons)	Changes to Fig 2 made in line with suggestions from the Ensemble Genome Annotation team at EMBL-EBI
v3.0	2019-12-10	Jeff Christiansen (Aus BioCommons)	Minor modifications made following feedback from Australian research IT providers D2 was split into two components (D2 and D3), and D3 was re-named as D4, to make the items included in the implementation timeline (in Section 4.4) clearer
v4.0	2020-07-14	Jeff Christiansen (Aus BioCommons)	V3.0 document was converted to pdf format for upload to zenodo repository. Modifications included: addition of the list of authors, and a change on page 3 to reflect the updated method for contributing input/comments