

Genome Assembly Infrastructure Roadmap for Australia

V4.0
31 July 2020

Tiffanie Nelson and Jeffrey H Christiansen

doi: 10.5281/zenodo.3967970

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



Contents

1. Executive Summary	2
2. Background and Context	3
3. Genome Assembly - methods and community	4
3.1 What is genome assembly, how is it done, and why?	4
3.2 Who in Australia is performing genome assembly, and which species are they tackling?	6
3.3 How is genome assembly being done in Australia?	8
3.3.1 Data	8
3.3.2 Tools	8
3.3.3 Compute infrastructure	9
3.3.3.1 Types used	9
3.3.3.2. Resourcing	10
3.3.3.3. Systems administration expertise	10
3.4 Challenges being faced	10
3.4.1 What technical challenges are researchers encountering?	10
3.4.2 Data/Methods Management, Standards and Interoperability challenges	11
3.4.3 Upskilling challenges	12
3.5 Is a shared national solution palatable to the research community?	12
4. Meeting the needs of Australian researchers for high-quality, accessible genome assembly infrastructure	14
4.1 Goal/Aim	14
4.2 Objectives	14
4.3 Outputs	15
4.4 Implementation Timelines	19
Appendix 1	21
Appendix 2	27

1. Executive Summary

Genome assembly is the process of taking large numbers of relatively short discrete DNA or nucleotide sequences, termed 'reads', and computationally ordering them correctly to create a representation of the original chromosome or genome from which the DNA originated. To comprehend this challenge further, genomes range in size from 1,000 to more than 100 billion base pairs, and sequence read length technology commonly ranges between 150 to 1,000 base pairs in length, but may go up to 10,000 base pairs. Assembling genomes has the potential to provide numerous insights which can lead to benefits in health, conservation, and agriculture.

In Australia, species targeted for genome assembly span the whole tree of life, often with a focus on endemic Australian species, or economically important agricultural plants and livestock species.

This document includes:

- a brief summary of genome assembly 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 assembly across Australia), and
- a high level description of key components of an envisaged shared national genome assembly infrastructure for Australia, which, when implemented, would enable Australian researchers from a wide range of institutions to perform genome assembly 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 assembly: *to provide all Australian researchers with access to a shared platform with tools and workflows for genome assembly, underpinned by sufficient compute resources and easily connectable to a variety of data storage locations and key datasets from public repositories.*

D2. Systems to enable sharing and visualisation of genome assemblies: *to make it easier for any Australian researcher to share their draft/reference genome assembly files with collaborators or the public.*

D3. Systems to enable submission of genome assembly files from Australia to appropriate global repositories: *to make it easier for any Australian researcher to publish their genome assembly files (raw, contigs and scaffolds) publicly in accordance with best-practice open science guidelines.*

This document has incorporated feedback on [earlier versions](#) received from the Genome Assembly SIG, other Australian researchers undertaking genome assembly, the COPO team at the Earlham Institute (UK), and numerous Australian research IT infrastructure partners. Implementation of the infrastructure outlined in this roadmap document will begin throughout 2020 according to the timelines indicated in [Section 4.4](#).

2. Background and Context

In Australia, investments to establish community-scale bioinformatics infrastructure have materialised in various forms and scales over the last decade under a range of funding schemes. One significant supporter is Bioplatforms Australia¹ which aims to support Australia's national bioinformatics infrastructure, and is funded under the National Collaborative Research Infrastructure Strategy (NCRIS).

Since 2019, Bioplatforms Australia has supported the Australian Biocommons², which is an initiative focussed on providing improved access to bioinformatics tools, methods, and training for Australia's molecular life scientists to underpin world class science. The Australian BioCommons is currently coordinating several national consultations with various communities of practice to gain input from life science researchers, bioinformaticians and infrastructure providers to identify, configure, connect and support bioinformatics infrastructure and resources that are relevant to these research communities.

To support the large (and growing) community of practice in Australia undertaking *de novo* genome assembly, in late January 2020, the Australian BioCommons convened a "Genome Assembly Special Interest Group (SIG)" and invited participation from over 200 researchers across Australia with either experience in, or interest in, genome assembly³.

The outcome of that meeting is this document, which summarises and represents the current or expected infrastructure roadblocks and challenges described by members of the community, and describes the potential broad features and requirements for shared, national infrastructure solution options that could help address these challenges.

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

This document has incorporated feedback on [earlier versions](#) received from the Genome Assembly SIG, other Australian researchers undertaking genome assembly, the COPO team at the Earlham Institute (UK), and numerous Australian research IT infrastructure partners. Implementation of some of the infrastructure outlined in this roadmap document began during Q1/2 2020, and is intended to continue through 2020/21 as indicated in [Section 4.4](#).

¹ Bioplatforms Australia, bioplatforms.com/

² Australian Biocommons, biocommons.org.au

³ see Section 3.2 for methodology employed for formation of the group and membership

3. Genome Assembly - methods and community

3.1 What is genome assembly, how is it done, and why?

Nucleic acid sequencing methods generally rely on approaches where regions of DNA are amplified from an organism and the nucleotide sequence of these DNA fragments (known as 'reads') are determined. All current sequencing technologies⁴ produce read lengths that are vastly shorter than most genomes (shot-gun approach), therefore computational methods are required to conceptually assemble these short sequences into the correct order to create a representation of the (very long) nucleotide sequence of the original chromosome or genome from which the DNA originated. This process is termed 'genome assembly'.

The reads range from 'short' (100-250 base pairs (bp)), through 'long' (10-20 kb) to 'ultra-long' (100-500kb) depending on the technology used. The assembly of small genomes, e.g. viruses and bacteria, ranging from 1 kbp (kilo bp) to 13 Mbp (mega bp)⁵ can often be performed with modest computational resources and a limited time commitment. However, eukaryotic genomes, e.g. animals and plants, whose sizes range from 100 Mbp to 150 Gbp (giga bp), may take months or years to complete high quality, near error free chromosome-level assemblies, representing a major challenge.

During the early 2000s when the first eukaryotic organisms were sequenced and assembled, the technology used⁶ required considerable scale in both time and money. As a result, genome projects were only the domain of a few large efforts and were primarily focused on 'model' organisms (i.e. species that are widely studied, usually because they are easy to maintain and breed in a laboratory setting and have particular experimental advantages). Since the mid-2000s, sequencing technologies have advanced rapidly and also decreased in cost⁷ allowing the method to be applied widely, including to determine the genomic characteristics of non-model organisms (i.e. species that are not necessarily widely studied in a laboratory context).

As a technique, genome assembly is made easier by the existence of a 'reference' genome⁸, which is a previously assembled genome from a related organism that can be used as a guide to inform how a new genome is built, and can enable assembly to be undertaken more quickly

⁴ Various sequencing platforms exist which have unique chemical and technological methods of identifying basepairs in a nucleic acid sequence. These include Illumina (illumina.com/), Pacific Biosciences (www.pacb.com/), Oxford Nanopore Technologies (nanoporetech.com/) and 10x Genomics (10xgenomics.com/)

⁵ Nucleic acid sequence length in numbers of base pairs: 1 kbp (kilo base pairs) = 1,000 bp; 1 Mbp (mega base pairs) = 1,000,000 bp; 1 Gbp (giga base pairs) = 1,000,000,000 bp. metagenomics.wiki/pdf/definition/tera-giga-base-pairs.

⁶ Sanger sequencing was the most common method for the first 40 years of sequencing, with high levels of accuracy. It was overtaken by high-throughput methods as accuracy became comparable but is still in use for low -throughput sequence analysis, small number of samples as it yields read lengths > 500 bp, McCombie et al, Cold Spring Harb. Perspect. Med. 2018, [9/11/a036798](https://doi.org/10.1101/a036798).

⁷ The 3 Gbp human genome was first sequenced in 2001 using Sanger sequencing. It cost billions of dollars and took years of machine time. In 2010 the Illumina HiSeq 2000 could sequence DNA more than 10,000 times faster than automated Sanger sequencing at less than 1/10,000 of the cost, Metzker, Nat. Rev. Genet. 2010, nature.com/articles/nrg2626.

⁸ Reference genomes or a reference assembly is a digital nucleic acid sequence representing an example species' set of genes. They are often assembled from the sequencing of DNA from a number of donors/individuals, Church et al. PLoS Biol. 2011, [10.1371/journal.pbio.1001091](https://doi.org/10.1371/journal.pbio.1001091).

and cheaply⁹. Most research projects however working on non-model organisms lack the availability of a reference genome from a comparable or closely-related species, so in this scenario the reads are assembled into longer sequences with no prior knowledge of the source genome layout or composition. In this process, known as *de novo* assembly (Figure 1), reads are joined through overlapping regions into long contiguous sequences known as 'contigs'. Contigs are then ordered and orientated to one another with gaps of known length to form 'scaffolds'. Where spaces between contigs exist (i.e. gaps), these can be filled by using independent reads to complete an assembly or left as NNNNNNNNNNN's. The scaffolding and gap-filling steps can be performed repeatedly to enhance the quality of the assembly.

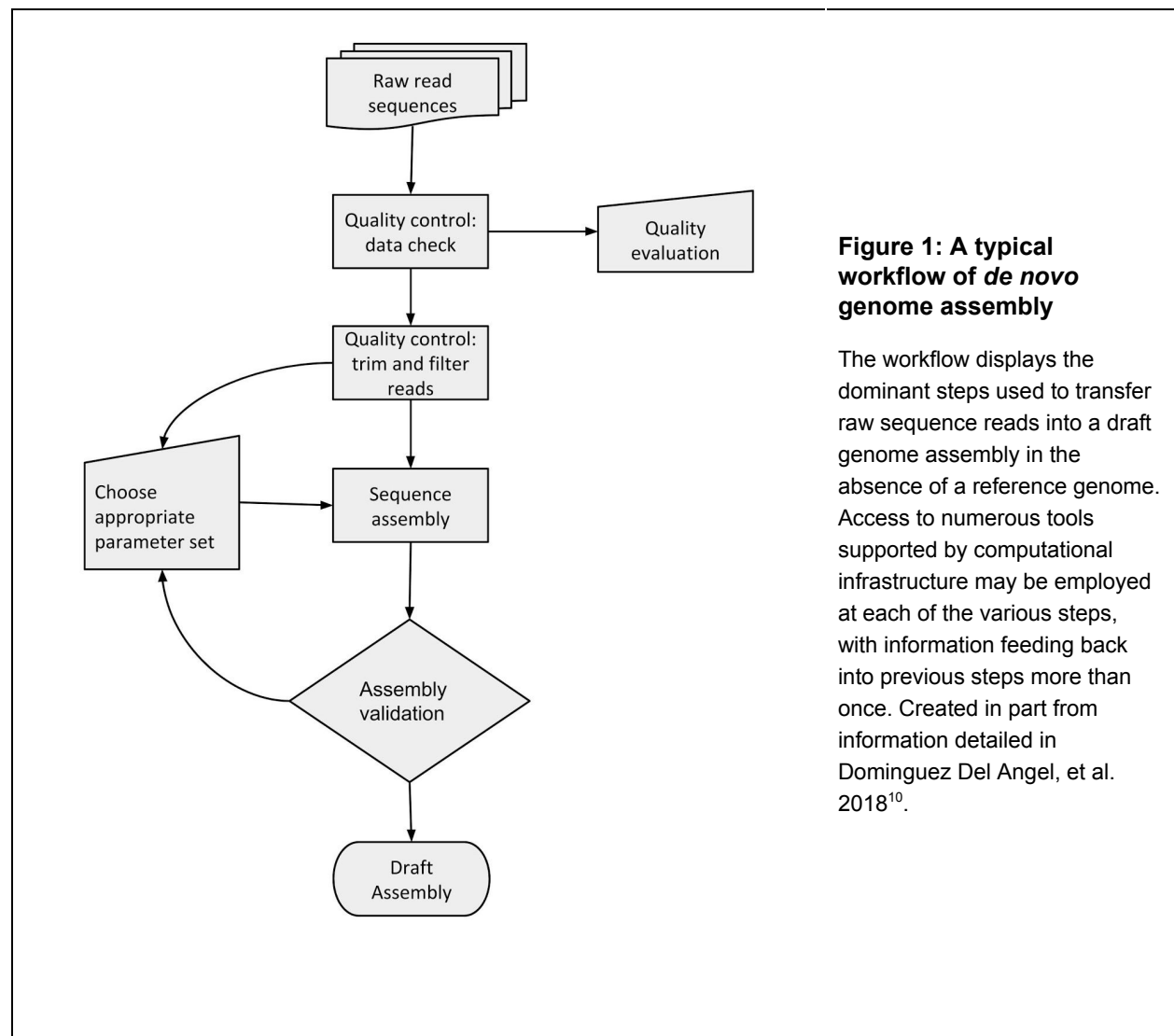


Figure 1: A typical workflow of *de novo* genome assembly

The workflow displays the dominant steps used to transfer raw sequence reads into a draft genome assembly in the absence of a reference genome. Access to numerous tools supported by computational infrastructure may be employed at each of the various steps, with information feeding back into previous steps more than once. Created in part from information detailed in Dominguez Del Angel, et al. 2018¹⁰.

⁹ *De novo* genome assemblies are orders of magnitude slower and more memory intensive than mapping assemblies, Rice and Green, Annu. Rev. Anim. Biosci. 2019, doi.org/10.1146/annurev-animal-020518-115344.

¹⁰ Dominguez Del Angel, et al. F1000Research, 2018, doi.org/10.12688/f1000research.13598.1.

In addition to these general steps, several other issues usually need to be overcome to obtain a high-quality genome assembly. These include: correction of incorrect base reads (sequencing errors), and dealing with the presence of large numbers of identical sequences known as repeats as well as uneven read depth due to biased sequencing of certain areas of DNA (e.g. areas of high guanine-cytosine content, known as GC-rich).

Recently, other technologies and approaches have also been developed to enable more complete genome assemblies to be produced with fewer artefacts (e.g. collapse of repeats or duplicated regions or false duplications of heterozygous regions). These include chromosome conformation capture or physical mapping providing chromosome level genome assemblies, or refinements developed for diploid/polyploid genomes to resolve individual parental haploid chromosomes.

3.2 Who in Australia is performing genome assembly, and which species are they tackling?

The benefits of sequencing and assembling genomes are far-reaching, impacting areas of health, agriculture, and the environment. Benefits include better diagnosis and early detection of disease and identification of treatment options for diseases or in managing risk, such as climate-induced changes^{11,12,13}. Hence, the critical importance of genomics as a key methodology to help to address challenges of strategic importance to Australia is outlined in several Australian Academy of Science Decadal Plans for science¹⁴: Biodiversity¹⁵, Agricultural Science¹⁶, Marine Science¹⁷ and Ecoscience¹⁸. Assembling 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.

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 now actively working on producing high-quality genome

¹¹ Clark, et al. Sci. Transl. Med. 2019, <https://stm.sciencemag.org/content/11/489/eaat6177>

¹² Johnson, et al. Nat. Genet. 2018, <https://doi.org/10.1038/s41588-018-0153-5>.

¹³ Scheben, et al. Curr. Plant Biol. 2016, <https://doi.org/10.1016/j.cpb.2016.09.001>

¹⁴ 10-year strategic plans for science disciplines, developed by the Australian Academy of Science's National Committees for Science.

¹⁵ science.org.au/support/analysis/decadal-plans-science/discovering-biodiversity-decadal-plan-taxonomy

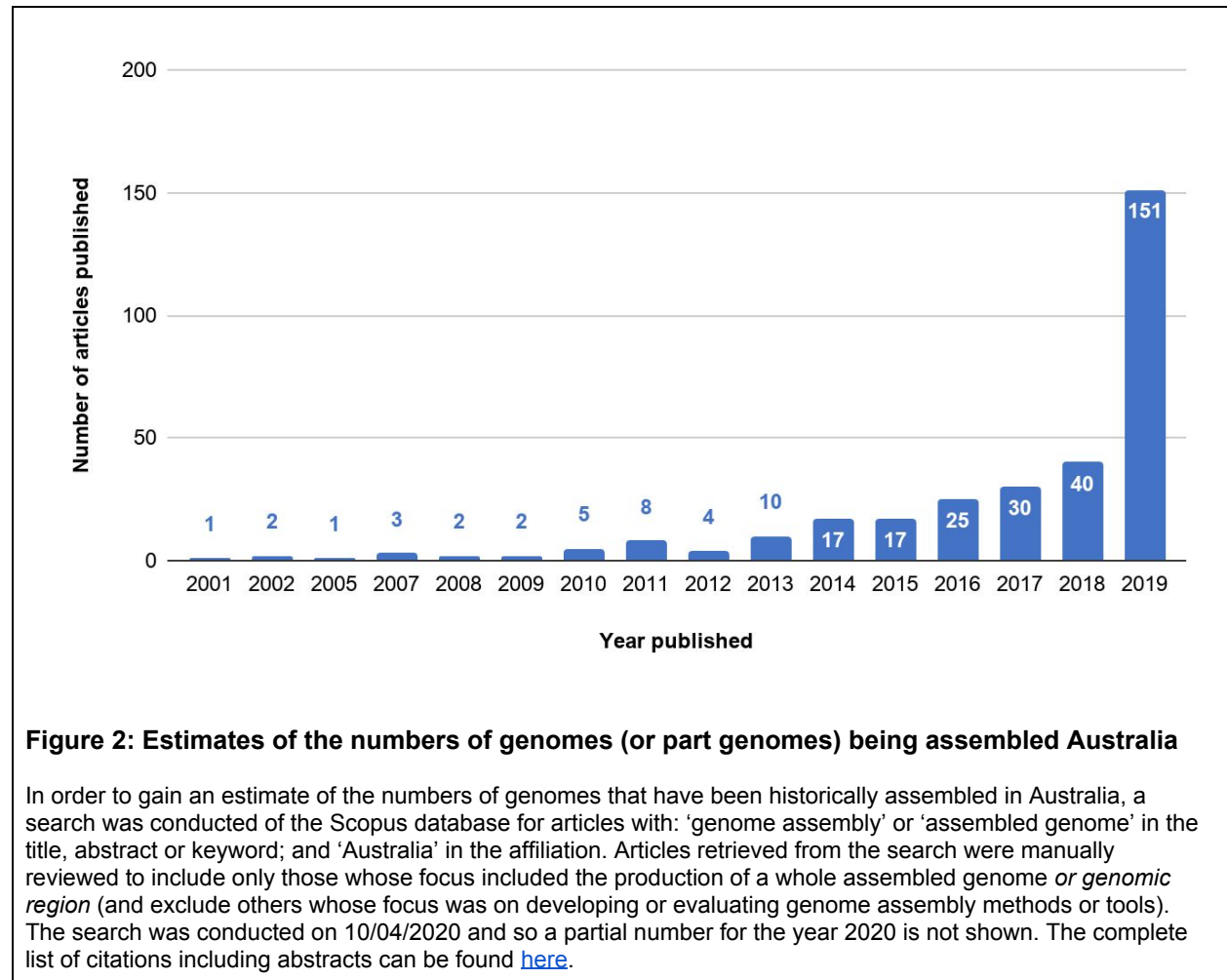
¹⁶ science.org.au/support/analysis/decadal-plans-science/decadal-plan-agricultural-sciences-2017-2026

¹⁷ science.org.au/support/analysis/reports/national-marine-science-plan

¹⁸ science.org.au/support/analysis/reports/foundations-future-long-term-plan-australian-ecosystem-science

assemblies, including (but not limited to): native mammals¹⁹, plants^{20,21}, reptiles and amphibians²²; agricultural crops^{23,24,25}, livestock²⁶; threatened species²⁷, or other representative organisms²⁸.

Searching the scientific literature indicates an approximate number of genome (or part genome) assemblies produced from Australian-based researchers (see Figure 2).



¹⁹ ozmammalsgenomics.com/whole-genomes/

²⁰ genomicsforaustralianplants.com

²¹ bioplatforms.com/projects/native-grasslands/

²² bioplatforms.com/projects/amphibians-reptiles/

²³ stories.uq.edu.au/qaafi/innovation-digital-agriculture/index.html#group-genomics-and-genetics-JwvYS6ADsl

²⁴ appliedbioinformatics.com.au/index.php/Main_Page

²⁵ scu.edu.au/southern-cross-plant-science/research/breeding-genetics-and-genomics/

²⁶ qaafi.uq.edu.au/centre-for-animal-science

²⁷ bioplatforms.com/projects/threatened-species/

²⁸ dnazoo.org/post/dna-zoo-australia-going-live

In late January 2020, the Australian BioCommons invited over 230 researchers across Australia to participate in a Genome Assembly Special Interest Group (SIG). These researchers were identified as having experience in, or interest in, *de novo* genome assembly. The Australian BioCommons sought information from the SIG about each member's level of expertise, current (and desired) practices and infrastructure used via an on-line survey²⁹ (number of respondents = 40), and also held an open video conference follow up to gain further information (minutes³⁰ and a recording³¹ of the meeting are available).

Respondents to the survey indicated they are producing or likely to produce genome assemblies from a wide range of organisms over the next 5 years including bacteria, fungi, flatworms, roundworms, corals, sponges, anemones, cephalopods, marine crustaceans, insects, agricultural crops (including cereals, solanum species, brassicas, legumes, banana, cannabis, macadamia), native trees/shrubs (e.g. eucalypts, proteaceae, acacia), amphibians, reptiles, birds and mammals.

3.3 How is genome assembly being done in Australia?

3.3.1 Data

Based on information received from the SIG members through the [online survey](#) ($n=40$), most researchers use a combination of sequencing platforms to generate their data with the most popular being Illumina³², PacBio³³, 10x³⁴ and Nanopore³⁵. The majority of researchers (87%) also access data to inform their assembly from public repositories or from existing private collections or collaborators.

3.3.2 Tools

Based on the [survey](#), ~70 software tools were identified as being used by respondents for various stages of the genome assembly process. These are listed in [Appendix 1](#) of this document.

There are several tools available for each step in the genome assembly process. While the choice of a specific tool is primarily dictated by the sequencing platform/s that was/were used for data generation, members of the SIG also noted that their choice of tool can also be influenced by whether software is already accessible to them or their collaborators and

²⁹ Presentation including survey results on genome assembly infrastructure needs and challenges conducted 01/12/2019 to 26/01/2020: drive.google.com/file/d/11MggTjmuUqBXx40l2FsdQv86EwxwnL9/view?usp=sharing

³⁰ Meeting minutes from genome assembly SIG meeting held 31/01/2020: [2020-01-31 Assembly Roadmap Meeting Minutes](#)

³¹ Recording of genome assembly SIG meeting held 31/01/2020: [20200131_Assembly_SIG_Meeting_recording.mp4](#)

³² Illumina, Inc. illumina.com/

³³ PacBio, Pacific Biosciences, pacb.com/

³⁴ 10x Genomics, 10xgenomics.com/

³⁵ Oxford Nanopore Technologies, nanoporetech.com/

best/accepted practice. It was also noted by the group that the development of genome assembly tools is rapid – matching the rapid development of sequencing technologies.

A number of researchers (33%, $n=13$) reported that they were not using their preferred tools and instead had resorted to a workaround solution due to difficulties they had encountered with resourcing, installing and/or administering certain tools on shared computational infrastructure that was available to their group. For example, several researchers noted that the high memory requirements of some tools, e.g. Cactus³⁶, Canu³⁷ and Falcon³⁸, were greater than the allocation available to them on their available compute infrastructure. One researcher noted that the I/O patterns of the tool Supernova³⁹ was incompatible with the data storage architecture of systems they had been using. Other issues include tools being incompatible with job scheduling systems on shared compute infrastructure (e.g. Falcon incompatibilities with PBS⁴⁰), a lack of available GPUs (graphics processing units) meaning that some tools needed to run (much more slowly) in CPU (central processing unit) mode which made their use unworkably slow ($n = 1$). Several respondents ($n = 3$) also noted that certain tools were very difficult/tedious to install properly, e.g. Falcon and Canu, and that waiting for system administrators to install tools could take weeks. Together these roadblocks increase the time to research.

3.3.3 Compute infrastructure

3.3.3.1 Types used

Survey respondents ($n = 40$) currently use a wide variety of compute infrastructure for genome assembly, with 80% accessing high performance computing provided by their institute (or a collaborator's institute); 43% utilising national (Tier 1) or state (Tier 2) computing centres (e.g. NCI⁴¹, Pawsey⁴² or QCIF⁴³); 40% using their own laptops or personal computers; 17% accessing NeCTAR cloud compute resources⁴⁴ and 8% using commercial cloud resource, such as Amazon Web Services (AWS)⁴⁵. Less than 1% reported using GUI-based platforms (including Galaxy⁴⁶ and Cyverse⁴⁷) for their genome assembly work. Most respondents (70%) use more than one of these compute-infrastructure types to support their assembly work and mix and match their use.

³⁶ <https://github.com/ComparativeGenomicsToolkit/cactus>

³⁷ <https://canu.readthedocs.io/en/latest/>

³⁸ <https://pb-falcon.readthedocs.io/en/latest/>

³⁹ Supernova, support.10xgenomics.com/de-novo-assembly/software/overview/latest/welcome

⁴⁰ PBS Pro (Portable Batch System Professional) Open Source Project, pbspro.org/

⁴¹ NCI, National Computational Infrastructure, nci.org.au/

⁴² Pawsey Supercomputing Centre, pawsey.org.au/

⁴³ QCIF, Queensland Cyberinfrastructure Foundation, qcif.edu.au/

⁴⁴ Nectar, National eResearch Collaboration Tools and Resources, nectar.org.au/cloudpage/

⁴⁵ Amazon Web Services, aws.amazon.com/

⁴⁶ Galaxy, usegalaxy.org.au/

⁴⁷ CyVerse, cyverse.org/

3.3.3.2. *Resourcing*

More than half of the respondents (63%, $n = 25$) said the infrastructure they currently had access to was not sufficient for their current *de novo* genome assembly work, due to limitations in memory available or data storage allocations, especially for bigger assembly projects. For the remainder (i.e. those who believe that their computational infrastructure is currently sufficient), about half of this group believe their computational infrastructure will not be sufficient for their needs in 2-5 years.

3.3.3.3. *Systems administration expertise*

Most respondents (95%) say that access to systems administration expertise is not a limiting factor to building and maintaining computational infrastructure currently (they access expertise within their group or at the institution housing the HPCs being used).

3.4 Challenges being faced

3.4.1 What technical challenges are researchers encountering?

A variety of limitations/roadblocks/challenges with current computational infrastructures were identified by the SIG, including:

- Computational resources available (even across a variety of infrastructures) can be insufficient to underpin the use of computationally demanding assembly tools or pipelines (e.g. Canu, Cactus and Falcon) - especially for large *de novo* genome⁴⁸ or metagenome assemblies;
- Access to well-resourced shared (Tier1 or 2) computational resources for *testing and trialing* software on these systems is often limited (since access is usually awarded through merit allocation schemes which have stringent requirements to access⁴⁹);
- Use of institutional-provided resources can make sharing data related objects (data, pipelines and results) with collaborators outside of the institution complicated and can be problematic for research collaborations with participants from many institutions;

⁴⁸ For an example of computational resources consumed to conduct a *de novo* genome assembly for a plant species (Golden Wattle) using Canu (#cores and memory required per task), see https://australianbiocommons.github.io/Canu/NCI_optimisation.html

⁴⁹ Application to national and state computational infrastructure is routinely awarded through NCMAS (National Computational Merit Allocation Scheme, ncmas.nci.org.au/2020/) can be challenging for researchers who do not have a sufficient track record in previous computational use and additionally require that applicants demonstrate that the pipelines work prior to access which is difficult to predict with novel datasets. Note however that most HPC centres do also offer "director shares" to allow POCs and to test scaling capabilities.

- There is a general lack of systems for data management in use that support collaboration, and with correct data versioning functionalities, and this can lead to version confusion and task repetition, e.g. re-sharing files with collaborators;
- For researchers not using tools such as [Cloudstor](#), data transfer (between collaborators, facilities, short- and long- term storage) of large amounts of data is seen as being very inefficient/slow and is done in an *ad hoc* manner, using a variety of tools;
- Data publishing to international repositories (e.g. GenBank/SRA at NCBI⁵⁰) is considered by some to be difficult (“*daunting*”, “*horrible*”, “*painful*”, “*time consuming*”) - primarily due to a lack of mechanisms/systems available locally to help ensure files are compliant for submissions, i.e. in the correct data formats and adhering to the metadata standards required by the repositories⁵¹, but also due to the complexities of the NCBI data submission process itself;
- Convenient short-medium term storage options (i.e. for the active lifespan of a research project - months to several years) for raw and intermediate data files and draft assemblies that could be used for repeat analyses on datasets (e.g. repeat base calling as methods improve for some sequence platforms), is lacking;
- A shared community space to help propagate best practices, exchange expertise and experiences in resolving bioinformatic challenges encountered when using software on various infrastructures is lacking. As one researcher put it “*because the field is rapidly moving, word of mouth is often how knowledge is transferred and this will help in resolving bioinformatics issues efficiently*”.

3.4.2 Data/Methods Management, Standards and Interoperability challenges

More than half of respondents (60%, $n = 24$) reported that no specific data or method management tools or frameworks are used to support their genome assembly work. The remaining 40% use a mixture of custom-made in-house data management solutions (e.g. code-written systems), as well as Jupyter Notebooks⁵², Nextflow⁵³ or Snakemake⁵⁴ for methods/code management. Three respondents cited the use of the LabArchives Electronic Lab Notebook⁵⁵, which is provided by their institution.

Most respondents (92%, $n = 38$) stated that best practices in tools, formats and metadata along with compliance with international data repositories would be crucial or important in a shared

⁵⁰ Sequence Read Archive at National Center for Biotechnology Information, ncbi.nlm.nih.gov/sra

⁵¹ Note however that others ($n=4$) explicitly state that submitting data to NCBI is unproblematic.

⁵² Jupyter Notebooks, jupyter.org/

⁵³ Nextflow, nextflow.io/

⁵⁴ Snakemake, snakemake.readthedocs.io/en/stable/

⁵⁵ LabArchives, labarchives.com/, au-mynotebook.labarchives.com/

genome assembly platform. “*Smart metadata handling*” was considered a crucial or important feature of a shared platform for the majority (78%, $n = 37$) of respondents, including assistance with metadata formats, transfer of metadata through pipeline and controlled vocabulary lookup. Mechanisms built-in to ensure sample metadata was collected in accordance with community-agreed standards, and prior to data generation was suggested by one researcher to help ensure later compliance of submission files to international repositories.

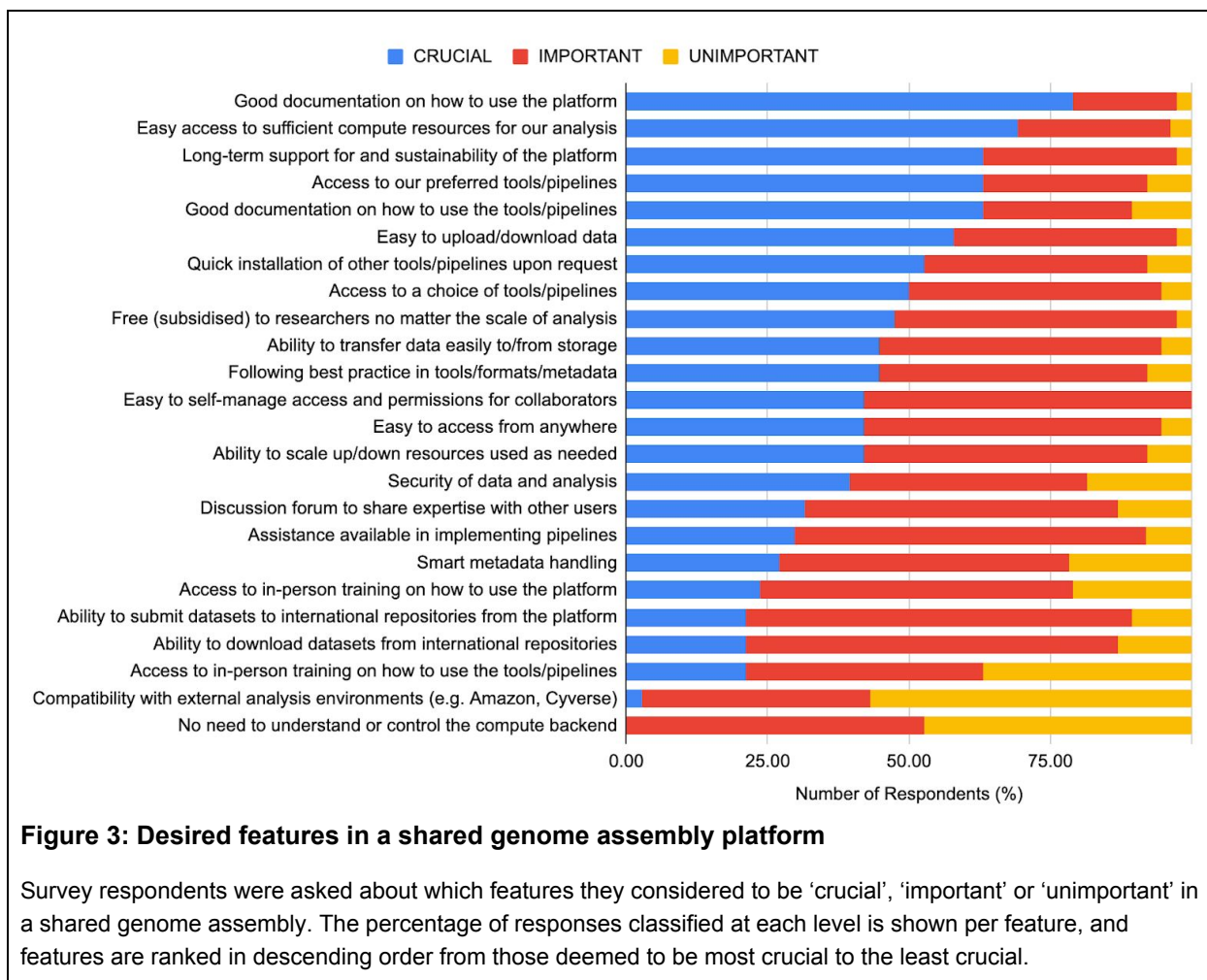
3.4.3 Upskilling challenges

There are many software tools available to perform various steps of the assembly process and it can be difficult or daunting as an inexperienced user to assess the suitability and compatibility of each tool, especially when technologies and available tools are released or changing rapidly. Most respondents (87%, $n = 38$) would welcome a discussion forum or knowledgebase to share expertise with other users. Experienced and inexperienced users alike identify benefits with a shared space to exchange ideas and experiences using tools and various infrastructures - highlighted by the fact that the field is moving rapidly with word of mouth being a strong method of knowledge transfer. In the words of one experienced bioinformatician “*individual research groups do not need to reinvent the wheel, and we can work more as a community*” through sharing pipelines and workflows. It was also noted that there are international as well as national communities that can and should be considered in this space.

Training opportunities on genome assembly workflows themselves, as well as in relation to computational infrastructure building and maintenance were identified by respondents as being of interest, e.g. in the area of software containers, or gaining a better understanding of the variety of application processes to access computational resources on shared computational systems.

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

All of the respondents (100%, $n = 40$) agreed that if a shared data collaboration / analysis platform for *de novo* genome assembly was available for use, they would use such a platform provided it was well designed and supported. This number included respondents who stated that their needs are currently met. 24 hypothetical features of such a system are listed in Figure 3, ranked according to how crucial respondents believe that feature would be (when asked would the feature be ‘crucial’, ‘important’ or ‘unimportant’). The top 5 features of a shared platform deemed the most crucial are: good documentation on how to use the platform (79%); easy access to sufficient to compute resources for our analysis (69%); long term support for and sustainability of the platform (63%); access to our preferred tools/pipelines on any such platform (63%) and good documentation on how to use the tools/pipelines (63%).



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

4.1 Goal/Aim

To develop a 'Genome Assembly Infrastructure Roadmap for Australia' that describes collaborative infrastructure, which, when implemented (from Q3 2020 onwards), will enable Australian researchers from a wide range of institutions to perform high-quality genome assembly 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 - May 2020.
- V2 - Content modified to incorporate feedback from SIG, other researchers undertaking genome assembly and international groups - May 2020.
- V3 - Content modified to incorporate feedback from various national computational infrastructure providers - June 2020.
- V4 - Content modified to incorporate final feedback from SIG - July 2020.

4.2 Objectives

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

1. To provide Australian researchers with access to a platform with:
 - a) A selection of tools and workflows that will allow *de novo* genome assembly to be performed across a wide range of taxa;
 - b) Sufficient computational infrastructure and resources; and,
 - c) Connectivity to a variety of data storage locations.
2. To make it easier for Australian researchers to share and improve draft genome assemblies; and,
3. To make it easier to publish high quality genome assembly files in accordance with best-practice open science guidelines.

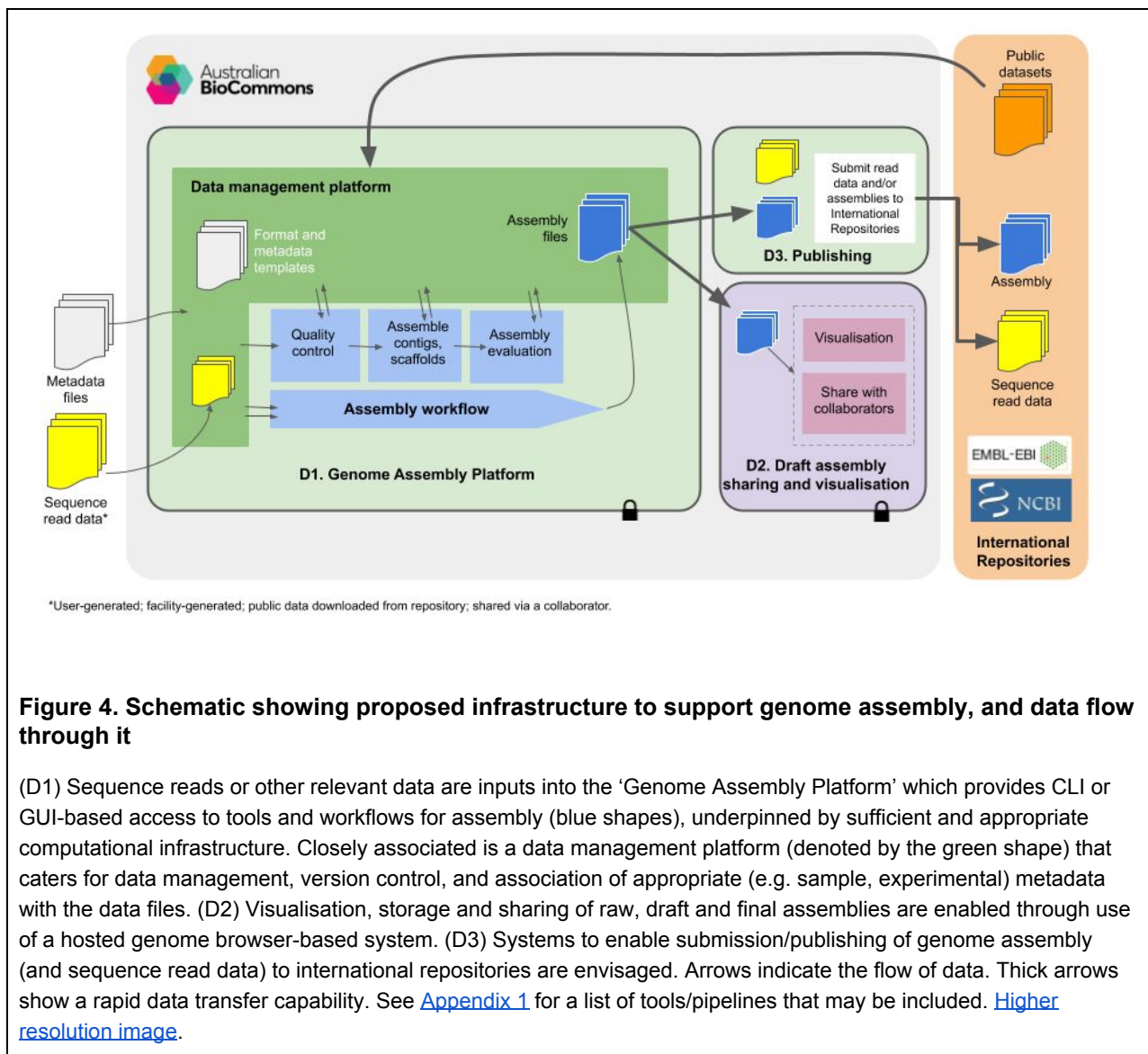
4.3 Outputs

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

D1 - A platform for performing genome assembly;

D2 - Systems to enable sharing and visualisation of draft genome assemblies; and,

D3 - Systems to enable submission of genome assembly files from Australia to appropriate global repositories.



D1 - A platform for performing genome assembly;

To address [objective 1](#) (i.e. providing Australian researchers with access to a selection of tools, workflows underpinned by computational resources that allow genome assembly to be performed), it is proposed to implement a platform in Australia⁵⁶, that:

- A. Includes a set of key tools for data preparation, quality control, genome assembly and evaluation:
 - a. Installed (plus all other dependencies) and optimised on a command line interface (CLI) analysis environment (i.e. across a variety of Tier 1 and 2⁵⁷ shared computational infrastructures) underpinned by appropriate computational resources;
 - b. Installed (plus all other dependencies) and optimised on a graphical user interface (GUI) web-based data analysis platform, (i.e. Galaxy Australia), underpinned by appropriate computational resources⁵⁸;
 - c. Available as high quality, trusted software containers for self-deployment on institutional or independent computational infrastructures.
- B. Has support available from experts for installation/containerisation of extra software tools and maintenance with version control and updates as required;
- C. Is easily connectable to a variety of data storage locations, i.e. international, national (i.e. Cloudstor), institutional or other data storage, and with the ability to upload/mount user generated or other datasets that are required as inputs for an assembly pipeline;
- D. Has appropriate user authorisation and sharing mechanisms to allow for data sharing, solely at the discretion of a data owner/custodian;
- E. Is tightly associated with a data management component that contains shared metadata templates that include all elements required to enable submission of assembly files to international repositories, when required;

⁵⁶ 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.

⁵⁷ The definition of Peak (Tier 1) High Performance Computing (HPC) is traditionally defined as a compute capability that is in the top 200 globally. Australia's current Tier 1 facilities are: [NCI](#) and [Pawsey](#). Examples of Tier 2 facilities include State-level systems such as [QRIScloud](#) operated by QCIF, and [Time](#) operated by Intersect, and many institutionally operated facilities.

⁵⁸ See biocommons.org.au/pathfinder-biocloud

- F. Support available from experts in formatting data and curating metadata to comply with NCBI and ENA repository format requirements⁵⁹;
- G. Includes documentation, including a knowledgebase with community contributed content; and,
- H. Includes training for all the above.

D2 - Systems to enable sharing and visualisation of draft genome assemblies

To address [objective 2](#) (i.e. to make it easier for Australian researchers to share draft genome assemblies and evaluate these by checking against known data, to identify and correct errors) it is proposed to implement:

- A. A hosted genome-browser based (e.g. JBrowse⁶⁰ or Ensembl-empowered⁶¹) framework to enable researchers to upload, share and visualise draft genome assemblies⁶².
- B. Appropriate user authorisation and sharing mechanisms to allow for public or private data sharing, solely at the discretion of a data owner/custodian;
- C. Documentation on how to use the system (including a knowledgebase with community contributed content); and,
- D. Training.

D3 - Systems to enable submission of genome assembly files from Australia to appropriate global repositories.

To address [objective 3](#) (i.e. to make it easier to publish high quality and share final genome assemblies (and relevant input data) in accordance with best-practice open science guidelines) it is proposed to implement:

- A. A temporary 'staging post' in Australia for genome assembly (and sequence read) files ready for public international release. The system should include data/metadata formatting checks (which would be enabled by the use of the data management platforms described in [D1-D](#)), and support as detailed in [D1-E](#);

⁵⁹ potentially building on the previous [data submission service](#) which was offered nationally by the EMBL-ABR: QCIF node, and is now available to researchers from QCIF/QFAB member organisations

⁶⁰ The JBrowse Genome Browser, jbrowse.org/

⁶¹ Projects using Ensembl, asia.ensembl.org/info/about/ensembl_powered.html

⁶² As well as to align data such as RNA-seq data and any subsequent genome annotations

- B. Includes a rapid data transfer from the data management platform or the sharing platform to NCBI and/or ENA; and,
- C. Documentation on how to use the system (including a knowledgebase with community contributed content).

4.4 Implementation Timelines

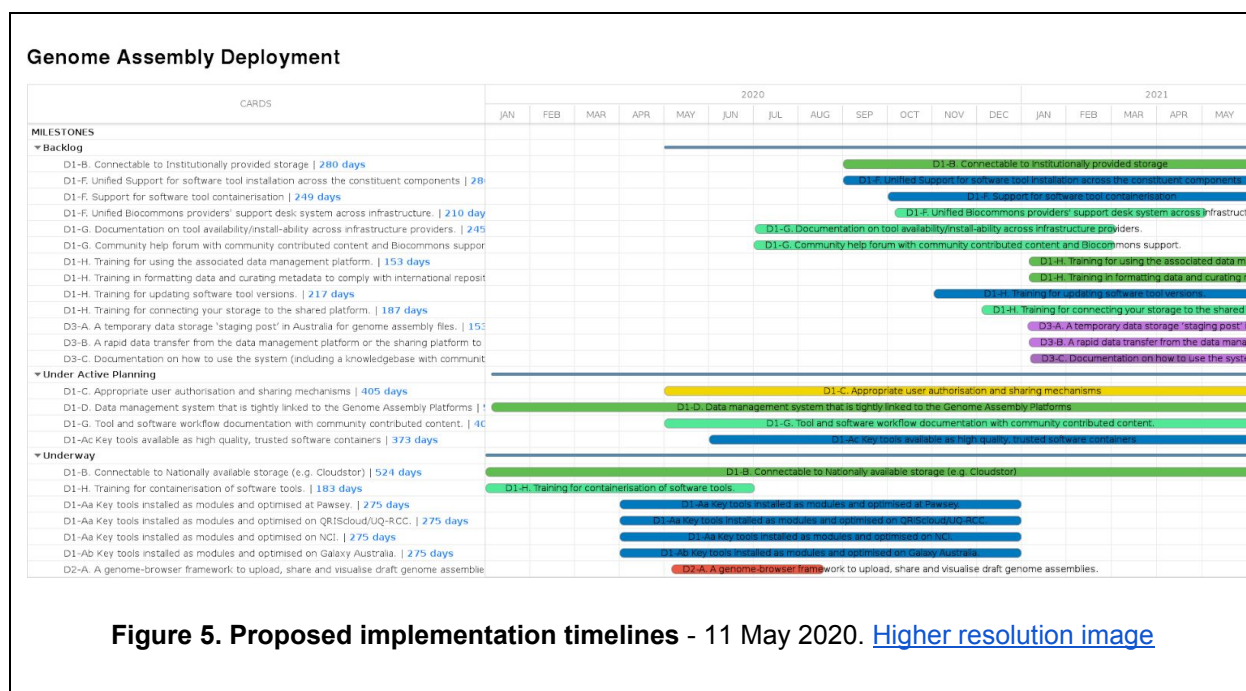


Figure 5. Proposed implementation timelines - 11 May 2020. [Higher resolution image](#)

As of May 2020, the following key activities are already underway:

Component	Planned date for delivery	Notes
D1-Aa Key tools installed as modules and optimised at Pawsey.	Ongoing	As of 11 May 2020, 9 of the genome assembly tools listed in Appendix 1 are installed as modules on Pawsey
D1-Aa Key tools installed as modules and optimised on QRIScloud/UQ-RCC.	Ongoing	As of 11 May 2020, 21 of the genome assembly tools listed in Appendix 1 are installed as modules on QRIScloud-UQ RCC
D1-Aa Key tools installed as modules and optimised on NCI.	Ongoing	As of 11 May 2020, 7 of the genome assembly tools listed in Appendix 1 are installed as modules on NCI
D1-Ab Key tools installed as modules and optimised on Galaxy Australia.	Ongoing	As of 11 May 2020, 12 of the genome assembly tools listed in Appendix 1 are installed as modules on Galaxy Australia
D1-B. Connectable to Nationally available storage (e.g. Cloudstor)	Ongoing	In 2019, work started to directly connect Cloudstor to Galaxy Australia. Streamlined connectivity of Cloudstor storage to Pawsey, QCIF, NCI compute resource will continue through the ARDC-funded "BioCommons Bring Your Own Data (BYOD) Expansion Project"

		https://www.biocommons.org.au/ardc-platforms-investment
D1-H. Training re. containerisation of software tools.	June/July 2020	Introductory level training (co-organised by BioCommons and Pawsey) will occur in June and July 2020. See https://www.biocommons.org.au/events/containers-intro
D2-A. A genome-browser framework to upload, share and visualise draft genome assemblies.	August 2020	QCIF is currently engaged by the BioCommons to establish a hosted Apollo/JBrowse-based service, and to explore the feasibility of establishing an Ensembl-empowered genome browser service.

As of May 2020, the following key activities are under active planning:

Component	Notes
D1-C. Appropriate user authorisation and sharing mechanisms	AAF is currently engaged by the BioCommons to explore Access and Authentication Frameworks that will be fit for purpose across all envisaged BioCommons-related platforms and services.
D1-D. Data management system that is tightly linked to the Genome Assembly Platforms	Considerations for what may be the best technical solution are ongoing. See Requirements of a Data Management Component of the Australian BioCommons
D1-G. Tool and software workflow documentation with community contributed content.	Tool and workflow documentation will be organised via the Australian BioCommons Github https://github.com/australianbiocommons
D1-Ac Key tools available as high quality, trusted software containers	This work will be enabled through the ARDC-funded "BioCommons Bring Your Own Data (BYOD) Expansion Project". https://www.biocommons.org.au/ardc-platforms-investment

Other components are planned for deployment from Q3 2020 and throughout 2021.

Appendix 1

Table 1. Genome assembly tools for consideration for inclusion in a shared analysis environment.

Note that a genome assembly protocol may also incorporate many other software tools not listed here.

Workflow Step	High-level component	Tool	Specific task / notes	Link to software
1	Quality Control	FastQC	Quality control tool for sequence data.	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
1	Quality Control	GenomeScope	Infers the global properties of a genome from unassembled sequenced data. GenomeScope can also be run on the Command Line. Note: a shiny app for batch running GenomeScope on multiple files and/or with multiple settings (like this http://qb.cshl.edu/genomescope/) is available (Richard Edwards, UNSW, pers comm)	https://github.com/schatzlab/genomescope
1	Quality Control	Jellyfish	Required as a preliminary step to GenomeScope. Jellyfish computes the histogram of k-mer frequencies	http://www.genome.umd.edu/jellyfish.html
1	Quality Control	Racon	Quality check of raw reads and assembled reads.	https://github.com/isovic/racon
1	Quality Control	seqstats	Seqstats is a small tool written in C on top of the klib library by Heng Li. It provides general summary statistics of both read and assembly files.	https://github.com/clwgg/seqstats
2	Trim and filter read sequences	BBTools	BBTools is a suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data.	https://jgi.doe.gov/data-and-tools/bbtools/
2	Trim and filter read sequences	BEDTools	Tools for a wide-range of genomics analysis tasks	https://github.com/arq5x/bedtools2
2	Trim and filter read sequences	Cutadapt	Detection and removal of adapter sequences, primers, polyA from HT sequencing reads	https://github.com/marcelm/cutadapt

2	Trim and filter read sequences	PRINSEQ	Filter and trim data.	http://prinseq.sourceforge.net/
2	Trim and filter read sequences	Trimmomatic	Assembly. Trims sequence data.	http://www.usadellab.org/cms/?page=trimmomatic
3	Sequence assembly	ABYSS	Assembles short and long read sequence data.	https://github.com/bcgsc/abyss
3	Sequence assembly	ALLMAPS	The ordering and orientation of genomic scaffolds to reconstruct chromosomes	https://github.com/tanghaibao/jcvi/wiki/ALLMAPS
3	Sequence assembly	ALLPATHS-LG	Whole genome shotgun assembler	http://software.broadinstitute.org/allpaths-lg/blog/
3	Sequence assembly	amos	A Modular, Open-Source whole genome assembler	http://amos.sourceforge.net/wiki/index.php/AMOS
3	Sequence assembly	Arrow	Arrow algorithm to map PacBio reads to get consensus and variant calls.	https://github.com/PacificBiosciences/GenomicConsensus
3	Sequence assembly	bbmap	Short read aligner	https://sourceforge.net/projects/bbmap/
3	Sequence assembly	bcftools	BCFtools is a set of utilities that manipulate variant calls in the Variant Call Format (VCF) and its binary counterpart BCF	http://samtools.github.io/bcftools/bcftools.html
3	Sequence assembly	Bionano hybrid scaffolding software		https://bionanogenomics.com/bionano-university/tags/hybrid-scaffolding/
3	Sequence assembly	Bowtie	Alignment of short reads for assembly	http://bowtie-bio.sourceforge.net/index.shtml
3	Sequence assembly	BWA	Alignment. Burrow-Wheeler Aligner for short-reads.	https://github.com/lh3/bwa
3	Sequence assembly	Cactus	whole genome multiple alignments	https://github.com/ComparativeGenomicsToolkit/cactus
3	Sequence assembly	Canu	Assembler for PacBio RS II/Sequel or Oxford Nanopore MinION data.	https://github.com/marbl/canu
3	Sequence assembly	cap3	Assemble sequences into contigs	http://seq.cs.iastate.edu/cap3.html

3	Sequence assembly	Cufflinks	RNAseq transcript assembly	http://cole-trapnell-lab.github.io/cufflinks/
3	Sequence assembly	edena	De novo short read assembler	http://www.genomic.ch/edena.php
3	Sequence assembly	Falcon	Assembly of PacBio sequence reads.	https://pb-falcon.readthedocs.io/en/latest/
3	Sequence assembly	Falcon Unzip	true diploid assembler - takes Falcon contigs	https://github.com/PacificBiosciences/pb-assembly
3	Sequence assembly	Falcon-Phase	maps HiC data to the FALCON-Unzip assembly	https://github.com/PacificBiosciences/pb-assembly
3	Sequence assembly	Flye	Assembly. De novo assembly for single molecule sequencing.	https://github.com/fenderglass/Flye
3	Sequence assembly	GATK: Genome Analysis Tool Kit	Variant identification for assembly.	https://software.broadinstitute.org/gatk/documentation/quickstart
3	Sequence assembly	HiRise	Assembly of scaffolds using Dovetail/Chicago	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4772016/
3	Sequence assembly	HMMER v3.2	Blasting Illumina reads	http://hmmer.org/
3	Sequence assembly	idba-ud	de Bruijn graph assembler	https://github.com/loneknightpy/idba
2	Sequence assembly	LINKS	LINKS is a genomics application for scaffolding genome assemblies with long reads, such as those produced by Oxford Nanopore Technologies Ltd.	https://github.com/bcgsc/LINKS
3	Sequence assembly	MAFFT	Alignment. Align multiple sequences.	https://mafft.cbrc.jp/alignment/software/
3	Sequence assembly	MaSuRCA	Assembly. Hybrid platforms Illumina, PacBio, MinION	https://github.com/alekseyezimin/masurca
3	Sequence assembly	Megahit	Ultra-fast and memory-efficient (meta-)genome assembler	https://github.com/voutcn/megahit
3	Sequence assembly	metabat	Metagenome assembly framework	https://bitbucket.org/berkeleylab/metabat/src/master/
3	Sequence assembly	minia	Minia is a short-read assembler based on a de Bruijn graph, capable of assembling a human genome on a desktop computer in a day.	https://github.com/GATB/minia

3	Sequence assembly	Minimap2	Alignment. Pairwise aligner.	https://github.com/lh3/minimap2
3	Sequence assembly	MIRA	Plastid assemblies.	http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html
3	Sequence assembly	MUMmer	MUMmer is a versatile alignment tool for DNA and protein sequences.	https://github.com/mummer4/mummer
3	Sequence assembly	Muscle	Alignment. Align multiple sequences.	https://www.ebi.ac.uk/Tools/msa/muscle/
3	Sequence assembly	nextgenmap	Read mapping program	http://cibiv.github.io/NextGenMap/
3	Sequence assembly	oases	De novo transcriptome assembler for very short reads	https://www.ebi.ac.uk/~zerbino/oases/
3	Sequence assembly	pb-assembly (falcon)	Falcon assembly pipeline	https://github.com/PacificBiosciences/pb-assembly
3	Sequence assembly	Pilon	Improves assemblies, detects variation among strains.	https://github.com/broadinstitute/pilon/wiki
3	Sequence assembly	Plink v1.9	Illumina	https://www.cog-genomics.org/plink2
3	Sequence assembly	PurgeHaplotigs	Pipeline to help with curating heterozygous diploid genome assemblies (for instance when assembling using FALCON or FALCON-unzip).	https://github.com/skinngan/purge_haplotigs_multiBAM
3	Sequence assembly	pyrad	Analyzing RADseq-like data	http://dereneaton.com/software/pyrad/
3	Sequence assembly	Quickmerge	quickmerge uses a simple concept to improve contiguity of genome assemblies based on long molecule sequences, often with dramatic outcomes	https://github.com/mahulichak/quickmerge
3	Sequence assembly	SAMTools	SNP calling. Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format	http://www.htslib.org/
3	Sequence assembly	scarpa	Stand-alone scaffolding tool for NGS data	http://compbio.cs.toronto.edu/hapsembler/scarpa.html
3	Sequence assembly	SMARTdenovo	SMARTdenovo is a de novo assembler for PacBio and Oxford Nanopore (ONT) data.	https://github.com/ruanjue/smartdenovo

3	Sequence assembly	smrtlink (pacbio)	Analysis platform for assembly, calling variants, discovering RNA transcripts, detecting epigenetic modifications	https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/
3	Sequence assembly	SOAPdenovo	Assembles short and long read sequence data.	https://github.com/aquaskeyline/SOAPdenovo2
3	Sequence assembly	SOAPSnp	Re-sequencing utility	https://sourceforge.net/projects/soapsnp/
3	Sequence assembly	SPAdes	paired-end reads, mate-pairs and single (unpaired) reads	https://github.com/ablab/spades
3	Sequence assembly	Stacks	Software pipeline for De novo assembly & SNP calling, building loci from short-read sequences	http://catchenlab.life.illinois.edu/stacks/
3	Sequence assembly	Supernova	Assembly. De novo assembly from chromium-linked reads.	https://support.10xgenomics.com/de-novo-assembly/software/overview/latest/welcome
3	Sequence assembly	Trinity	RNAseq de novo assembly.	https://github.com/trinityrnaseq/trinityrnaseq/wiki
3	Sequence assembly	Unicycler	Unicycler is an assembly pipeline for bacterial genomes. It can assemble Illumina-only read sets where it functions as a SPAdes-optimiser.	https://github.com/rwwick/Unicycler
3	Sequence assembly	VCFTools	Variant identification for assembly.	https://vcftools.github.io/index.html
3	Sequence assembly	Velvet	De novo genome assembler specially designed for short read sequencing technologies, such as Solexa or 454	https://www.ebi.ac.uk/~zerbino/velvet/
3	Sequence assembly	vsearch	De novo and reference based chimera detection, clustering, full-length and prefix dereplication, rereplication, reverse complementation, masking, all-vs-all pairwise global alignment, exact and global alignment searching, shuffling, subsampling and sorting	https://github.com/torognes/vsearch
3	Sequence assembly	WTDBG2	Assembly. Fuzzy approach to long noisy reads.	https://github.com/ruanjue/wtdbg2
4	Assembly validation	Blobtools	Modular command-line solution for visualisation, quality control and taxonomic partitioning of genome datasets	https://github.com/DRL/blobtools
4	Assembly validation	Busco	Assessing assemblies with single copy orthologs.	https://busco.ezlab.org/

4	Assembly validation	Nanopolish	Software package for signal-level analysis of Oxford Nanopore sequencing data. Nanopolish can calculate an improved consensus sequence for a draft genome assembly, detect base modifications, call SNPs and indels with respect to a reference genome and more	https://github.com/jts/nanopolish
5	Other	Nextflow	Workflow tool	https://www.nextflow.io/
5	Other	Snakemake	Workflow management	https://snakemake.readthedocs.io/en/stable/index.html

A complete list of tools with more details is available [here](#) (see Tool Level Table ASSEMBLY tab).

Appendix 2

Survey⁶³ questions posed to the Research Community

- ☐ Email address
- ☐ What's your name?
- ☐ How would you describe your level of experience with genome assembly?
 - ☐ Very experienced
 - ☐ Some experience
 - ☐ Beginner
 - ☐ Interested but no direct experience
 - ☐ Other:
- ☐ Which part(s) of the genome assembly process do you / group members perform, or envisage performing in the next 5 years?
 - ☐ Collecting sample material for nucleic acid sequencing, e.g. field work or wet lab DNA extraction work
 - ☐ Nucleic acid sequencing, e.g. running sequencing machine
 - ☐ Check of sequence quality
 - ☐ Trim and quality filter sequence reads
 - ☐ Assemble sequences into contigs
 - ☐ Assemble contigs into scaffolds
 - ☐ Use genomes after assembly by someone else
 - ☐ Other:
- ☐ With respect to genome assembly, which organism(s) have you worked on, or will work on in the next 5 years?
- ☐ Which (if any) tools do you / group members use? Please only indicate those you'd currently recommend for use (i.e. no deprecated examples)
 - ☐ FastQC
 - ☐ Trimmomatic

⁶³ [Genome Assembly Poll/Survey](#)

- ☐ SPADES
- ☐ Flye
- ☐ Canu
- ☐ MaSuRCa
- ☐ Falcon
- ☐ Supernova
- ☐ SamTools
- ☐ BWA
- ☐ WTDBG2
- ☐ Pilon
- ☐ ABySS
- ☐ HiRise
- ☐ SOAPdenovo
- ☐ Stacks
- ☐ Racon
- ☐ GATK
- ☐ VCFTools
- ☐ Busco
- ☐ Trinity
- ☐ Custom tool developed in our group or by collaborators
- ☐ Other:
- ☐ Are there pipelines or tools you'd like to use but that aren't suitable for your study taxon? If so, what are they and why aren't they suitable?
- ☐ Are there pipelines or tools you'd like to use but can't because of technical limitations (e.g. installation, compute requirements, dataset access requirements)? If so, what are the tools and what are the roadblocks you've encountered? What is your workaround and why is it inadequate?

- ☐ Do you require custom or proprietary tools for your genome assembly approach? If so, what are they?
- ☐ What sequencing platform/s are you currently using to generate data?
 - ☐ Illumina
 - ☐ PacBio
 - ☐ 10 X
 - ☐ Nanopore
 - ☐ Ion Torrent
 - ☐ Other:
- ☐ Do you make use of existing datasets from the same taxon or closely related taxa?
 - ☐ Yes, public datasets from the same taxon
 - ☐ Yes, private datasets from the same taxon (from my previous work or that of collaborators)
 - ☐ Yes, public datasets from closely related taxa
 - ☐ Yes, private datasets from closely related taxa (from my previous work or that of collaborators)
 - ☐ No, because no relevant data exists from my taxon or a closely-related taxon
 - ☐ No - some data exists but it's too low quality for this purpose
 - ☐ No - some data exists but it's too difficult to integrate because of poor/outdated format or metadata
 - ☐ No - some data exists but it's too difficult to integrate because of a lack of suitable tools/pipelines
 - ☐ No - some private data exists but I can't access it
 - ☐ Other:
- ☐ Do you use a data management tool/framework within your genome assembly project(s)? If so, what?
- ☐ How do you share data within your group and with collaborators? Where are your collaborators based? What difficulties have you encountered?
- ☐ Do you make your assemblies publicly available? If so, where? Have you encountered any difficulties in doing so?

- ☐ If you don't make your genome assembly results publicly available, why not?
 - ☐ Commercial confidence issues
 - ☐ I don't see a benefit in sharing my genome assembly results publicly
 - ☐ I don't know how to make my genome assembly results publicly available
 - ☐ Other:
- ☐ What kind of compute infrastructure setup do you use for genome assembly?
 - ☐ Local desktop/PC
 - ☐ High-performance computing at my institution
 - ☐ High-performance computing at a collaborator's institution
 - ☐ State or national high-performance computing infrastructure (e.g. NCI, Pawsey, QCIF)
 - ☐ NeCTAR cloud instance
 - ☐ Other research cloud resources (e.g. Amazon)
 - ☐ Cyverse
 - ☐ Galaxy
 - ☐ Other:
- ☐ Do you have access to the expertise you need to build and maintain this compute infrastructure (e.g. installing and updating software)?
 - ☐ Yes, within our group
 - ☐ Yes, via collaborators
 - ☐ Yes, within our institution
 - ☐ Yes, via partner high-performance computing infrastructure (e.g. NCI, Pawsey, QCIF)
 - ☐ No, we would like to set some up or update our current approach but can't access expertise
 - ☐ Other:
- ☐ Is your current compute infrastructure sufficient for your current needs? If not, why not?
- ☐ Will this compute infrastructure setup be sufficient for your needs in 2 years' time?

- ☐ Yes, we expect to be doing genome assembly at a similar scale in 2 years' time
 - ☐ Yes, as we expect to be doing less genome assembly in 2 years' time
 - ☐ No, we expect to be doing more genome assembly by then and will need more resources
 - ☐ No, this infrastructure will be shut down or deprecated by then and we need to find a replacement
 - ☐ No, the responsible lab member will be moving on by then and we will need an alternative
 - ☐ I don't know
 - ☐ Other:
- ☐ Will this compute infrastructure setup be sufficient for your needs in 5 years' time?
- ☐ Yes, we expect to be doing genome assembly at a similar scale in 5 years' time
 - ☐ Yes, as we expect to be doing less genome assembly in 5 years' time
 - ☐ No, we expect to be doing more genome assembly by then and will need more resources
 - ☐ No, this infrastructure will be shut down or deprecated by then and we need to find a replacement
 - ☐ No, the responsible lab member will be moving on by then and we will need an alternative
 - ☐ I don't know
 - ☐ Other:
- ☐ Would you / group members use a shared compute infrastructure platform to perform genome assembly?
- ☐ Yes - we can't currently perform genome assembly without such a platform
 - ☐ Yes - our needs are currently met but we'd consider a shared platform if it was suitable
 - ☐ No - we will always prefer to perform genome assembly locally no matter how good a shared platform is
 - ☐ Other:

- ☐ How important are these general factors to you in a shared genome assembly platform?
[Following best practice in tools, formats and metadata; compliant with requirements of international data repositories]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these general factors to you in a shared genome assembly platform?
[Free (subsidised) to researchers no matter the scale of analysis]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these general factors to you in a shared genome assembly platform?
[Easy to access from anywhere]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these general factors to you in a shared genome assembly platform?
[Easy to self-manage access and permissions for collaborators]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these general factors to you in a shared genome assembly platform?
[Easy to upload/download data]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these general factors to you in a shared genome assembly platform?
[Security of data and analysis]
 - ☐ Crucial

- ☐ Important
- ☐ Unimportant
- ☐ How important are these general factors to you in a shared genome assembly platform? [Long-term support for and sustainability of the platform]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these data-related factors to you in a shared genome assembly platform? [Smart metadata handling (e.g. assistance with metadata formats, transfer of metadata through pipeline, controlled vocabulary lookup)]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these data-related factors to you in a shared genome assembly platform? [Ability to submit datasets to international repositories from the platform]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these data-related factors to you in a shared genome assembly platform? [Ability to download datasets from international repositories within the platform]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these data-related factors to you in a shared genome assembly platform? [Ability to transfer data easily to/from storage]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant

- ☐ How important are these training-related factors to you in a shared genome assembly platform? [Good documentation on how to use the platform]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these training-related factors to you in a shared genome assembly platform? [Good documentation on how to use the tools/pipelines]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these training-related factors to you in a shared genome assembly platform? [Access to in-person training on how to use the platform]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these training-related factors to you in a shared genome assembly platform? [Access to in-person training on how to use the tools/pipelines]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these training-related factors to you in a shared genome assembly platform? [Discussion forum to share expertise with other users]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these tool/pipeline-related factors to you in a shared genome assembly platform? [Access to our preferred tools/pipelines]
 - ☐ Crucial
 - ☐ Important

- ☐ Unimportant
- ☐ How important are these tool/pipeline-related factors to you in a shared genome assembly platform? [Access to a choice of tools/pipelines]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these tool/pipeline-related factors to you in a shared genome assembly platform? [Quick installation of other tools/pipelines upon request]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these tool/pipeline-related factors to you in a shared genome assembly platform? [Assistance available in implementing pipelines]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ What are the top 1-5 tools/pipelines you would absolutely require in a shared genome assembly platform?
- ☐ How important are these compute-related factors to you in a shared genome assembly platform? [Easy access to sufficient compute resources for our analysis]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these compute-related factors to you in a shared genome assembly platform? [Ability to scale up/down resources used as needed]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant

- ☐ How important are these compute-related factors to you in a shared genome assembly platform? [No need to understand or control the compute backend]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ How important are these compute-related factors to you in a shared genome assembly platform? [Compatibility with external analysis environments (e.g. Amazon, Cyverse)]
 - ☐ Crucial
 - ☐ Important
 - ☐ Unimportant
- ☐ Are there any other factors you consider crucial in a shared genome assembly platform? If so, what?
- ☐ Is there anything else you'd like to tell us or suggest? Please let us know here.

Document Control

VERSION	DATE	AUTHOR(S)	DESCRIPTION
V1.0	19/02/2020	Tiffanie Nelson	A preliminary document detailing the outline of the roadmap draft including software list obtained from researchers.
V1.1	11/05/2020	Tiffanie Nelson, Jeff Christiansen	Draft document defined.
V1.2	15/05/2020	Tiffanie Nelson, Jeff Christiansen	Draft ready for distribution to the community.
v4.0	31/07/2020	Tiffanie Nelson, Jeff Christiansen	Modifications made based on feedback from Genome Annotation SIG and Australian computational infrastructure providers