Comparative Genomics Infrastructure Roadmap for Australia

V4.0

05 September 2022

Tiffanie M Nelson and Jeffrey H Christiansen

Contents

1. Executive Summary

Comparative genomics is defined here to include phylogenetics, phylogenomics and pangenomics. The comparative analyses of genomes or genome components are steps on a workflow to understand common ancestors or conserved genes in phylogenomics or phylogenetics, or to identify the complete genetic repertoire in the entire species in pangenomics.

Different questions can be addressed through the comparison of genomes or genetic sequences, such as identifying broad insights into gene types and their evolution, or pinpointing genes that are related to a particular biological system. These developments have the potential to translate into novel approaches for treating disease in humans, animals and plants, conserving rare and endangered species or advancing the economic value and sustainability of agricultural species. In Australia, comparative genomics is conducted by researchers across a wide variety of life science domains.

This document includes:

- a brief summary of comparative genomics tools and methodologies,
- how the Australian community currently undertakes this work and their common data-, software- or compute-related infrastructure challenges (information obtained through consultation with a 'Special Interest Group' (SIG) of researchers undertaking various comparative genomic approaches across Australia), and
- a high-level description of key components of an envisaged shared national comparative genomic analysis infrastructure for Australia, which, when implemented, would enable Australian researchers from a wide range of institutions to perform comparative genomic work they would otherwise be unable to undertake because of the reported roadblocks, i.e.

  D1. A platform for performing comparative genomics analyses: to provide all Australian researchers with access to a shared platform with tools and workflows for comparative genomics analysis, including phylogenetics, phylogenomics and pangenomics, 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 visualisation and statistical analysis for comparative genomics, phylogenetics, phylogenomics and pangenomics data and their data products: to make it easier for Australian researchers to perform relevant statistical and/or visualisation-based analyses of phylogenetic, phylogenomic and pangenomic data.

  D3. Systems to enable submission of raw sequencing reads, phylogenetic/phylogenomic trees and pangenome files from Australia to appropriate repositories: to make it easier for any Australian researcher to share and publish their phylogenetic, phylogenomic or pangenomic data files publicly and in accordance with best-practice open science guidelines.

Feedback on the proposed components outlined in this initial draft plan is now sought from the SIG and any other Australian researchers undertaking comparative genomics analyses. Following engagement with other stakeholder groups (i.e. international entities operating comparative genomics infrastructure

elsewhere and Australian research IT infrastructure partners), further iterations of this document will be produced with a final version of the plan scheduled for February 2022.

2. Background and Context

In Australia, investments to establish community-scale bioinformatics infrastructure to support life science research have materialised in various forms and scales over the last decade under a range of funding schemes. One significant supporter is Bioplatforms Australia[1], which aims to develop and support Australia's national bioinformatics infrastructure and is funded under the National Collaborative Research Infrastructure Strategy (NCRIS)[2].

Since 2019, Bioplatforms Australia has supported the Australian BioCommons[3], which is an initiative focussed on establishing improved access to bioinformatics tools, methods, datasets, computational infrastructure, along with training and support 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 infrastructure to support bioinformatics-based research and resources that are relevant to these research communities.

To support the large (and growing) community of practice in Australia undertaking comparative genomics, in late May 2021, the Australian BioCommons convened a "Comparative Genomics Special Interest Group (SIG)" and invited participation from over 100 researchers across Australia with either experience in, or interest in comparative genomics[4].

The outcome of the survey and that meeting is this document, which summarises and represents the current or expected infrastructure roadblocks and challenges described by members of the community, and identifies 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 the nomination of additional members of the SIG, by either adding comments directly to this google document or by emailing communities@biocommons.org.au

Feedback on the proposed components outlined in this initial draft plan is now sought from the SIG and any other Australian researchers or their collaborators undertaking comparative genomics, inclusive of phylogenomics, pangenomics or molecular phylogenetics.

Following engagement with other stakeholder groups (i.e. international entities operating comparative genomics infrastructure elsewhere and Australian research IT infrastructure

[1] Bioplatforms Australia
[2] National Collaborative Infrastructure Strategy (NCRIS)
[3] Australian BioCommons
[4] see Section 3.2 for methodology employed for formation of the group and membership

partners), further iterations of this document will be produced with a final version of the plan scheduled for February 2022.

3. Comparative Genomics, Phylogenomics and Pangenomics - Methods and Community

3.1 What is comparative genomics and how is it done, and why?

We use the term ‘comparative genomics’ as a catchall for a number of techniques including molecular phylogenetics, phylogenomics, and pangenomics. For the purpose of this community consultation (and therefore within this document), we define comparative genomics as a practice:

- Where two or more sets of gene sequences or whole or partial genome sequences from multiple organisms are aligned, compared and analysed for relatedness, and
- Where this is done:
    - for the purpose of understanding the evolutionary lineage,
    - to infer putative functions for DNA or protein sequences,
    - to identify common genetic/genomic features between organisms, or
    - to identify genetic differences among and within a species[5].

Prior to the 1980s, phylogenetic analysis (i.e. reconstructing the evolutionary history of life on Earth) was primarily conducted through the comparison of anatomical (i.e. morphological) structures and features between individuals and/or organisms. With the introduction of nucleic acid sequencing technologies in the 1980s, short stretches of molecular genetic information was incorporated into phylogenetic analysis approaches which transformed the methodology and speed at which such phylogenetic studies can be undertaken[6].

[5] National Human Genome Research Institute: Comparative Genomics Fact Sheet genome.gov/about-genomics/fact-sheets/Comparative-Genomics-Fact-Sheet
[6] Young and Gilung, 2019, Systematic Entomology, onlinelibrary.wiley.com/doi/10.1111/syen.12406

The advancement of affordable ‘next generation’ sequencing technologies since the mid-2000s has similarly transformed the field of phylogenetics by enabling a comparative genomics approach to be employed which allows for a much greater richness of information to be fed into an analysis. The ability to readily obtain genome-scale data from multiple samples[7] and their equivalent part- or whole- assembled genomes[8] has resulted in an explosion of sequence data generated by large and small research groups relating to a wide range of model and non-model species[6]. Increasingly powerful methods for sequence-based approaches capable of drawing more accurate inferences (in part by using more data per species) have been developed along with faster or lower memory methods that can process more data allowing for phylogenetic analyses from a wider range of species, including those previously inaccessible due to range, endangerment or extinction.

With respect to the three broad approaches we classify as ‘comparative genomics’: **molecular phylogenetics** is focused on the generation of a phylogenetic tree from relatedness calculations based on a single gene or region across a range of species within a group. **Phylogenomics** is an extension of this approach whereby the inclusion of the entire genome (or very large regions of it, e.g. chromosomes etc) are used to inform the construction of a phylogenetic tree.

Input data (an assembled genome or gene regions) for molecular phylogenetic or phylogenomic tree construction are assessed for homology or relationship to a common ancestor[6]. Within homology assessment, detecting orthologs (i.e. similar genes that are a result of speciation, and that originated by vertical descent from a single gene of the last common ancestor), is conducted via methodologies that rely on pairwise comparison of available sequences[9]. The order of homologous genes on chromosomes, known as synteny, further provides insight into the relationship between organisms and their linkage to a common ancestor.

Increasing the number of samples and the range of genetic variation in an analysis can be used to improve the accuracy of a phylogenetic tree[10]. Large pairwise comparisons can take a significant amount of compute, for example, a set of ~100 taxa with ~1,000 genes in transcriptomes can take more >60 h runtime on multi-core HPCs without seeing either a significant decrease in tree robustness or complete failure to run depending on the choice of tool[11]. As increasing numbers of sequences become available, the comparison and complexity of alignments have grown quadratically (i.e. proportional to the square of the function argument) and as a result, phylogenomic and comparative genomic analyses have developed a large computational time burden[12].

[7] Mardis, 2011, Nature, nature.com/articles/nature09796
[8] See also: Nelson and Christiansen (2020). Genome Assembly Infrastructure Roadmap for Australia (4.0). Zenodo. doi.org/10.5281/zenodo.3967970
[9] Nichio, et al, 2017, Frontiers in Genetics, doi.org/10.3389/fgene.2017.00165
[10] Zwickl and Hillis, 2002, Systematic Biology, 10.1080/10635150290102339
[11] Zhou, et al, 2018, Molecular Biology and Evolution,doi.org/10.1093/molbev/msx302
[12] Dieckmann, et al, 2021, Nucleic Acid Research, https://doi.org/10.1093/nar/gkab341

Pangenomics (i.e. the construction of a pangenome for a species - i.e. the core genes/sequences found in all individuals of that species as well as accessory genes/sequences found in some individuals only) by contrast, requires a whole-genome assembly and high-quality annotation and can not rely on smaller targeted regions of the genome. Construction of an organism's pangenome includes genomes from multiple individuals within a species to understand the complete repertoire of genes available to a species. Previously a single reference genome from a single or a few individuals was used as an indication of the standard genome, known as the 'reference' for the species[13]. However, focussing on the reference genome based on only a few individuals limits the ability to identify the diversity of genomic-based traits across a species[14]. The pangenome is made up of 'core' genes and 'accessory' genes, also known as variable or dispensable. Core genes are those found in all individuals, where accessory genes are absent from one or more individuals[15].

Pangenome construction requires the selection of genotypes from genetically diverse populations. The inference of a species' 'core' and 'variable' genes requires the pangenome is created from multiple, diverse individuals [16]. Long-read sequencing greatly improves the ability to assemble large, complex genomes, and particularly complex genomes. With the availability of technology that produces long sequence reads, it is now commonplace for a research group to produce 10 genomes from genomically complex species or 100s of genomes from genomically simpler species over the course of a year[15]. Many pangenomes to date, with the exception of humans[17,18], have been created on less than 50 representatives from the species, for example, barley (n = 20)[19] wheat (n = 15)[20], pig (n = 12)[21], soybean (n = 26)[22]. Core and variable regions of the pangenome are identified through various toolkits and visualisation techniques with considerably different methods required for more complex genomes (and therefore pangenomes), such as plants, compared to organisms with relatively simple and small genomes, such as bacteria. In plant and animal genomes, a large portion of genetic diversity is present in intergenic sequences and derived from transposable elements or genes that move around, or regions that are repetitive regions, hence establishing what is shared versus unique can be difficult to determine[17].

High-level conceptual workflows showing the general steps that are required to generate a phylogenetic tree, or construct a pangenome are shown in Figure 1.

[13] Morneau, 2021, Nature Milestones, nature.com/articles/d42859-020-00115-3
[14] Golicz, et al, 2020, Trends in Genetics, pubmed.ncbi.nlm.nih.gov/31882191/
[15] Golicz, et al, 2020, Trends in Genetics, pubmed.ncbi.nlm.nih.gov/31882191/
[16] Jayakodi, et al, 2021, DNA Research, https://doi.org/10.1093/dnares/dsaa030
[17] Sherman, et al, 2018, Nature, nature.com/articles/s41588-018-0273-y
[18] Duan, et al, 2019, Genome Biology, genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1751-y
[19] Jayakodi, et al, 2020, Nature, nature.com/articles/s41586-020-2947-8?faodatalab=2020-11-26-1
[20] Walkowiak, et al, 2020, Nature, nature.com/articles/s41586-020-2961-x
[21] Tian, et al, 2020, Science China Life Sciences, 10.1007/s11427-019-9551-7
[22] Liu, et al, 2020, Cell, doi.org/10.1016/j.cell.2020.05.023

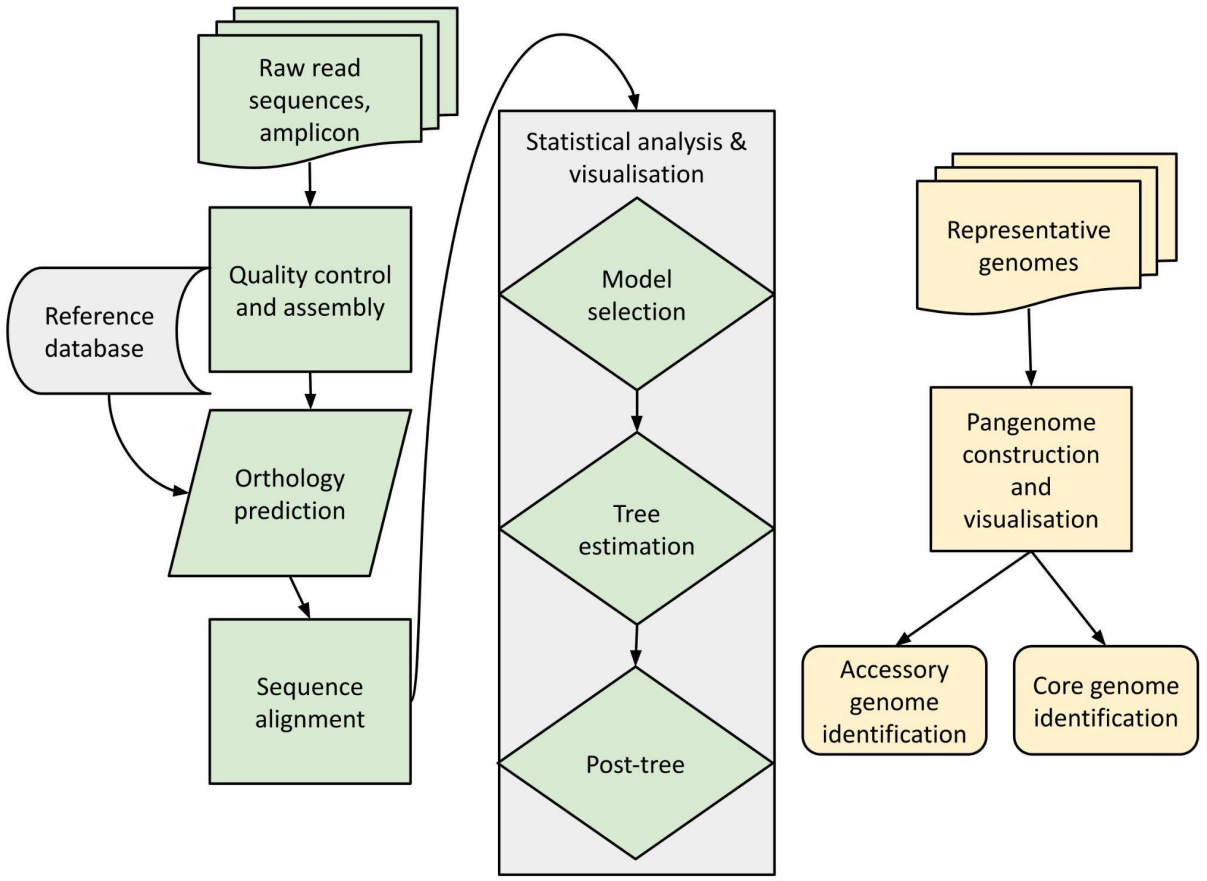

Figure 1: A general phylogenomics and pangenomics workflow

The phylogenetic tree workflow is shown on the left in green and the pangenome construction workflow on the right in yellow. The workflow displays the dominant steps used to transfer raw sequence reads from a whole-genome sequencing project into an assembled genome with appropriate quality control and evaluation prior to identifying if homologous genes are orthologs[23]. Following ortholog detection, sequences are aligned prior to model selection and phylogeny generation. This workflow was adapted from information detailed in Young and Gillung, 2019[24].

3.2 Who in Australia is performing comparative genomics analysis, and which species are they tackling?

The reduction in nucleic acid sequencing cost and greater data availability has reduced the barrier for many researchers to carry out comparative genomics and related projects. These advances have greatly progressed the global understanding of the evolutionary Tree of Life, including extant and extinct organisms[25]. The benefits of next generation sequencing have a

[23] Phylogenetic relationships should be assessed with the identification of sequences related by orthology, i.e. whose common ancestor diverged as a result of speciation. See the full discussion in Nichio et al, 2017, Frontiers in Genetics, frontiersin.org/articles/10.3389/fgene.2017.00165/full

[24] Young and Gillung, 2019, Systematic Entomology, onlinelibrary.wiley.com/doi/full/10.1111/syen.12406

[25] Hug et al, 2016, Nature Microbiology, nature.com/articles/nmicrobiol201648

far-reaching impact on the environment, agriculture, and health.

Broad uses of comparative genomics approaches include classification of species, tracking disease outbreaks or identifying the origin of pathogens, informing conservation policy when predicting and preventing extinction and in forensic science for solving crimes and identifying parentage[26]. Specific examples of the benefits in pangenome construction include finding 'core' genes in pangenomes that are indispensable to an organism's survival and can be a promising target for the development of medical or agricultural treatments[27],[28] or finding 'variable' accessory genes that are associated with virulence/defense responses[29]. When conducting comparative sequence analyses in phylogenomic and molecular phylogenetics, some specific examples include tracking virus mutations in disease outbreaks or pandemics[30],[31], developing a greater understanding of basic biological processes[32] such as photosynthesis[33] and elucidating host-parasite interactions[34].

Hence, the critical importance of comparative genomics and the generation of phylogenetic trees or pangenomes is a natural progression from the production of sequencing data and associated data products, such as genomes. Supporting the generation of genomic data will help to address challenges of strategic importance to Australia, and as such is touched on in several Australian Academy of Science Decadal Plans for Science[35], including in the plans for Biodiversity[36], Agricultural Science[37], Marine Science[38], Ecoscience[39], and Nutrition Science[40]. Undertaking comparative genomics analyses 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 sequencing, as well as readily available data, compute and tools to enable access to previously sequenced genomic data or data products have facilitated comparative genomic techniques becoming somewhat routine and accessible. There are many groups and consortia across Australia who are actively working on comparative genomics projects with the goals of creating phylogenies and pangenomes, with a general focus on Australian ecosystems. These include:

---

[26] Grueber, 2015, Computational and Structural Biotechnology Journal, ncbi.nlm.nih.gov/pmc/articles/PMC4475778/
[27] Young and Gillung, 2019, Systematic Entomology, onlinelibrary.wiley.com/doi/full/10.1111/syen.12406
[28] Poulsen, 2019, Proceedings of the National Academy of Science, pnas.org/content/116/20/10072
[29] Golicz, et al, 2020, Trends in Genetics, pubmed.ncbi.nlm.nih.gov/31882191/
[30] Seeman, et al, 2020, Nature Communications, nature.com/articles/s41467-020-18314-x
[31] Jungreis, et al, 2021, Nature Communications, nature.com/articles/s41467-021-22905-7
[32] Young and Gillung, 2019, Systematic Entomology, onlinelibrary.wiley.com/doi/full/10.1111/syen.12406
[33] Rubin, et al, 2015, Proceedings of the National Academy of Science, /pubmed.ncbi.nlm.nih.gov/26508635/
[34] Foth, et al, 2014, Nature Genetics, nature.com/articles/ng.3010
[35] 10-year strategic plans for science disciplines, developed by the Australian Academy of Science's National Committees for Science.
[36] science.org.au/support/analysis/decadal-plans-science/discovering-biodiversity-decadal-plan-taxonomy
[37] science.org.au/support/analysis/decadal-plans-science/decadal-plan-agricultural-sciences-2017-2026
[38] science.org.au/support/analysis/reports/national-marine-science-plan
[39] science.org.au/support/analysis/reports/foundations-future-long-term-plan-australian-ecosystem-science
[40] science.org.au/supporting-science/science-policy-and-analysis/decadal-plans-science/nourishing-australia-decadal-plan

- Australian Centre for Ecogenomics (ACE) Genome Taxonomy Database which houses >258,000 genomes from Archaea and Bacteria[41] with a goal of providing a standardised microbial taxonomy[42].

- The Genomics for Australian Plants (GAP) consortium has the goal of resolving the Australian Angiosperm Tree of Life to genus level (Phase 1) using a set of molecular markers. The consortium is also working on Phase 2 to generate datasets with denser sampling within genera. A longer term goal of resolving the Tree of Life down to species level (Phase 3) is outside the current GAP project and requires additional investment[43].

- Oz Mammals Genomics consortium which aims to resolve the phylogeny of all mammals native to the Australo-Papuan region (around 500 species)[44].

- Australian Amphibian and Reptile Genomics (AusARG) consortium who are currently generating exon capture data for orthologous genes from all reptiles and amphibians native to the Australo-Papuan region to produce highly resolved phylogenies[45].

- Taxonomy Australia[46] which has an objective to implement the recommendations of the Australian Academy of Sciences' decadal plan for taxonomy and biosystematics[37].

- Atlas of Living Australia[47] which is a digital infrastructure platform that aggregates numerous data types including species occurrence data and their taxonomy integrated with spatial and environmental data.

[41] Parks, et al, 2021, Nucleic Acids Research, doi.org/10.1093/nar/gkab776
[42] Genome Taxonomy Database, gtdb.ecogenomic.org/
[43] Genomics for Australian Plants Phylogenomics, genomicsforaustralianplants.com/phylogenomics/
[44] Oz Mammals Genomics Phylogenomics, ozmammalsgenomics.com/phylogenomics/
[45] Australian Amphibian and Reptile Genomics Phylogenomics, ausargenomics.com/phylogenomics/
[46] Taxonomy Australia, taxonomyaustralia.org.au/home-static
[47] Atlas of Living Australia, ala.org.au/

There are also a number of laboratory groups and divisions within institutes and universities pursuing molecular evolutionary projects[48,49,50,51,52], as well as others who focus on human comparative genomics[53,54,55] including those specifically related to disease proliferation, immunity or treatment and forensic sciences[56,57,58,59,60,61].

The scientific literature indicates an approximate number of studies using comparative genomics, phylogenomics, phylogenetics (molecular) and pangenomics produced from Australian-based researchers (see Figure 2).

[48] Moritz Group - Evolutionary Biogeography and Conservation, Australian National University, biology.anu.edu.au/research/groups/moritz-group-evolutionary-biogeography-conservation
[49] Lanfear Group - Mutation, molecular Evolution and Phylogenetics, Australian National University, biology.anu.edu.au/research/groups/lanfear-group-mutation-molecular-evolution-and-phylogenetics
[50] Molecular Ecology, Evolution and Phylogenetics (MEEP) Lab, University of Sydney, meep.sydney.edu.au/
[51] Centre for Conservation Ecology and Genomics, canberra.edu.au/research/faculty-research-centres/cceg
[52] Centre for Australian National Biodiversity Research and Australian National Herbarium, https://www.cpbr.gov.au/cpbr/
[53] Garvan Institute of Medical Research - Human Comparative Genomics, garvan.org.au/research/genomics-epigenetics/human-comparative-and-prostate-cancer-genomics/human-comparative-genomics
[54] Australian Genomics, australiangenomics.org.au/
[55] Victor Chang Cardiac Research Institute, victorchang.edu.au/about-us/our-scientists/dr-emily-wong
[56] LanLab - Evolutionary Microbiology, University of New South Wales, lanlab.unsw.edu.au/
[57] Queensland Health's Forensic and Scientific Services, Queensland Government, health.qld.gov.au/public-health/forensic-and-scientific-services
[58] Kathy Belov Lab Group - Comparative Genomics, University of Sydney, sydney.edu.au/science/about/our-people/academic-staff/kathy-belov
[59] Institute for Molecular Bioscience, University of Queensland, imb.uq.edu.au/about
[60] QIMR Berghofer Medical Research Institute, qimrberghofer.edu.au/
[61] Diamantina Institute, University of Queensland, di.uq.edu.au/

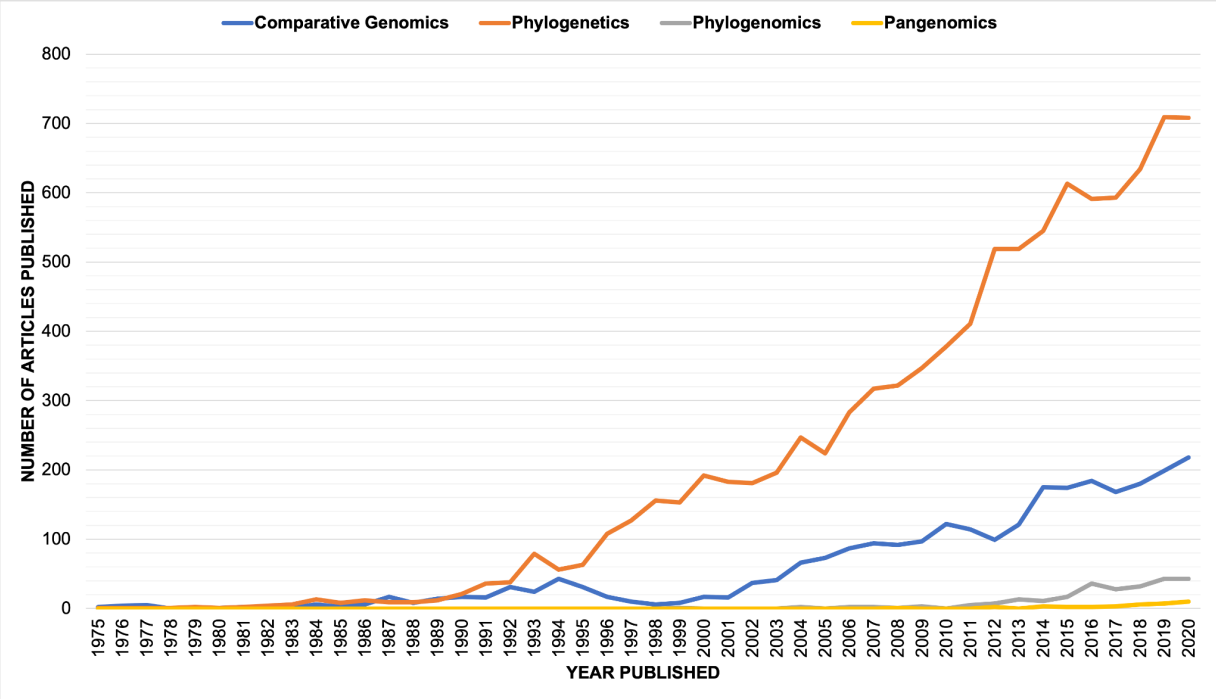

Figure 2: Estimates of the increasing number of comparative genomics studies conducted in Australia

To gain an estimate of the number of different types of comparative genomics studies that have been conducted historically in Australia, a search was conducted of the Scopus database for articles with: either (A) comparative genomic/s, (B) phylogenetic/s, (C) phylogenomic/s or, (D) pangenomic/s' 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 analysis of molecular data, therefore excluding phylogenetic studies looking at morphology or other attributes, as well as excluding others whose focus was on developing or evaluating analysis methods or tools. Duplicate articles that were retrieved during multiple searches were limited to include only one representative article categorised to either category. The complete list of citations including authors, titles and journals can be found here.

In mid and late May 2021, the Australian BioCommons invited over 100 researchers across Australia to participate in a Comparative Genomics Special Interest Group (SIG). These researchers were identified as having experience in, or interest in, comparative genomics, phylogenomics, molecular phylogenetics or pangenomics. The Australian BioCommons sought information from the SIG about each member's level of expertise, current (and desired) practices and infrastructure used via an online survey[62] (number of respondents = 18), and also held two open video conferences to follow-up and gain further information (minutes[63] and recordings[64] of the meetings are available).

[62] Presentation including survey results on comparative genomics infrastructure needs and challenges conducted from 22/03/2021 to 12/05/2021 is here.
[63] Meeting minutes from the comparative genomics SIG meeting held 13/05/2021 are here and a follow-up meeting held 27/05/2021 are here.
[64] Recording of comparative genomics SIG meeting held 13/05/2021 is here and a follow-up meeting held 27/05/2021 is here.

Respondents to the survey and attendees at the meeting collectively indicated they are performing comparative genomics analyses on a wide variety of organisms including prokaryotes such as viruses and bacteria as well as eukaryotes such as vertebrate animals and vascular plants. Project foci cover agriculture, health and medicine and ecological domains of science. The collective responses also indicated that all of the following approaches are being undertaken by Australian researchers: comparative genomics, phylogenetics, pangenomics, phylogenomics and multi-omics studies.

3.3 How are comparative genomics analyses being done in Australia?

The field of comparative genomics covers a broad range of methodologies. Therefore, the data, tools, compute and other needs of researchers performing comparative genomic analyses are also broad. Most researchers using an 'omic technology will compare sequence data in some way to interpret their data biologically. This may be for the purpose of aligning a novel organism's sequence reads to the genome of a closely related species to perform genome assembly. Comparative sequence analysis can also be part of a workflow in the fields of microbiome analysis or genome assembly. In the fields of phylogenomics, molecular phylogenetics and pangenomics, performing comparative sequence analyses is a central part of the workflow.

3.3.1 Data

Based on information received from the SIG members through the survey (*n*=18), most researchers use a combination of sequencing platforms to generate their data with the most popular being Illumina[65], PacBio[66] and Nanopore[67].

Researchers depend on access to up-to-date databases that house information for aligning and comparing gene sequences to identify taxonomy, function and evolutionary lineage. Collectively, the SIG identified that researchers access more than 9 different databases[68] either directly, or through collaborators and colleagues.

To aid with their analyses, 83% of respondents indicated the use of the National Centre for Biotechnology Information (NCBI) database. The next databases each being used by 44% of respondents are the Protein Family Database (Pfam)[69] (which provides information for classification of protein sequences) and the European Molecular Laboratory's European Bioinformatics Institute (EMBL-EBI) numerous databases[70]. In particular, Ensembl Compara[71] is a valuable resource for accessing cross-species whole genome and gene sequence databases.

---

[65] Illumina, Inc.
[66] PacBio, Pacific Biosciences
[67] Oxford Nanopore Technologies
[68] Complete list of databases used by survey respondents with a number of responses shown in brackets: NCBI (15), Pfam (8), EBI (8), shared by colleagues under a non-disclosure agreement or pre-publication (2), KEGG (1), BUSCO (1), COG (1), databases integrated with InterPro (1) and databases integrated with CoGe (1).
[69] Pfam: Protein family database
[70] European Bioinformatics Institute, EBI is part of the European Molecular Biology Laboratory, EMBL, and is sometimes referred to as EMBL-EBI.
[71] https://asia.ensembl.org/info/genome/compara/index.html

Most researcher respondents are accessing the publicly available datasets of closely related taxa (67%) or the same (61%) taxon as their species or organism of interest. A little over a quarter of respondents used private datasets from collaborators of closely related taxa (28%) or the same taxon (33%) and about one fifth of respondents could not access the datasets they needed because either no data existed (17%) or weren't able to access the datasets as they remained private (6%).

3.3.2 Tools

Based on the survey, approximately 47 software tools, pipelines, packages or platforms were identified as being used by respondents for various stages[72] of the comparative genomics analysis process. These are listed in Appendix 1 of this document.

The top tools being used by respondents (28-44%, n=5-8) are related to phylogenetic inference and model choice (e.g. IQ-TREE[73], BEAST2[74], MrBayes[75], RAxML[76]) and multiple sequence alignment (e.g. MAFFT[77], Mauve[78], MUSCLE[79], DIAMOND[80]). Some respondents (22%, n=4) noted that custom tools were developed within their group with one respondent saying that they "*routinely develop comparative genomics tools*", and another stating that they create "*custom scripts to convert the output of one software to another*".

A small number of researchers (n=2) reported that they were not using their preferred tools/pipelines (i.e. G-PhoCS[81], DiscoVista[82], and HybPiper[83]) due to either not having access to sufficient computational memory to run these tools or not having the essential prior knowledge to run the tool.

One respondent echoed this sentiment stating that "*the greatest roadblock has been understanding how to use these programs*" that are only accessible through the command line interface[84]. Access to a complete pangenomic and phylogenetic pipeline that performs all the necessary steps in the analysis workflow[85] was identified by one respondent as a desirable tool

[72] e.g. quality control, preprocessing, sequence assembly, gene prediction and alignment, homology/ortholog prediction, annotation prediction, assembly validation, pangenome construction, phylogenetic tree or phylogenomics tree construction, statistical analysis and visualisation.
[73] IQ-TREE
[74] BEAST2: Bayesian evolutionary analysis by sampling trees
[75] MrBayes: Bayesian Inference of Phylogeny
[76] RAxML: Randomized Axelerate Maximum Likelihood
[77] MAFFT: Multiple alignment program for amino acid or nucleotide sequences
[78] Mauve: Multiple genome alignment
[79] MUSCLE: Multiple Sequence Comparison by Log- Expectation
[80] DIAMOND
[81] G-PhoCS
[82] DiscoVista: Discordance Visualization Tool
[83] HybPiper
[84] From the respondent: "*My work revolves around comparing the secondary metabolite potential of bacteria in a certain genus. Though there are multiple tools available for this, most of them use a command line interface which is not the most user friendly. So the greatest roadblock has been understanding how to use these programs.*"
[85] From the respondent: "*A pangenomic pipeline that is scalable that takes in multiple genome assemblies, annotates these the exact same way and performs a pangenomic analysis would be good. Minimal preprocessing of input files separately would be ideal. Same for Phylogenetics, a pipeline that takes in multiple genome assemblies, annotates these, pulls out gene orthologs, builds a gene matrix and MSA ready for phylogenetic analysis would be a big advantage. The continuous increase in number of sequenced genomes necessitates use of standardised methods that are benchmarked for accuracy, this makes working on large genomic possible and reproducible*"

to allow them to perform their research. One respondent suggested a pangenomic pipeline for eukaryotes similar to Roary[86] would be beneficial and another respondent suggested having access to automatic genome annotation pipelines, such as FGENESH++[87] would be beneficial so as to perform this step prior to pangenome construction.

3.3.3 Compute infrastructure

3.3.3.1 Types used

Survey respondents (*n* = 18) currently use a variety of computational infrastructure for their analyses.

Most access laptops or personal computers (61%) or high-performance computing provided by their host institute (50%), with fewer respondents using shared high-performance computers managed by national (e.g. NCI[88], Pawsey[89]) or state (e.g. QCIF/QRIScloud[90]) computing centres (31%), or accessing commercial cloud resources, such as Amazon Web Services (AWS)[91], Google Cloud[92] or Microsoft Azure[93] (5%).

All respondents (100%) use more than one of these compute-infrastructure types to support their work and mix and match their use to the problem at hand.

3.3.3.2. Resourcing

More than half of the respondents (70%, *n* = 7) said the infrastructure they currently had access to was not sufficient for their current comparative genomics work, due to limitations in available memory, data storage allocations, limited access (due to busy clusters) or old hardware systems. One respondent stated that their institutional HPC takes a long time to use due to long queues and often 'crashes'[94]. Another respondent stated that they were able to handle one project at a time but since computational resources are limited, when there are multiple projects to run, their available allocation is insufficient[95]. And one respondent stated that the older system hardware that is available to them limits their ability to perform complete analyses[96].

3.4 Challenges being faced

A variety of limitations/roadblocks/challenges/issues with current infrastructure were identified

[86] Roary: the pangenome pipeline
[87] FGENESH++
[88] NCI, National Computational Infrastructure, nci.org.au/
[89] Pawsey Supercomputing Centre, pawsey.org.au/
[90] Queensland Cyber Infrastructure Foundation, QCIF
[91] Amazon Web Services
[92] Google Cloud
[93] Microsoft Azure
[94] From the respondent: "*The HPC in our institute could use more memory, the jobs take long to start (long SLURM queues), and it often crashes.*"
[95] From the respondent: "*We can handle one project at a time (a crop species), but struggle with multiple projects that need to run within the same time frame*"
[96] From the respondent: "*Because we have not access to up to date unix systems. Our current systems are over 6 years old.*"

by the SIG.

3.4.1 Computational resourcing and set-up challenges

- Computational resources available (even across a variety of infrastructures) can be insufficient with requirements dependent on each project. Processing phylogenetic or phylogenomic inference trees requires access to many CPU nodes with just a small amount of RAM, but with extended wall times (e.g. up to 100 16GB nodes and 2GB RAM for *weeks or months*) to allow for multiple alignments and large parallelisation of numerous sequences. In contrast, processing pangenomes require access to many CPUs with a large amount of RAM and extended wall times (e.g. up to 20 cores and 1TB RAM for weeks). Workarounds include either limiting the dataset size or accessing commercial Google Cloud Platform (GCP)[97] and Amazon Web Services (AWS) clouds which incur a financial cost with each analysis;

- Some respondents (n=2) suggested having access to flexible workflows or pipelines that can take in multiple genome assemblies or sequences to minimise individual sample data processing to produce a pangenome, phylogenetic tree or associated outputs would be valuable. Also having access to standard tools that have been benchmarked for accuracy and available in a shared format or repository was viewed as being beneficial by three SIG group members.

- Some respondents (n=2) stated that they were limited by the lack of understanding about how to use programs and software with the command line. So although tools are readily available, as one respondent stated "*most of them use a command line interface which is not the most user friendly. So the greatest roadblock has been understanding how to use these programs.*"

- Sharing and organising data was identified by respondents (n=3) as a large hurdle with a desire to have access to better "*methods to share large files with collaborators*" and "*main problems are that external collaborators do not have access to our institutional high performance computing infrastructure, and that large files are difficult to share any other way*".

3.4.2 Data related challenges

- Data publishing from Australia to international repositories (e.g. GenBank/SRA) is considered by some (n=4) to be difficult - primarily due to an unclear submission process, changing input requirements, and issues with uploading the data to the repositories. As one respondent stated "*I find the Genbank submission process increasingly difficult*".

  More generally, one SIG member stated that the requirements for metadata required by

[97] Google Cloud

the international repositories are "*getting out of hand*", often with dozens of fields whose only realistic entries are '*does not apply in my case*' or '*I have no idea what is meant with this field*'. Another respondent wants: "*Streamlined methods/tools/processes for submission to Genbank*".

3.4.3. Other challenges

- Access to a user-friendly interface with appropriate computational resourcing for teaching was indicated as being desirable by one respondent. As the respondent stated: I would like to "*teach students about access to a graphical user interface for teaching phylogenomics to non-bioinformaticians and students*";

- Another challenge was to enable greater data linkage for conservation projects with one respondent stating they would like "*to have access to web accessible tools to allow conservation analysis data to be shared with relevant conservation organisations and linked into other species occurrence registries/data resources*".

- One respondent, who is the coordinator of a project where ~40 unpublished genomes are being assembled, requires them to be privately accessed by consortium members for comparative analyses via a BLAST function but lacks the IT ability to establish such a system.

- Although most respondents had access to expertise to build and maintain their computational infrastructure, including installing and updating software, 25% of respondents stated that they did not have access to expertise to maintain and update their infrastructure. Maintaining tools via a set of managed modules, common on high-performance computing, is burdensome. The high task load required to maintain an ever-growing list of tools including old versions as well as the latest version is generally no longer justified for staffing on institutional infrastructures. In addition a number of tools require a specific computational architecture set up for the program to actually run successfully. So, despite the relatively easy install of a complex tool, getting the tool to run with real data may be extremely challenging.

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

All but two of the respondents (94%, *n* = 16) agreed that if a shared data collaboration/analysis platform for comparative genomics analysis was available for use, they would use such a platform provided it was easy to use. This number included respondents who stated that their needs are currently met.

Twenty-three hypothetical features of such a shared 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 several features of a shared platform deemed the most crucial are: (1) ease of uploading/downloading data; (2) access to preferred

tools/pipelines; (3) an ability to transfer data easily to/from storage; (4) good documentation on platform use; (5) ease of access from anywhere; and (6) quick installation of additional tools/pipelines on request.

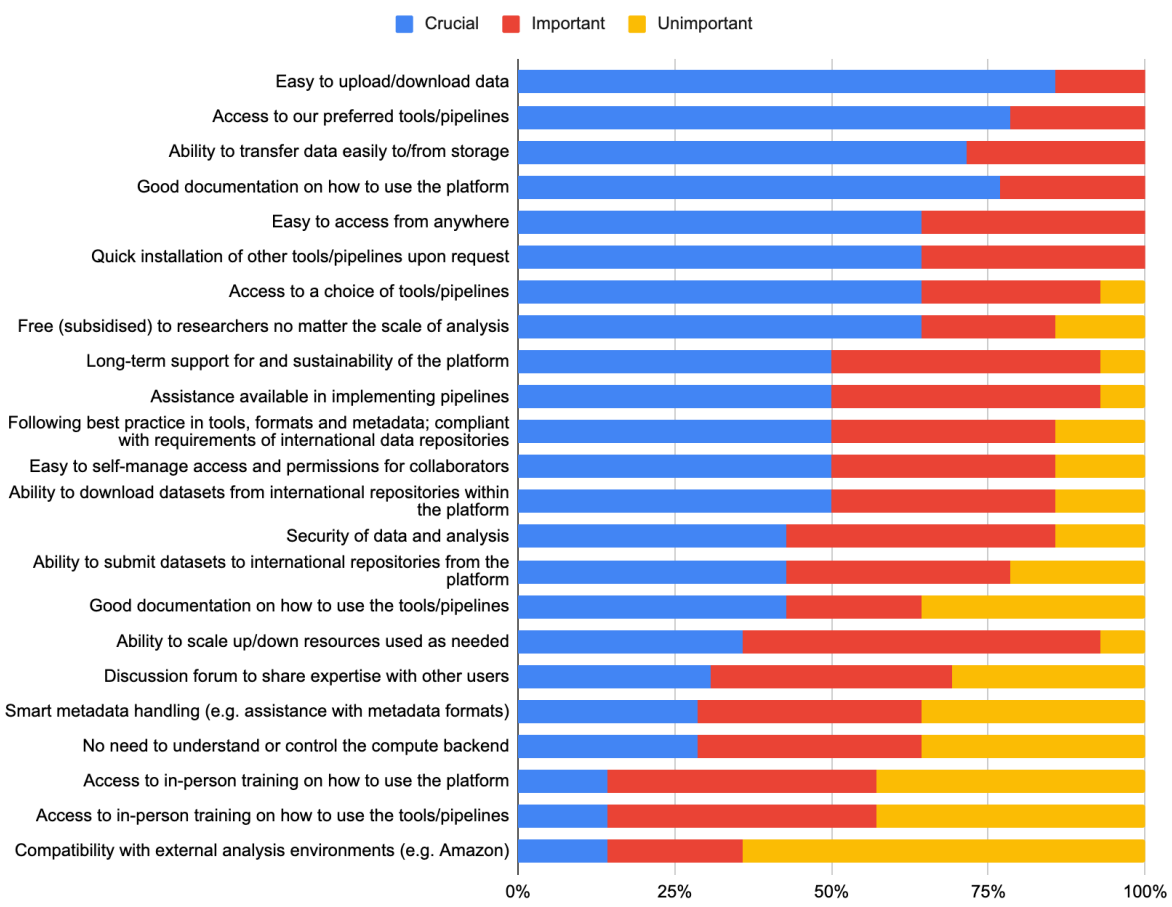

Figure 3: Desired features in a shared comparative genomics analysis platform

Survey respondents were asked about which features they considered to be ‘crucial’, ‘important’ or ‘unimportant’ in a shared comparative genomics analysis workspace. The number of responses classified at each level is shown per feature, and features are ranked in descending order from those deemed to be most crucial to the least crucial.

4. Meeting the Needs of Australian Researchers for High-quality, Accessible Comparative Genomics Analysis Infrastructure

4.1 Goal

The Australian BioCommons aims to develop a ‘Comparative Genomics Infrastructure Roadmap for Australia’ that describes collaborative infrastructure, which, when implemented (from Q1/Q2 2022 onwards), will enable Australian researchers from a wide range of institutions to perform high-quality comparative genomic analysis work (including phylogenetic tree and pangenome

construction) 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 the development of the Roadmap are as follows:

- V1 (this document) - Content-based on SIG survey results and input from SIG meeting - December 2021.
- V2 - Content modified to incorporate feedback from various national computational infrastructure providers - December 2021
- V3 -Content modified to incorporate feedback from SIG, other researchers undertaking comparative genomics analysis, and international groups - January/February 2022.
- V4 - Content modified to incorporate final feedback from SIG - February 2022.

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 comparative genomics analyses (whether they be molecular phylogenetics, phylogenomics, or pangenomic based) 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 (locally and internationally).
2. To make it easier for Australian researchers to perform visualisation and statistical -based analyses of genomic, pangenomic and phylogenomic data; and,
3. To make it easier to publish high-quality comparative genomic-associated data 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 comparative genomic analysis

D2. Systems to enable visualisation and statistical analyses of comparative genomics, phylogenetic, phylogenomic and pangenomic data and their data products

D3. Systems to enable submission of raw sequencing reads, phylogenetic/phylogenetic trees and pangenomes files from Australia to appropriate repositories

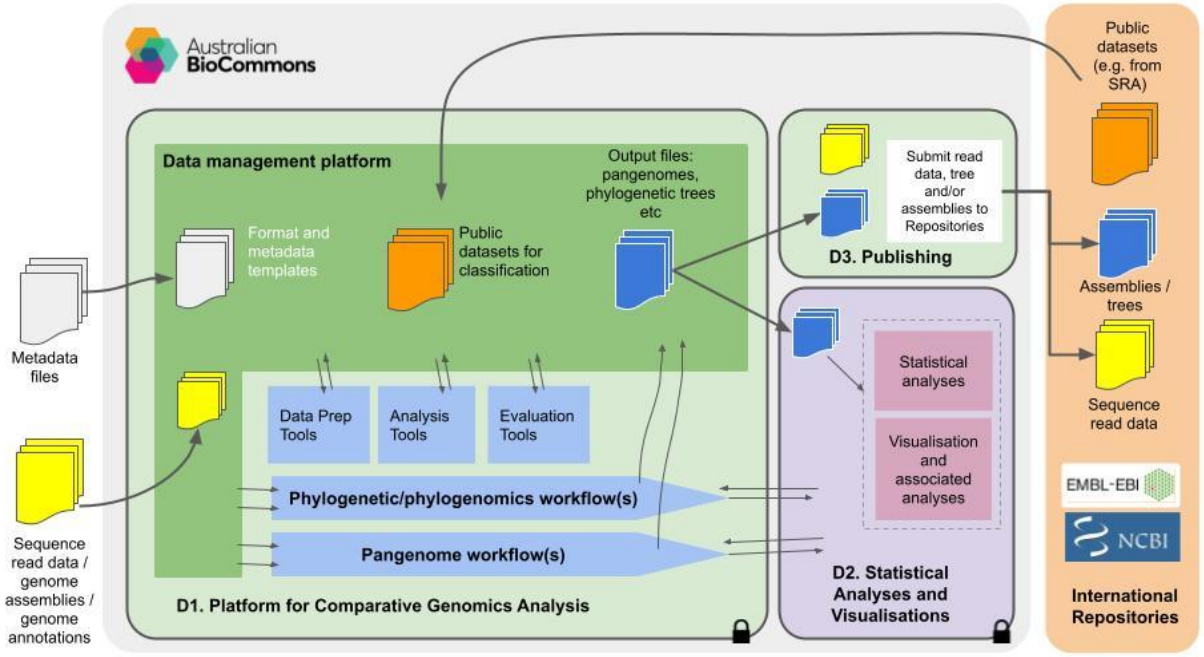

Figure 4. Schematic diagram showing the proposed infrastructure to support comparative genomics analyses, and data flow

D1 - Sequence reads, assembled genomes or other relevant data (either locally sourced, or obtained from various public nucleic sequence repositories) are inputs into the Platform for Comparative Genomics Analysis which provides a command-line interface (CLI) or graphical user interface (GUI) access to a variety of tools and workflows for performing comparative genomic analyses, including phylogenetic/phylogenomic tree or pangenome construction (blue shapes). It is underpinned by sufficient and appropriate computational infrastructure. Closely associated is a data management platform (denoted by the darker green shape) that caters to data management, version control, and association of appropriate (e.g. sample, experimental) metadata with the data files. Outputs of D1 are accessible to both: **D2 -** hosted frameworks to enable researchers to utilise common packages for statistical analysis, visualisation, and exploration of comparative genomic, pangenomic, phylogenetic and phylogenomic datasets and their data products; and **D3** - systems to enable submission/publishing of phylogenetic/phylogenomic trees and pangenome files (and sequence read data) to international and national repositories. Arrows indicate the general flow of data. See Appendix 1 for a list of tools/pipelines that may be included in D1. Higher resolution image.

D1 - A platform for comparative genomics analyses;

To address objective 1 (*i.e. providing Australian researchers with access to a selection of tools and workflows underpinned by computational resources that allow phylogenetic and phylogenomic tree construction, pangenome generation and other comparative genomic analyses to be performed*), it is proposed to implement a platform in Australia[98], that:

[98] 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.

A. Includes a set of key tools[99] and/or pipelines for various types of comparative genomic analyses (including for example, tools for data preparation, quality control, tree construction, pangenome assembly etc):

    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 2[100] shared computational infrastructures) underpinned by appropriate computational resources[101];

    b. Installed (plus all other dependencies) and optimised on a graphical user interface (GUI) web-based data analysis platform where possible, (i.e. Galaxy Australia[102]), underpinned by appropriate computational resources; and,

    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, including public national and/or international databases (e.g. those hosted by NCBI/EMBL-EBI), national (i.e. CloudStor[103]), institutional or other data storage, and with the ability to upload/mount user-generated or other datasets[104] that are required as inputs for a comparative genomics analysis pipeline. This should include a feature that enables a user to easily search and locate relevant data for their comparative analysis across a variety of public national and/or international sequence databases;

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 files to international repositories, when required;

F. Has support available from experts in formatting data and curating metadata to comply with any international repository format requirements[105];

---

[99] e.g. a selection of the tools listed in Appendix 1

[100] 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 many institutionally operated facilities.

[101] Including necessary high memory nodes (>1TB RAM) for performing large coassembly research and phylogenetic tree creation. See also biocommons.org.au/pathfinder-biocloud

[102] Galaxy Australia

[103] CloudStor Service

[104] Including necessary datasets for classification, comparison of genome sequences.

[105] 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

G. Includes documentation, including a knowledgebase with community-contributed content; and,

H. Includes training[106] for all the above.

D2. Systems to enable statistical analyses and visualisations of comparative genomics analysis results:

To address objective 2 (*i.e. to make it easier for Australian researchers to perform statistical and visualisation analyses of phylogenetic, phylogenomic, pangenomic and comparative genomic data*), it is proposed to implement:

A. Hosted frameworks to enable researchers to utilise common packages for statistical analysis, visualisation, and exploration of comparative genomic associated datasets;

B. Appropriate user authorisation and sharing mechanisms to allow for public or private data and associated data product(s) 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.

[106] Training will be developed and delivered by the Australian BioCommons bioinformatics training program (biocommons.org.au/biocommons-training) in conjunction with the objectives identified in this Roadmap. The goals identified by the Australian BioCommons' National Training Strategy are:

- Goal #1 - to produce life scientists who can confidently use GUI-based bioinformatics tools
- Goal #2 - to produce life scientists who can confidently begin to utilise scripting-based bioinformatics tools/methodologies
- Goal #3 - to produce bioinformaticians who routinely observe best practice in algorithm and tool development
- Goal #4 - to produce a national network of bioinformatics training instructors and training material developers who are integrated into the global bioinformatics training community

D3 - Systems to enable submission of phylogenetic, phylogenomic, pangenomic and comparative genomic data and associated output files from Australia to appropriate global repositories:

To address objective 3 (*i.e. to make it easier to publish high quality and share final raw phylogenetic trees, pangenomes and associated data products (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 draft pangenomes and phylogenetic trees (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-E), and support as detailed in D1-F;

B. Includes a rapid data transfer from the data management platform or the sharing platform to relevant national and/or international repositories; and,

C. Documentation on how to use the system (including a knowledgebase with community-contributed content).

4.4 Implementation timeframes

It is intended that the components identified in Section 4.3 will be implemented throughout 2022-2023.

As of December 2021, several key activities that are relevant to the proposed infrastructure are already underway:

Component	Planned dates for delivery	Notes
D1-Aa/Ab. Key tools[107] and/or pipelines for data preparation, quality control, tree construction, pangenome assemblies and associated data products from comparative genomic analyses.	Ongoing	Researchers can easily identify which bioinformatic tools are currently installed as modules across several national, state and institutional infrastructures, including Pawsey, NCI, QRIScloud/UQ-RCC, UNSW and Galaxy Australia, as well as finding links to downloadable software containers for installation anywhere through the searchable interactive webpage the BioCommons ToolFinder. Tool installation request mechanisms on these systems are also highlighted through the ToolFinder.
D1-Aa. Key tools installed (plus all other dependencies) and optimised on a command line interface (CLI) analysis environment (i.e. across a variety of Tier 1 and 2[108] shared computational infrastructures)	Ongoing	As of November 2021, six of the tools listed in Appendix 1 (DendroPy, CheckM, GraPhlAn, BEAST, Cactus, bbmap) are installed as modules on QRIScloud/UQ-RCC HPC machines (Tinaroo[109], Awoonga[110], FlashLite[111]); three of the tools (ExaBayes, beagle-lib, BEAST) are installed as modules at Pawsey; and four of the tools (beagle-lib, BEAST, bbmap, Cactus) are installed as modules at NCI. Installation of further tools as modules across NCI, Pawsey, and QRIScloud/UQ-RCC infrastructures to

[107] e.g. a selection of the tools listed in Appendix 1
[108] 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 many institutionally operated facilities.
[109] Tinaroo high performance computer.
[110] Awoonga high performance computer.
[111] FlashLite high performance computer.

		support phylogenetic, phylogenomic, pangenomic and other comparative genomics analysis is being undertaken in the BioCommons 'BYOD' Expansion Project. Users of these systems can also request tool installation on these systems (see 'Requesting tool installations' in the ToolFinder).
D1-Aa. CLI platform appropriately resourced for performing comparative genomics analyses	Ongoing	BioCommons partner infrastructures at NCI, Pawsey, and QCIF include computational systems that are capable of performing any part of phylogenetic, phylogenomic, pangenomic and comparative genomics analysis. This includes FlashLite at QCIF/UQ which can be structured to allow 'supernodes' of up to 8TB). Enabling increased researcher access to partner HPC systems via mechanisms other than through the National Computational Merit Allocation Scheme (NCMAS) or partner shares are under active exploration by NCI (Adapter scheme) and the BioCommons (Australian BioCommons Leadership Share (ABLeS) which is currently being established to support the generation of reference datasets of national importance.
D1-Ab. Key tools installed (plus all other dependencies) and optimised on a graphical user interface (GUI) web-based data analysis platform where possible, (i.e. Galaxy Australia[112]), underpinned by appropriate computational resources; and,	Ongoing	As of November 2021, 12 of the tools listed in Appendix 1 (Roary, Chromeister, HyPhy-aBSREL, GenomeScope, PhyML, Gubbins, RAxML, Pangolin, HyPhy-GARD, IQ-TREE, FastTree, GraPhlAn) are installed on Galaxy Australia. Installation of further tools on Galaxy Australia can be requested by any member of the community at any time.
D1-Ab. Galaxy Australia appropriately resourced for performing comparative genomics analyses	Ongoing	As of December 2021, Galaxy Australia is underpinned by a total of 1476 vCPUs and 20TB RAM, including one 2TB and four 4TB high memory nodes that are reserved for specific tools requiring high memory.
D1-Ac. Key tools available as high quality, trusted software containers for self-deployment on institutional or independent computational infrastructures.	Ongoing	As of November 2021, 13 of the tools listed in Appendix 1 (Anvi'o, Beast2, DendroPy, GraPhlAn, CheckM, IQ-Tree, Cactus, bbmap, beagle-lib, BEAST, BEAST2, ExaBayes, MrBayes) are

[112] Galaxy Australia

		available as containers (either Bioconda, Docker or Singularity). The development of containerised tools to support various life science research communities in Australia (including comparative genomics) is being undertaken in the BioCommons 'BYOD' Expansion Project.
D1-C. Connectable to public national and/or international databases	Q1/2 2022	A web-accessible system ("ARGA - the Australian Reference Genome Atlas") that enables a user to easily search and locate relevant genomic data from macroscopic Australian native or agriculturally relevant species for comparative genomics analyses from a variety of public national and/or international sequence data repositories will be jointly developed during 2022 by the Australian BioCommons, Bioplatforms Australia and the Atlas of Living Australia. The ARGA system will enable integrated searching for genomic data across multiple repositories via taxonomic groupings, species occurrence or functional traits as well as access to the associated genomic sequence files (made possible through API calls to the public sequence data repositories). It will allow a user to either download the data for local analysis or to directly push data from the repository to online systems such as Galaxy Australia for further analysis.
D1-C. Connectable to Nationally available storage (e.g. Cloudstor)	Ongoing	In late 2020, a direct connection between Cloudstor and Galaxy Australia was implemented. Streamlined connectivity of storage to Pawsey, QCIF, NCI, and other computational resources will continue in the BioCommons 'BYOD' Expansion Project.
D1-D/D2-B. Appropriate user authorisation and sharing mechanisms	Ongoing	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. Technical solution options for future deployment include CILogon.
D1-D. A data management		Considerations for what may be the best technical

system that is tightly linked to the Comparative Genomics Platform		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.	Ongoing	Tool and workflow documentation and discovery are available via the BioCommons ToolFinder and WorkflowFinder Services, respectively. These Services aim to cover all research communities to provide a 'hub' for central resources and provide a more flexible mechanism for software and workflow access and availability on infrastructure.
D1-G, D2-C and D3-C. Documentation on how to use the system (including a knowledgebase with community-contributed content) and Training		Support to researchers in the form of Frequently Asked Questions and documentation are available on the Australian BioCommons Support Pages. Further support and documentation are provided in the format of the ToolFinder, WorkflowFinder BioCommons Training Program and Galaxy Training Program. The growth of technical documentation and support is ongoing and currently being explored in a number of formats beyond the current offerings in the BioCommons.
D1-H. Training for all aspects of the Comparative Genomics Platform, Statistics and Visualisation Platform.	Ongoing	Introductory level training has occurred for a number of relevant skills, including building phylogenetic trees[113], software containerisation[114]; getting started with command line[115], Galaxy Australia[116] and R[117] along with >60 webinar and workshop recordings. You can search the Australian BioCommons Training Materials Zenodo Repository for materials as well as the Australian BioCommons YouTube channel for recordings from many of these events. Further training material for comparative genomics can be located through international bioinformatics training repositories such as TeSS or GOBLET, and

[113] Phylogenetic Trees Workshop: Phylogenetic Trees: Back to Basics and Phylogenetic Trees: Back to Basics Q & A, delivered by Professor Michael Charleston
[114] Containers in Bioinformatics Workshop: Containerising a Pipeline, Building Containers and BYO Pipeline, delivered by Dr Marco de la Pierre
[115] Webinar: Getting Started with Command Line Bioinformatics, delivered by Parice Brandies
[116] Webinar: Galaxy Australia: a Strengthened National Life Science Platform that Engages Globally, delivered by Dr Gareth Price
[117] Webinar: Getting Started with R, delivered by Dr Saskia Freytag

		the development and launch of the DReSA (Digital Research Skills Australasia) platform in November 2021 provides a further way of discovering digital research and training resources from a number of informatics training providers in Australasia.
D2-A. Hosted frameworks to enable researchers to utilise common packages for statistical analysis, visualisation, and exploration of comparative genomic associated datasets.		'Interactive environments' offered through the Galaxy Australia platform include R-Studio[118] and Jupyter notebooks[119] are being trialled with planned implementation in 2022. AARNet's CloudStor also offers a web-accessible Jupyter notebook through their SWAN[120] service. Options for displaying phylogenetic trees within a broader context (e.g. within the context of species distribution or environmental data) include the Atlas of Living Australia's Phylolink[121] which represents one option for inclusion in the Comparative Genomics platform but requires further exploration. As of December 2022, systems to enable researchers within a collaboration to better search across multi-genome datasets, are being explored by the Australian BioCommons and its key operational partner QCIF. In this project, https://sequenceserver.com/ is being investigated as a system that could be used to enable researchers and their collaborators to perform sequence similarity searches across private (pre-publication) genome or sequence collections. The European Galaxy service also enables access to HiGlass through a Galaxy Interactive Environment, which could be installed on Galaxy Australia if desired by the community.
D3-A and D3-B. A temporary 'staging post' in Australia for draft pangenomes and phylogenetic trees (and sequence read) files ready for public international release.		COPO is a GUI-based metadata platform for brokering life science data submissions to various repositories including the ENA (see https://f1000research.com/articles/9-495). It is being adopted by the Darwin Tree of Life

[118] R-studio, rstudio.com/
[119] Jupyter Notebook, /jupyter.org/
[120] CloudStor SWAN
[121] Phylolink, phylolink.ala.org.au/

		project in the UK as the tool to enable the data and metadata submission to ENA to be completed for genome assemblies of over 60,000 species native to the British Isles. The Australian Biocommons is currently exploring whether a locally supported COPO instance can fulfil the requirements of D3-A/D3-B. The recently developed Galaxy ENA Reads Submission tool which has been developed by the ELIXIR Belgium Node to support the streamlined submission of COVID-19 data to ENA is another tool that will be explored for inclusion on Galaxy Australia to support this feature.

Appendix 1

Table 1. Comparative genomics analysis tools for consideration for inclusion in a shared analysis environment.

Note that a comparative genomics analysis protocol may also incorporate many other software tools not listed here. Specific genome assembly, genome annotation and microbiome analysis tools are listed elsewhere[122].

Workflow Step	High-level component	Tool	Brief description	Link to data/software or article
1	Quality Control	BUSCO	Based on evolutionarily-informed expectations of gene content of near-universal single-copy orthologs, BUSCO metric is complementary to technical metrics like N50. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.	https://busco.ezlab.org/
1	Quality Control	CheckM	CheckM provides a set of tools for assessing the quality of genomes recovered from isolates, single cells, or metagenomes.	https://ecogenomics.github.io/CheckM/
1	Quality Control	FastQC	Provides a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines.	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
1	Quality Control	HTStream	A quality control and processing pipeline for High Throughput Sequencing data.	https://s4hts.github.io/HTStream/
1	Quality Control	samblaster	samblaster is a fast and flexible program for marking duplicates in read-id grouped1 paired-end SAM files. It can also optionally output discordant read pairs and/or split read mappings to separate SAM files, and/or unmapped/clipped reads to a separate FASTQ file.	https://github.com/GregoryFaust/samblaster
2	Assembly Validation	CompareM	A software toolkit that supports performing large-scale comparative genomic analyses. It provides statistics across sets of genomes (e.g., amino acid identity) and for individual genomes.	https://github.com/dparks1134/CompareM

122

2	Preprocessing	BLAST+	A suite of command line tools to run BLAST which is to search for nucleotide similarities.	https://blast.ncbi.nlm.nih.gov/Blast.cgi
2	Preprocessing	FASTX-Toolkit	A collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.	http://hannonlab.cshl.edu/fastx_toolkit/
2	Preprocessing	Gblocks	Gblocks is a computer program written in ANSI C language that eliminates poorly aligned positions and divergent regions of an alignment of DNA or protein sequences. These positions may not be homologous or may have been saturated by multiple substitutions and it is convenient to eliminate them prior to phylogenetic analysis. Gblocks selects blocks in a similar way as it is usually done by hand but following a reproducible set of conditions.	http://molevol.cmima.csic.es/castresana/Gblocks.html
2	Preprocessing	gubbins	Gubbins (Genealogies Unbiased By recomBinations In Nucleotide Sequences) is an algorithm that iteratively identifies loci containing elevated densities of base substitutions while concurrently constructing a phylogeny based on the putative point mutations outside of these regions.	https://sanger-pathogens.github.io/gubbins/
2	Preprocessing	MultiQC	A reporting tool that parses summary statistics from results and log files generated by other bioinformatics tools.	https://multiqc.info/docs/
2	Preprocessing	Trimmomatic	A flexible read trimming tool for Illumina NGS data.	http://www.usadellab.org/cms/?page=trimmomatic
2	Preprocessing	UCHIME/ UCHIME2	Chimera detection tool.	https://www.drive5.com/usearch/manual/uchime2_algo.html https://www.biorxiv.org/content/10.1101/074252v1.full
2	Preprocessing	GenomeScope	Fast genome analysis from unassembled short reads. We have developed an analytical model and open-source software package GenomeScope that can infer the global properties of a genome from unassembled sequenced data. GenomeScope uses the k-mer count distribution, e.g. from Jellyfish, and within seconds produces a report and several informative plots describing the genome properties. We validate the	http://qb.cshl.edu/genomescope/

			approach on simulated heterozygous genomes, as well as synthetic crosses of related strains of microbial and eukaryotic genomes with known reference genomes. GenomeScope was also applied to study the characteristics of several novel species, including pineapple, pear, the regenerative flatworm Macrostomum lignano, and the Asian sea bass.	
3	Haplotype Estimator	SHAPEIT	SHAPEIT is a fast and accurate method for estimation of haplotypes (aka phasing) from genotype or sequencing data.	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
3	Ortholog Detection	HybPiper	We developed HybPiper as a user-friendly platform for assembly of gene regions, extraction of exon and intron sequences, and identification of paralogous gene copies. We test HybPiper using baits designed to target 333 phylogenetic markers and 125 genes of functional significance in Artocarpus (Moraceae).	https://github.com/mossmatters/HybPiper
3	Ortholog Prediction	COG Clusters of Orthologous Groups of proteins	A developed system for delineation of Clusters of Orthologous Groups of proteins (COGs) from the sequenced genomes of prokaryotes and unicellular eukaryotes and the construction of clusters of predicted orthologs.	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC222959/
3	Ortholog Prediction	eggNOG	A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.	http://eggnog5.embl.de/#/app/home
3	Ortholog Prediction	eggNOG- mapper	A tool for fast functional annotation of novel sequences.	https://github.com/eggnogdb/eggnog-mapper
3	Ortholog Prediction	KAAS - KEGG Automatic Annotation Server	Provides functional annotation of genes by BLAST or GHOST comparisons against the manually curated KEGG GENES database.	https://www.genome.jp/kegg/kaas/
3	Ortholog Prediction	KofamKOALA	A web server to assign KEGG Orthologs (KOs) to protein sequences by homology search.	https://www.genome.jp/tools/kofamkoala
3	Ortholog Prediction	OMA: Orthologous MAtrix	The OMA (“Orthologous MAtrix”) project is a method and database for the inference of orthologs among complete genomes.	https://omabrowser.org/

3	Ortholog Prediction	orthoDB	The hierarchical catalogue of orthologs.	https://www.orthodb.org/
3	Ortholog Prediction	orthofinder	OrthoFinder is a fast, accurate and comprehensive platform for comparative genomics. It finds orthogroups and orthologs, infers rooted gene trees for all orthogroups and identifies all of the gene duplication events in those gene trees.	https://github.com/davidemms/OrthoFinder
3	Ortholog Prediction	orthograph	Orthology prediction using a graph-based, reciprocal approach with profile hidden Markov models	https://mptrsen.github.io/Orthograph/
3	Ortholog Prediction	OrthoMCL	OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences. Such orthologous sequences not only share evolutionary history but also share function.	https://orthomcl.org/orthomcl/app
3	Ortholog Prediction	PhyloTreePruner	Phylogenomics, phylogenetic inference using large multigene datasets, relies on accurate identification of orthologous sequences among the taxa of interest. In order to improve orthology determination for phylogenomics using a tree-based approach, we have developed PhyloTreePruner, a utility that uses a phylogenetic approach to refine orthology inferences made by graph-based (or other) methods.	
3	Ortholog Prediction	PROKKA	Annotation tool for bacterial, archaeal, and viral genomes.	http://www.metagenomics.wiki/tools/annotation/prokka
3	Model Selection	aBSREL (adaptive Branch-Site Random Effects Likelihood)	aBSREL (adaptive Branch-Site Random Effects Likelihood) is an improved version of the commonly-used "branch-site" models, which are used to test if positive selection has occurred on a proportion of branches. As such, aBSREL models both site-level and branch-level ω heterogeneity. aBSREL, however, does not test for selection at specific sites. Instead, aBSREL will test, for each branch (or branch of interest) in the phylogeny, whether a proportion of sites have evolved under positive selection. Implemented in the software package HyPhy (see below).	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4408413/
3	Recombination Identification	GARD (Genetic Algorithm for	GARD (Genetic Algorithm for Recombination Detection) is a method to screen a multiple sequence analysis for	https://academic.oup.com/mbe/article/23/10/1891/1096946

		Recombination Detection)	the presence of recombination and is extremely useful as a pre-processing step for selection inference. Because recombinant sequences cannot be adequately described with a single phylogenetic history, selection inference on recombinant data often leads to a significant increase in false positives. GARD alleviates this concern by comprehensively screening an alignment for recombination breakpoints and inferring a unique phylogenetic history for each detected recombination block. Implemented in the software package HyPhy (see below).	
3	Synteny Detection	MCScanX	The MCScanX toolkit implements an adjusted MCScan algorithm for detection of synteny and collinearity that extends the original software by incorporating 14 utility programs for visualisation of results and additional downstream analyses. Applications of MCScanX to several sequenced plant genomes and gene families are shown as examples. MCScanX can be used to effectively analyse chromosome structural changes, and reveal the history of gene family expansions that might contribute to the adaptation of lineages and taxa.	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3326336/
4	Clustering Sequences	UCLUST	The UCLUST algorithm divides a set of sequences into clusters. UCLUST is not designed for OTU clustering. See recommended protocols for OTU analysis.	https://drive5.com/usearch/manual/uclust_algo.html
4	Gene prediction and alignment	BBMap	Splice-aware global aligner for DNA and RNA sequencing reads. It can align reads from all major platforms.	https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/
4	Gene prediction and alignment	BLAT	Accurate and 500 times faster than popular existing tools for mRNA/DNA alignments.	https://genome.cshlp.org/content/12/4/656
4	Gene prediction and alignment	BMGE - Block Mapping and Gathering with Entropy	Designed to select regions in a multiple sequence alignment that are suited for phylogenetic inference.	https://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-10-210
4	Gene prediction and alignment	BWA	A software package for mapping low-divergent sequences against a large reference genome, such as the human genome.	http://bio-bwa.sourceforge.net/

4	Gene prediction and alignment	CD-HIT/CD-HIT_EST	A very widely used program for clustering and comparing protein or nucleotide sequences.	http://weizhongli-lab.org/cd-hit/
4	Gene prediction and alignment	DIAMOND	A sequence aligner for protein and translated DNA searches, designed for high performance analysis of big sequence data.	http://www.diamondsearch.org/index.php
4	Gene prediction and alignment	HMMER	Biosequence analysis using profile hidden Markov models.	http://hmmer.org/
4	Gene prediction and alignment	IQ-TREE	Phylogenetic tree inference by maximum likelihood.	http://www.iqtree.org/
4	Gene prediction and alignment	mauve	A system for constructing multiple genome alignments in the presence of large-scale evolutionary events such as rearrangement and inversion.	http://darlinglab.org/mauve/mauve.html
4	Gene prediction and alignment	PhyloSift	A suite of software tools to conduct phylogenetic analysis of genomes and metagenomes.	https://github.com/gjospin/PhyloSift
4	Gene prediction and alignment	PSORTm / PSORTb	For protein subcellular localization prediction (SCL).	https://www.psort.org/psortm/
4	Gene prediction and alignment	pyani	a Python package and standalone program for calculation of whole-genome similarity measures.	https://pyani.readthedocs.io/_/downloads/en/latest/pdf/
4	Gene prediction and alignment	tRNAscan-SE	The de facto tool for predicting tRNA genes in whole genomes.	http://trna.ucsc.edu/tRNAscan-SE/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6768409/
4	Sequence Alignment	Clustal Omega	Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.	https://www.ebi.ac.uk/Tools/msa/clustalo/
4	Sequence Alignment	Cactus	Cactus is a reference-free whole-genome multiple alignment program.	https://github.com/ComparativeGenomicsToolkit/cactus

4	Sequence Alignment	Ensembl	Ensembl Compara provides cross-species resources and analyses, at both the sequence level and the gene level. These data can be accessed in various ways.	https://asia.ensembl.org/info/genome/compara/index.html
4	Sequence Alignment	Harvest	Harvest is a suite of core-genome alignment and visualisation tools for quickly analysing thousands of intraspecific microbial genomes. Harvest includes Parsnp, a fast core-genome multi-aligner, and Gingr, a dynamic visual platform. Combined they provide interactive core-genome alignments, variant calls, recombination detection, and phylogenetic trees.	
4	Sequence alignment	MAFFT - Multiple Alignment with Fast Fourier Transform	MAFFT (Multiple Alignment using Fast Fourier Transform) is a high speed multiple sequence alignment program. We have recently changed the default parameter settings for MAFFT. Alignments should run much more quickly and larger DNA alignments can be carried out by default. Please click the 'More options' button to review the defaults and change them if required.	https://www.ebi.ac.uk/Tools/msa/mafft/
4	Sequence Alignment	Muscle	MUSCLE stands for MUltiple Sequence Comparison by Log- Expectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.	https://www.ebi.ac.uk/Tools/msa/muscle/
4	Sequence Alignment	Chromeister	An ultra fast, heuristic approach to detect conserved signals in extremely large pairwise genome comparisons.	https://github.com/estebanpw/chromeister
4	Sequence Alignment	BLAST+	Suite of command line tools to run BLAST which is to search for nucleotide similarities.	https://blast.ncbi.nlm.nih.gov/Blast.cgi
4	Variant Identification	SNVPhyl: Single Nucleotide Variant PHYLogenomics	SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology	https://snvphyl.readthedocs.io/en/latest/
5	Phylogenetic Inference	AMPHORA	We have developed an automated pipeline for phylogenomic analysis (AMPHORA) that overcomes the existing bottlenecks limiting large-scale protein phylogenetic inference. We demonstrated its high throughput capabilities and high quality results by constructing a genome	https://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-10-r151

			tree of 578 bacterial species and by assigning phylotypes to 18,607 protein markers identified in metagenomic data collected from the Sargasso Sea.	
5	Phylogenetic Inference	ASTRAL	ASTRAL is a tool for estimating an unrooted species tree given a set of unrooted gene trees. ASTRAL is statistically consistent under the multi-species coalescent model (and thus is useful for handling incomplete lineage sorting, i.e., ILS).	https://github.com/smirarab/ASTRAL
5	Phylogenetic Inference	BAMM	BAMM (Bayesian Analysis of Macroevolutionary Mixtures) is a program for modelling complex dynamics of speciation, extinction, and trait evolution on phylogenetic trees.	http://bamm-project.org/
5	Phylogenetic Inference	BEAST	BEAST is a cross-platform program for Bayesian analysis of molecular sequences using MCMC.	https://beast.community/
5	Phylogenetic Inference	BEAST2	BEAST 2 is a cross-platform program for Bayesian phylogenetic analysis of molecular sequences. It estimates rooted, time-measured phylogenies using strict or relaxed molecular clock models.	https://www.beast2.org/
5	Phylogenetic Inference	Beagle-lib	BEAGLE is a high-performance library that can perform the core calculations at the heart of most Bayesian and Maximum Likelihood phylogenetics packages.	https://github.com/beagle-dev/beagle-lib
5	Phylogenetic Inference	BiG-Scape	The 'Biosynthetic Gene Similarity Clustering and Prospecting Engine' (BiG-SCAPE), which facilitates fast and interactive sequence similarity network analysis of biosynthetic gene clusters and gene cluster families.	https://bigscape-corason.secondarymetabolites.org/
5	Phylogenetic Inference	ClustAGE	We have developed the open-source software package ClustAGE. This program, written in Perl, uses BLAST to cluster nucleotide accessory genomic elements from the genomes of multiple bacterial strains and to identify their distribution within the study population. The program output can be used in combination with strain phenotype data or other characteristics to detect associations.	https://pubmed.ncbi.nlm.nih.gov/29678129/

5	Phylogenetic Inference	cogent3	cogent3 is a mature python library for analysis of genomic sequence data. We endeavour to provide a first-class experience within Jupyter notebooks, but the algorithms also support parallel execution on compute systems with 1000's of processors.	https://pypi.org/project/cogent3/
5	Phylogenetic Inference	CORASON	The ‘Core Analysis of Syntenic Orthologs to Prioritise Natural Product Gene Clusters’ (CORASON), which elucidates phylogenetic relationships within and across these families.	https://bigscape-corason.secondarymetabolites.org/
5	Phylogenetic Inference	DendroPy	DendroPy is a Python library for phylogenetic computing. It provides classes and functions for the simulation, processing, and manipulation of phylogenetic trees and character matrices, and supports the reading and writing of phylogenetic data in a range of formats, such as NEXUS, NEWICK, NeXML, Phylip, FASTA, etc.	https://dendropy.org/
5	Phylogenetic Inference	ExaBayes	ExaBayes is a software package for Bayesian tree inference. It is particularly suitable for large-scale analyses on computer clusters.	https://cme.h-its.org/exelixis/web/software/exabayes/
5	Phylogenetic Inference	FastML	The FastML server is a bioinformatics tool for the reconstruction of ancestral sequences based on the phylogenetic relations between homologous sequences.	http://fastml.tau.ac.il/overview.php
5	Phylogenetic Inference	fastSTRUCTURE	fastStructure is an algorithm for inferring population structure from large SNP genotype data. It is based on a variational Bayesian framework for posterior inference and is written in Python2.x. Here, we summarise how to set up this software package, compile the C and Cython scripts and run the algorithm on a test simulated genotype dataset.	https://rajanil.github.io/fastStructure/
5	Phylogenetic Inference	FastTree	FastTree infers approximately-maximum-likelihood phylogenetic trees from alignments of nucleotide or protein sequences. FastTree can handle alignments with up to a million of sequences in a reasonable amount of time and memory.	http://www.microbesonline.org/fasttree/

5	Phylogenetic Inference	FreeBayes	The direct detection of haplotypes from short-read DNA sequencing data requires changes to existing small-variant detection methods. Here, we develop a Bayesian statistical framework that is capable of modelling multiallelic loci in sets of individuals with non-uniform copy number. We then describe our implementation of this framework in a haplotype-based variant detector, FreeBayes.	https://github.com/ekg/freebayes
5	Phylogenetic Inference	MrBayes	MrBayes is a program for Bayesian inference and model choice across a wide range of phylogenetic and evolutionary models. MrBayes uses Markov chain Monte Carlo (MCMC) methods to estimate the posterior distribution of model parameters.	http://nbisweden.github.io/MrBayes/
5	Phylogenetic Inference	PANGOLIN: Phylogenetic Assignment of Named Global Outbreak Lineages	Pangolin was developed to implement the dynamic nomenclature of SARS-CoV-2 lineages, known as the Pango nomenclature. It allows a user to assign a SARS-CoV-2 genome sequence the most likely lineage (Pango lineage) to SARS-CoV-2 query sequences.	https://cov-lineages.org/resources/pangolin.html
5	Phylogenetic Inference		PhyloBayes is a Bayesian Monte Carlo Markov Chain (MCMC) sampler for phylogenetic reconstruction using protein alignments. Compared to other phylogenetic MCMC samplers (e.g. MrBayes), the main distinguishing feature of PhyloBayes is the underlying probabilistic model, CAT. It is particularly well suited for large multigene alignments, such as those used in phylogenomics.	http://www.atgc-montpellier.fr/phylobayes/
5	Phylogenetic Inference	PhyML	PhyML is a software package that uses modern statistical approaches to analyse alignments of nucleotide or amino acid sequences in a phylogenetic framework.	https://github.com/stephaneguindon/phyml
5	Phylogenetic Inference	plink	PLINK is a comprehensive genome analysis toolset with an extensive list of functions. It was originally developed for human data (hence has all those human terms like “family” and “father”), but the new PLINK 1.9 can also be used with genomic data of non-model organisms.	https://www.cog-genomics.org/plink2/

5	Phylogenetic Inference	Pplacer	Pplacer places query sequences on a fixed reference phylogenetic tree to maximise phylogenetic likelihood or posterior probability according to a reference alignment. Pplacer is designed to be fast, to give useful information about uncertainty, and to offer advanced visualisation and downstream analysis.	http://matsen.fhcrc.org/pplacer/
5	Phylogenetic Inference	RaxML	A standard tool for Maximum-likelihood based phylogenetic inference.	https://cme.h-its.org/exelixis/software.html
5	Phylogenetic Inference	TASSEL: Trait Analysis by aSSociation, Evolution and Linkage	TASSEL is a software package used to evaluate traits associations, evolutionary patterns, and linkage disequilibrium.	https://www.maizegenetics.net/tassel
6	Pangenome	BGDMdocker	We introduce here a complete 16 and accurate bioinformatics workflow based on Docker to analyse and visualise pangenomes and biosynthetic 17 gene clusters of bacteria.	https://core.ac.uk/download/pdf/186968079.pdf
6	Pangenome	panX	panX is a software package for comprehensive analysis, interactive visualisation and dynamic exploration of bacterial pangenomes. The analysis pipeline is based on DIAMOND, MCL and phylogeny-aware post-processing.	https://pangenome.org/
6	Pangenome	Roary	Roary is a high speed stand alone pan genome pipeline, which takes annotated assemblies in GFF3 format (produced by Prokka (Seemann, 2014)) and calculates the pan genome.	https://sanger-pathogens.github.io/Roary/
7	Visualisation and Statistics	JSpecies	JSpecies is an easy to use, biologist-centric software designed to measure the probability of two genomes belonging to the same species or not.	http://www.imedea.uib.es/jspecies/about.html
7	Visualisation and Statistics	GraPhlAn: Graphical Phylogenetic Analysis	GraPhlAn is a software tool for producing high-quality circular representations of taxonomic and phylogenetic trees. GraPhlAn focuses on concise, integrative, informative, and publication-ready representations of phylogenetically- and taxonomically-driven investigation.	https://github.com/biobakery/graphlan
7	Visualisation and Statistics	HiGlass	HiGlass is a tool for exploring and comparing genomic contact matrices and tracks.	http://higlass.io/

7	Visualisation and Statistics	Parsnp	Here we present Parsnp and Gingr for the construction and interactive visualisation of massive core-genome alignments. For alignment, Parsnp combines the advantages of both whole-genome alignment and read mapping. Like whole-genome alignment, Parsnp accurately aligns microbial genomes to identify both structural and point variations, but like read mapping, Parsnp scales to thousands of closely related genomes. To achieve this scalability, Parsnp is based on a suffix graph data structure for the rapid identification of maximal unique matches (MUMs), which serve as a common foundation to many pairwise and multiple genome alignment tools.	
7	Visualisation and Statistics	iTOL: Interactive Tree Of Life	The is an online tool for the display, annotation and management of phylogenetic and other trees. Manage and visualise your trees directly in the browser, and annotate them with various datasets.	https://itol.embl.de/
8	Databases	EBI: European Bioinformatics Institute		https://www.ebi.ac.uk/
8	Databases	FunGuild	A python-based tool that can be used to taxonomically parse fungal OTUs by ecological guilds independent of sequencing platforms or analysis pipelines.	http://www.funguild.org/
8	Databases	GO	The Gene Ontology (GO) project (http://www. geneontology.org/) provides structured, controlled vocabularies and classifications that cover several domains of molecular and cellular biology and are freely available for community use in the annotation of genes, gene products and sequences. The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.	http://www.geneontology.org/

8	Databases	InterPro	Functional analysis of proteins by classifying them into families and predicting domains and important sites.	https://www.ebi.ac.uk/interpro/
8	Databases	KEGG: Kyoto Encyclopedia of Genes and Genomes KEGG	KEGG is a database resource for understanding high-level functions and utilities of the biological system	https://www.genome.jp/kegg/
8	Databases	KOG eukaryotic orthologous groups (KOGs)	A eukaryote-specific version of the Clusters of Orthologous Groups (COG) tool for identifying ortholog and paralog protein	https://mycocosm.jgi.doe.gov/Tutorial/tutorial/kog.html https://www.hsls.pitt.edu/obrc/index.php?page=URL1144075392
8	Databases	NCBI	National Center for Biotechnology Information.	www.ncbi.nlm.nih.gov
8	Databases	Pfam	A large collection of protein families.	https://pfam.xfam.org/
8	Databases	PR2	A reference database of carefully annotated 18S rRNA sequences using eight unique taxonomic fields.	https://pr2-database.org/
8	Databases	SEED	To provide consistent and accurate genome annotations across thousands of genomes and as a platform for discovering and developing de novo annotations.	https://pubseed.theseed.org/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965101/
8	Databases	TIGRFAM	patterns, and linkage disequilibrium. Strengths of this software include:	http://tigrfams.jcvi.org/cgi-bin/index.cgi
9	Platform	CoGe (Comparative Genomics)	CoGe is an online system for making the retrieval and comparison of genomic information and sequence data quick and easy.	https://genomevolution.org/coge/
9	Platform	NGPhylogeny.fr	Free, simple to use web service dedicated to reconstructing and analysing phylogenetic relationships between molecular sequences.	https://ngphylogeny.fr/
9	Platform	PGAP-X	PGAP-X is a microbial comparative genomic analysis platform with a graphic interface. Serials of algorithms and methodologies have been developed and integrated to analyse and visualise genomics structure variation, gene distribution with different conservative	https://pgapx.zhaopage.com/

			levels, and genetic variation from pangenome sight.	
9	Platform	Anvi'o	Anvi'o is an open-source, community-driven analysis and visualisation platform for microbial omics. It brings together many aspects of today's cutting-edge strategies including genomics, metagenomics, metatranscriptomics, pangenomics, metapangenomics, phylogenomics, and microbial population genetics in an integrated and easy-to-use fashion through extensive interactive visualisation capabilities.	https://merenlab.org/software/anvio/
9	Platform	Phylolink	Phylolink is a collection of tools through which biodiversity can be explored from a phylogenetic (or tree of life) perspective.	https://phylolink.ala.org.au/
9	Platform	Atlas of Living Australia	Species occurrence records with spatial and environmental data.	https://www.ala.org.au/
9	Platform	HyPhy (Hypothesis Testing using Phylogenies)	An open-source software package or comparative sequence analysis using stochastic evolutionary models	https://stevenweaver.github.io/hyphy-site/
9	Platform	Galaxy Australia	Galaxy is a web-based analysis and workflow platform.	https://usegalaxy.org.au/
9	Platform	R/R Studio	A development environment for R and Python, with a console, syntax-highlighting editor.	https://rstudio.com/

A complete list of tools with more details is available here.

Appendix 2

Survey[123] questions posed to the Comparative Genomics Research Community

1. How would you describe your level of experience with comparative genomics Analysis?

- ❑ Very experienced
- ❑ Some experience
- ❑ Beginner
- ❑ Interested but no direct experience
- ❑ Other:

2. Which types of comparative genome analysis do you / group members perform, or envisage performing in the next 5 years? Tick all that apply.

- ❑ Phylogenetics
- ❑ Phylogenomics
- ❑ Comparative Genomics
- ❑ Pangenomics
- ❑ Other:

3. With respect to comparative genomics analyses, which broad taxonomic groups will you work on in the next 5 years (choose all that apply)? Tick all that apply.

- ❑ Humans
- ❑ Vertebrate Animals
- ❑ Invertebrate Animals
- ❑ Vascular Plants
- ❑ Non-vascular Plants
- ❑ Fungi
- ❑ Ciliates
- ❑ Flagellates
- ❑ Other Eukaryotes not indicated above
- ❑ Archaea
- ❑ Bacteria
- ❑ Viruses
- ❑ Other:

4. Which reference databases do you use (choose all that apply)? **NB. this list is non-exhaustive so please note preferences not listed in 'other'. Tick all that apply.

- ❑ EBI
- ❑ NCBI

[123] Comparative Genomics Poll/Survey Results

- ❑ PFAM
- ❑ Other:

5. Which (if any) tools / software / pipelines / programs / platforms do you or group members use (choose all that apply)? Please only indicate those you'd currently recommend for use . **NB. this list is non-exhaustive so please note preferences not listed in 'other'. Tick all that apply.

- ❑ A custom tool developed in our group or by collaborators
- ❑ AGEnt
- ❑ Bacterial Pangenome Analysis (BPGA)
- ❑ bamm
- ❑ BEAST
- ❑ BEAST2
- ❑ BGDMdocker
- ❑ CD-HIT-EST
- ❑ CoGe
- ❑ ClustAGE
- ❑ ClustalW
- ❑ dendropy
- ❑ DeNoGAP
- ❑ diamond
- ❑ EDGAR
- ❑ EUPAN
- ❑ ExaBayes
- ❑ export graphlan
- ❑ fastml
- ❑ faststructure
- ❑ freebayes
- ❑ GARLI
- ❑ gblocks
- ❑ gcta
- ❑ GET_HOMOLOGUES
- ❑ GraPhIAn
- ❑ gubbins
- ❑ Harvest
- ❑ IQ-TREE
- ❑ ITEP
- ❑ jags
- ❑ LS-BSR
- ❑ mach
- ❑ MAFFT
- ❑ Mauve
- ❑ MeShClust
- ❑ micropan
- ❑ Minimac4
- ❑ Muscle
- ❑ MrBayes

- ❑ NGSPanPipe
- ❑ PanACEA
- ❑ Panaconda
- ❑ PanCake
- ❑ PanFunPro
- ❑ PanGeT
- ❑ PanGFR-HM
- ❑ PanGP
- ❑ PANINI
- ❑ PANNOTATOR
- ❑ PanOCT
- ❑ Panseq
- ❑ Pan-Tetris
- ❑ PanTools
- ❑ PanViz
- ❑ PanWeb
- ❑ panX
- ❑ Parsnp
- ❑ PGAdb-Builder
- ❑ PGAP
- ❑ PGAP-X
- ❑ Phylemon
- ❑ PhyML
- ❑ PICRUSt
- ❑ Piggy
- ❑ PipMaker
- ❑ plink
- ❑ pplacer
- ❑ probabel
- ❑ pyseer
- ❑ RAxML
- ❑ Roary
- ❑ rjags
- ❑ samblaster
- ❑ seq-seq-pan
- ❑ shapeit
- ❑ SNAPPy
- ❑ Spine
- ❑ SplitMEM
- ❑ Svtyper
- ❑ tassel
- ❑ TreePuzzle
- ❑ UCLUST
- ❑ VISTA
- ❑ Other:

6. Are there tools / software / pipelines / programs / platforms you'd like to use but that aren't suitable for your study taxon/taxa? If so, what are they and why aren't they suitable?

7. Are there tools / software / pipelines / programs / platforms 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?

8. Do you require custom or proprietary tools / software for your analysis approach? If so, what are they?

9. What sequencing platform/s are you currently using to generate data (choose all that apply)? Tick all that apply.

- ❑ Illumina
- ❑ PacBio
- ❑ 10 X
- ❑ Nanopore
- ❑ Ion Torrent
- ❑ Other:

10. Do you make use of existing datasets from the same taxon of closely related taxa (choose all that apply)? Tick all that apply.

- ❑ 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:

11. Do you use a data management tool/framework within your comparative genomics project(s)? If so, what?

12. How do you share data within your group and with collaborators? Where are your collaborators based? What difficulties have you encountered?

13. Do you make your comparative genomics datasets publicly available? If so, where? Have you encountered any difficulties in doing so?

14. If you don't make your comparative genomics datasets publicly available,

why not? Tick all that apply.

- ❑ Commercial confidence issues
- ❑ I don't see a benefit in sharing my comparative genomic datasets publicly available
- ❑ I don't know how to make my comparative genomic datasets publicly available
- ❑ No international repository exists in which to deposit the data
- ❑ Other:

15. What kind of compute infrastructure setup do you use for comparative genomics (choose all that apply)? Tick all that apply.

- ❑ Local desktop/PC
- ❑ High-performance computing at my institution
- ❑ High-performance computing at a collaborator's institution
- ❑ High-performance computing within my research group
- ❑ High-performance computing within my department
- ❑ National or state high-performance computing infrastructure (e.g. NCI, Pawsey, QCIF/QRIScloud)
- ❑ NeCTAR cloud instance
- ❑ Commercial cloud (e.g. Amazon Web Services, Microsoft Azure, Google Cloud)
- ❑ Galaxy
- ❑ Other:

16. Do you have access to the expertise you need to build and maintain this compute infrastructure (e.g. installing and updating software)? Tick all that apply.

- ❑ 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:

17. Is your current compute infrastructure sufficient for your current needs? If no, why not?

18. Will this compute infrastructure setup be sufficient for your needs in 2 years' time? Tick all that apply.

- ❑ Yes, we expect to be doing comparative genomics at a similar scale in 2 years' time
- ❑ Yes, as we expect to be doing less comparative genomics in 2 years' time
- ❑ No, we expect to be doing more comparative genomics 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:

19. Will this compute infrastructure setup be sufficient for your needs in 5 years' time? Tick all that apply.

- ❑ Yes, we expect to be doing comparative genomics at a similar scale in years' time
- ❑ Yes, as we expect to be doing less comparative genomics in 5 years' time
- ❑ No, we expect to be doing more comparative genomics 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:

20. Would you / group members use a shared compute infrastructure platform to perform comparative genomics? Tick all that apply.

- ❑ Yes - we can't currently perform comparative genomics 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 comparative genomics locally no matter how good a shared platform is
- ❑ Other:

21. How important are these general factors to you in a shared comparative genomics platform?

- ❑ Following best practice in tools, formats and metadata; compliant with requirements of international data repositories
- ❑ Free (subsidised) to researchers no matter the scale of analysis
- ❑ Easy to access from anywhere
- ❑ Easy to self-manage access and permissions for collaborators
- ❑ Easy to upload/download data
- ❑ Security of data and analysis
- ❑ Long-term support for and sustainability of the platform

22. How important are these data-related factors to you in a shared comparative genomics platform?

- ❑ Smart metadata handling (e.g. assistance with metadata formats, transfer of metadata through pipeline, controlled vocabulary lookup)
- ❑ Ability to submit datasets to international repositories from the platform
- ❑ Ability to download datasets from international repositories within the platform
- ❑ Ability to transfer data easily to/from storage

23. How important are these tool/pipeline-related factors to you in a shared

comparative genomics platform?

- ❑ Access to our preferred tools/pipelines
- ❑ Access to a choice of tools/pipelines
- ❑ Quick installation of other tools/pipelines upon request
- ❑ Assistance available in implementing pipelines

24. What are the top 1-5 tools/pipelines you would absolutely require in a shared comparative genomics platform?

25. How important are these compute-related factors to you in a shared comparative genomics platform?

- ❑ Ability to scale up/down resources used as needed
- ❑ No need to understand or control the compute backend
- ❑ Compatibility with external analysis environments (e.g. Amazon, Cyverse)

26. How important are these training-related factors to you in a shared comparative genomics platform?

- ❑ Good documentation on how to use the platform
- ❑ Good documentation on how to use the tools/pipelines
- ❑ Access to in-person training on how to use the platform
- ❑ Access to in-person training on how to use the tools/pipelines
- ❑ Discussion forum to share expertise with other users

27. Are there any other factors you consider crucial in a shared comparative genomics platform? If so, what? Please let us know here.

Document Control

VERSION	DATE	AUTHOR(S)	DESCRIPTION
V1.0	30/11/2021	Tiffanie Nelson, Jeff Christiansen	A preliminary document detailing the outline of the roadmap draft including the software list obtained from researchers.
V2.0	20/12/2021	Tiffanie Nelson, Jeff Christiansen	Revised draft version. Preliminary sharing with the broader BioCommons team and expert leaders in the field of comparative genomics.
V3.0	22/03/2022	Tiffanie Nelson, Jeff Christiansen	Revised draft version. Sharing with the broader group of comparative genomics researchers and infrastructure providers and international experts[124].
V4.0	05/09/2022	Tiffanie Nelson	Finalised draft incorporating feedback from international experts.

[124] This version of the Roadmap has been reviewed by the following international experts: Dr Paul Kersey, Deputy Director of Science-Research, Royal Botanic Gardens, Kew, UK and Dr Tim Sackton, Director of Bioinformatics, Informatics Group Faculty of Arts and Sciences, Harvard University, USA.